ỨNG DỤNG THUẬT TOÁN MÁY HỌC ĐỂ XÂY DỰNG MÔ HÌNH PHÂN BỐ LOÀI SAO ĐEN (*Hopea odorata*) TRÊN GOOGLE EARTH ENGINE

Nguyễn Thanh Tuấn[1*], Phùng Văn Phê[2]
[1]*Trường Đại học Lâm nghiệp - Phân hiệu Đồng Nai*
[2]*Trường Đại học Lâm nghiệp*

https://doi.org/10.55250/jo.vnuf.2023.3.091-100

TÓM TẮT

Sao đen (*Hopea odorata*) là loài cây có vai trò quan trọng về sinh thái cũng như kinh tế ở rừng tự nhiên ở Việt Nam. Hệ sinh thái rừng bị suy thoái đã đe dọa đến sự tồn tại của loài trong tự nhiên. Từ những lý do trên cần thiết phải có những chiến lược để bảo tồn loài cây có giá trị này. Nghiên cứu đã sử dụng 4 thuật toán máy học khác nhau (rừng ngẫu nhiên, độ dốc tăng cường, véc-tơ hỗ trợ, cây phân loại và hồi quy) để mô hình hóa vùng phân bố tiềm năng của Sao đen dựa vào dữ liệu được thu thập từ 117 điểm xuất hiện loài kết hợp với 19 nhân tố khí hậu cùng với dữ liệu tài nguyên đất và địa hình. Kết quả nghiên cứu chỉ ra rằng vùng sinh thái phù hợp cho loài Sao đen khoảng 1.227.788 km^2 thuộc 13 quốc gia: Bangladesh, Myanmar, Camphuchia, Sri Lanka, Trung Quốc, Indonesia, Ấn độ, Lào, Malaysia, Singapore, Thái Lan và Việt Nam. Ngoài ra, nhiệt độ mùa (bio4), hàm lượng đạm (nitrogen), trữ lượng carbon hữu cơ (ocs), mảnh thô lẫn trong đất (cfvo), dung trọng của đất (bdod) là các nhân tố quan trọng ảnh hưởng đến phân bố tiềm năng loài này. Thuật toán rừng ngẫu nhiên trên nền tảng Google Earth Engine đã thể hiện được những ưu thế vượt trội trong việc xây dựng bản đồ môi trường sống phù hợp cho phục hồi và bảo tồn loài Sao đen trên phạm vi khu vực Đông Á, Nam Á và Đông Nam Á.

Từ khóa: **biến khí hậu, cây họ Dầu, rừng ngẫu nhiên, sinh cảnh phù hợp, tính chất đất.**

1. ĐẶT VẤN ĐỀ

Không gian phân bố và thời gian xuất hiện loài là những câu hỏi thuộc kiến thức nền tảng của sinh thái học. Những câu hỏi này đã thúc đẩy sự phát triển của các mô hình phân bố loài (species distribution models-SDMs), còn gọi là mô hình thích hợp sinh thái hoặc mô hình môi trường sống phù hợp [1]. Các mô hình này dựa trên dữ liệu các điểm ghi nhận loài và biến dự báo để đánh giá phân bố không gian của các loài, cùng với sự phù hợp với môi trường sống của chúng. Mô hình phân bố loài đã trở thành một nội dung quan trọng của các nghiên cứu trong các lĩnh vực sinh thái học, tiến hóa sinh học, quản lý và bảo tồn động vật hoang dã [2-4]. Hiện nay, mô hình phân bố loài được sử dụng rộng rãi để phân tích dữ liệu xuất hiện và dự đoán sự phù hợp với môi trường sống cho các loài động vật và thực vật [5]. Sự tăng lên SDMs có liên quan đến sự phát triển của công nghệ máy tính và phần mềm để thực hiện một lượng lớn các thuật toán khác nhau [2, 6-8]. Hiện nay, các thuật toán được sử dụng trong SDMs chủ yếu như: tuyến tính (Linear model), mô hình hồi quy cộng tính tổng quát, rừng ngẫu nhiên (Random forest), mô hình độ dốc tăng cường cấp cao (Grandient extrem boot), Entropy cực đại, véc-tơ hỗ trợ, Cây phân loại và hồi quy (CART) [2]. Trong số các thuật toán máy học thì Support Vector Machine (SVM) và Random Forest (RF) đã nổi lên như là các phương pháp tiềm năng để xây dựng SDMs.

Các thuật toán máy học phát triển đã cho phép các nhà nghiên cứu sử dụng được nguồn dữ liệu to lớn và miễn phí từ viễn thám và hệ thống thông tin địa lý trong SDMs [9-11]. Tuy nhiên, việc truy cập dữ liệu raster lớn và từ nhiều nguồn khác nhau đòi hỏi sức mạnh tính toán và thời gian xử lý, đó là những yếu tố hạn chế đối với nhiều nhà nghiên cứu, bảo tồn và các nhà quản lý [1].

**Corresponding author: nttuan@vnuf2.edu.vn*

Google Earth Engine (GEE) chạy trên nền tảng đám mây có thể xử lý và phân tích hàng loạt các hình ảnh vệ tinh và dữ liệu không gian địa lý, giúp cho các quy trình công việc đòi hỏi tính toán số lượng lớn có thể được truy cập và xử lý dễ dàng hơn. Ngoài ra, dựa trên công nghệ phân tích địa lý đám mây được lưu trữ bởi Google và miễn phí cho tất cả người dùng với mục đích nghiên cứu, cung cấp quyền truy cập với hiệu suất cao, hệ thống tính toán song song độc lập chia nhỏ các phép tính trên Google làm tăng tốc quá trình tính toán [12]. Google Earth Engine đã tạo ra hướng nghiên cứu mới cho các nhà khoa học, với hàng nghìn bộ dữ liệu và thuật toán trong một nền tảng trực tuyến duy nhất, giúp tăng tốc quá trình xử lý dữ liệu và cải thiện hiệu quả tính toán [13].

Sao đen (*Hopea odorata*) chất lượng gỗ tốt, không bị mối mọt, kích thước lớn, thân gỗ cao nên rất được yêu thích để đóng đồ đạc, sàn nhà, toa xe hay tàu thuyền. Ngoài ra, vỏ cây Sao đen chứa nhiều tanin nên được dùng trong dược phẩm [14]. Mặc dù đã có rất nhiều nỗ lực trong phát triển rừng trồng Sao đen, tuy nhiên sự suy giảm phân bố của loài ngoài tự nhiên vẫn diễn ra hết sức nghiêm trọng. Do vậy, cần có những chiến lược để bảo tồn vùng phân bố tự nhiên của loài, cùng với gây trồng phục hồi loài ở những vùng sinh thái phù hợp. Xuất phát từ những vấn đề đặt ra ở trên, nghiên cứu đã sử dụng thuật toán máy học để xây dựng mô hình phân bố tự nhiên của loài Sao đen trên Google earth engine phục vụ công tác bảo tồn và phát triển loài.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Phương pháp thu thập dữ liệu

Các điểm ghi nhận loài Sao đen xuất hiện được tải từ trang Global Biodiversity Information Facility.

Ngoài ra, nghiên cứu sử dụng dữ liệu khí hậu (từ năm1970 đến 2000) được tải từ trang web của Worldclim (www.worldclim.org) có độ phân giải là 30 arc-seconds (độ phân giải không gian 1km^2) để xác định vùng phân bố thích hợp cho loài, dữ liệu bao gồm 19 biến khí hậu: Nhiệt độ trung bình hàng năm (bio1:°C); khoảng nhiệt độ trung bình ngày đêm (bio2: °C); đường đẳng nhiệt (bio3); nhiệt độ mùa (bio4: °C); nhiệt độ cao nhất của tháng ấm nhất (bio5: °C); Nhiệt độ thấp nhất của tháng lạnh nhất (bio6: °C); nhiệt độ dao động hàng năm (bio7: °C); nhiệt độ trung bình của quý ẩm nhất, khô nhất, ấm nhất và lạnh nhất (bio8, bio 9, bio10, bio11: °C); lượng mưa hàng năm (bio12: mm); lượng mưa của tháng ẩm ướt nhất (bio13: mm); lượng mưa của tháng khô nhất (bio14: mm); lượng mưa của mùa (bio15: mm); lượng mưa của quý ẩm nhất, khô nhất, ấm nhất và lạnh nhất (bio16, bio17, bio18, bio19: mm). Mô hình cao độ (Digital Elevation Model-DEM) được thu thập từ dữ liệu cao độ số toàn cầu ASTER với độ phân giải 30 arc-seconds từ trang WorldClim 2.1.

Đặc điểm thổ nhưỡng bao gồm 11 biến được tải từ trang Soilgrid (https://soilgrids.org) với mức độ sâu tầng đất là 0-5 cm, bao gồm: Mật độ cacbon hữu cơ (ocd: g/dm^3), trữ lượng cacbon hữu cơ (ocs: tấn/ha), dung trọng (bdod: cg/cm^3), thành phần sét (clay: g/kg), mảnh thô lẫn trong đất (cfvo: cm^3/dm^3), cát (sand: g/kg), đất thịt (silt: g/kg), khả năng trao đổi cation (cec: mmol(c)/kg), đạm (nitrogen: cg/kg), cacbon hữu cơ (soc: dg/kg), pH (phh2o).

2.2. Phương pháp xây dựng mô hình phân bố loài

(1) Thuật toán rừng ngẫu nhiên (Random forest – RF)

Mô hình rừng ngẫu nhiên (RF) là một số lượng lớn các cây hồi quy được xác định hoàn toàn ngẫu nhiên từ các biến đầu vào (có thể là biến liên tục hoặc rời rạc) để xác định giá trị đầu ra. Các giá trị đầu ra sau đó được xác định bằng trung bình cộng kết quả đầu ra từ tất cả các cây hồi quy. Hai tham số cần được xác định trong thuật toán phân loại này là ntree (số lượng cây

được phát triển) và mtry (số lượng biến để phân chia tại mỗi node). Số ntree được lựa chọn phụ thuộc vào khoảng thời gian xử lý ngắn nhất để kết quả đạt được độ sai số thấp nhất, ntree chạy từ 1 đến 1000 cây và mtry biến động từ số biến độc lập tối thiểu bằng 1 đến số biến độc lập tối đa được sử dụng trong phân loại [15].

(2) Thuật toán véc tơ hỗ trợ (Support vector machine – SVM)

Phương pháp máy vec-tơ hỗ trợ (SVM) là thuật toán thống kê dựa trên phương pháp hạt nhân (kernel) để chuyển hồi quy phi tuyến sang tuyến tính trong không gian nhiều chiều, có thể sử dụng cho thuật toán phân lớp đối tượng hoặc hồi quy [16] . Chẳng hạn mẫu huấn luyện ban đầu: (x_i, y_i), (i=1,2…n), trong đó x_i là đa biến số đầu vào, y_i là đầu ra vô hướng và n là số mẫu huấn luyện. Với phương pháp SVM mô hình ban đầu sẽ được chuyển thành mô hình tuyến tính trong không gian nhiều chiều mới [17]:

$$y = f(x) = \langle w.\varphi(x)\rangle + b = \sum_{i=1}^{n} w_i\varphi_i(x) + b \qquad (1)$$

Trong đó:

w là trọng số của vector;

b là độ dịch;

φ là hàm phi tuyến chuyển đổi từ không gian đầu vào thành không gian nhiều chiều mới. Thay vì xác định chính xác dạng hàm của φ, chúng ta sử dụng hàm thức hạt nhân:

$$K(x_i, x) = \langle \varphi(x_i).\varphi(x)\rangle \qquad (2)$$

Thông thường hàm thức hạt nhân bao gồm hàm tuyến tính, đa thức bậc cao và hàm cơ sở bán kính. Mặt khác, chúng ta cần xác định w và b dựa vào sai số nhỏ nhất của hồi quy dựa vào công thức:

$$minimize\ \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\xi_i + \xi_i^*$$
$$subject\ to \begin{cases} y_i - f(x_i) \le \varepsilon + \xi_i^* \\ f(x_i) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, i = 1, \dots, n \end{cases} \qquad (3)$$

Để lựa chọn hàm thức hạt nhân và giá trị C tối ưu cho mô hình SVM, nghiên cứu tiến hành chạy thử nghiệm giá trị C và các hàm thức hạt nhân khác nhau để chọn ra giá trị phù hợp với sai số của mô hình là nhỏ nhất.

(3) Thuật toán độ dốc tăng cường (Grandient extrem boot-Boosted)

Mô hình độ dốc tăng cường cấp cao (Boosted) là một kỹ thuật đồng bộ nhằm mục đích tạo ra các phương pháp phân loại mạnh từ các phương pháp phân loại yếu. Điều đó được thực hiện bằng cách xây dựng các mô hình từ dữ liệu đào tạo và từ đó tạo ra một mô hình thứ hai sửa lỗi từ mô hình đầu tiên. Các mô hình sẽ tiếp tục được thêm vào cho đến khi tập đào tạo được dự đoán hoàn hảo hoặc thêm một số mô hình tối đa [18] .

Xuất phát từ mô hình hiện tại, xây dựng một cây quyết định cố gắng khớp phần dư từ mô hình trước. Điểm đặc biệt của mô hình này đó là thay vì cố gắng khớp giá trị biến mục tiêu là y thì chúng ta sẽ tìm cách khớp giá trị sai số của mô hình trước đó. Sau đó sẽ đưa thêm mô hình huấn luyện vào hàm dự báo để cập nhật dần dần phần dư. Mỗi một cây quyết định trong chuỗi mô hình có kích thước rất nhỏ với chỉ một vài nodes quyết định được xác định bởi tham số độ sâu d trong mô hình. Bằng cách khớp trên những cây quyết định có kích thước rất nhỏ trên những phần dư, sẽ từ từ cải hiện hàm dự báo $\hat{f}$ trong vùng mà nó không được dự báo tốt. Giả định $\hat{f}(x)$ là hàm dự báo từ *phương pháp tăng cường* được áp dụng trên một tác vụ dự báo với ma trận đầu vào X và biến mục tiêu là véc tơ y. Tại mô hình thứ b trong chuỗi mô hình dự báo, kí hiệu là $\hat{f}^b$, ta tìm cách khớp một giá trị phần dư r_i từ cây quyết định tiền nhiệm $\hat{f}^{b-1}$. Các bước trong quá trình huấn luyện mô hình theo phương

pháp tăng cường được tóm tắt như sau:

Ban đầu ta thiết lập hàm dự báo $\hat{f}(x)$=0 và số dư r_0=y cho toàn bộ quan sát trong tập huấn luyện.

Lặp lại quá trình huấn luyện cây quyết định theo chuỗi tương ứng với b=1,2,…B. Với một lượt huấn luyện gồm các bước con sau đây:

+Khớp một cây quyết định $\hat{f}^b$ có độ sâu làtrên tập huấn luyện (X, r_b)

+ Cập nhật $\hat{f}$ bằng cách cộng thêm vào giá trị dự báo của một cây quyết đinh, giá trị này được nhân với hệ số co λ:

$$\hat{f}(x)= \hat{f}(x)+ \lambda\hat{f}^b(x) \quad (4)$$

+ Cập nhật phần dư cho mô hình:

$r_{b+1}=r_b-\lambda\hat{f}^b(x)$

Thuật toán sẽ dừng cập nhật khi số lượng cây quyết định đạt ngưỡng tối đa B hoặc toàn bộ các quan sát trên tập huấn luyện được dự báo đúng.

Kết quả dự báo từ chuỗi mô hình sẽ là kết hợp của các mô hình con:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda\hat{f}^b(x) \quad (5)$$

(4) Thuật toán cây phân loại và hồi quy (Classification and regression tree- CART)

Thuật toán CART chia không gian n chiều thành các hình chữ nhật không chồng lên nhau bằng phép đệ quy. Đầu tiên, một biến độc lập x_i được chọn, và sau đó xác định một giá trị ui tương ứng. Không gian n chiều được chia thành hai phần. Một số điểm thỏa mãn $x_i \leq u_i$, và những điểm khác thỏa mãn $x_i > u_i$. Đối với một biến không liên tục, chỉ có hai giá trị là bằng hoặc không bằng nhau. Trong quá trình xử lý đệ quy, hai phần này dựa vào bước đầu tiên để chọn lại một thuộc tính và tiếp tục phân vùng cho đến khi chia hết không gian n chiều. Các thuộc tính có giá trị hệ số GINI tối thiểu được sử dụng làm chỉ mục phân vùng. Đối với tập dữ liệu D, hệ số GINI được xác định như sau:

$$GINI \times (D) = 1 - \sum i = kp_i^2 \quad (6)$$

Trong đó k là số loại mẫu và p_i biểu thị xác suất một mẫu được xếp vào loại i. Giá trị GINI càng nhỏ có nghĩa là chất lượng của mẫu càng cao và hiệu ứng phân loại càng tốt.

Cây quyết định bao gồm các nút nhiều cấp và nhiều lá. Các nút tối đa đề cập đến số lượng lá tối đa trên mỗi cây và quần thể lá tối thiểu là số lượng nút tối thiểu chỉ được tạo cho tập huấn luyện. Để xây dựng một cây phù hợp, phải tạo đủ các nút và nhánh. Giá trị nút tối đa là không giới hạn nếu nó không được chỉ định [19].

2.3. Phương pháp đánh giá độ chính xác của mô hình phân bố loài

Để so sánh độ chính xác của các thuật toán học máy trong xây dựng SDMs, nghiên cứu sử dụng các chỉ tiêu:

Độ chính xác (Precision-PR): Là tỉ lệ số điểm true positive trong số những điểm được phân loại là positive [20].

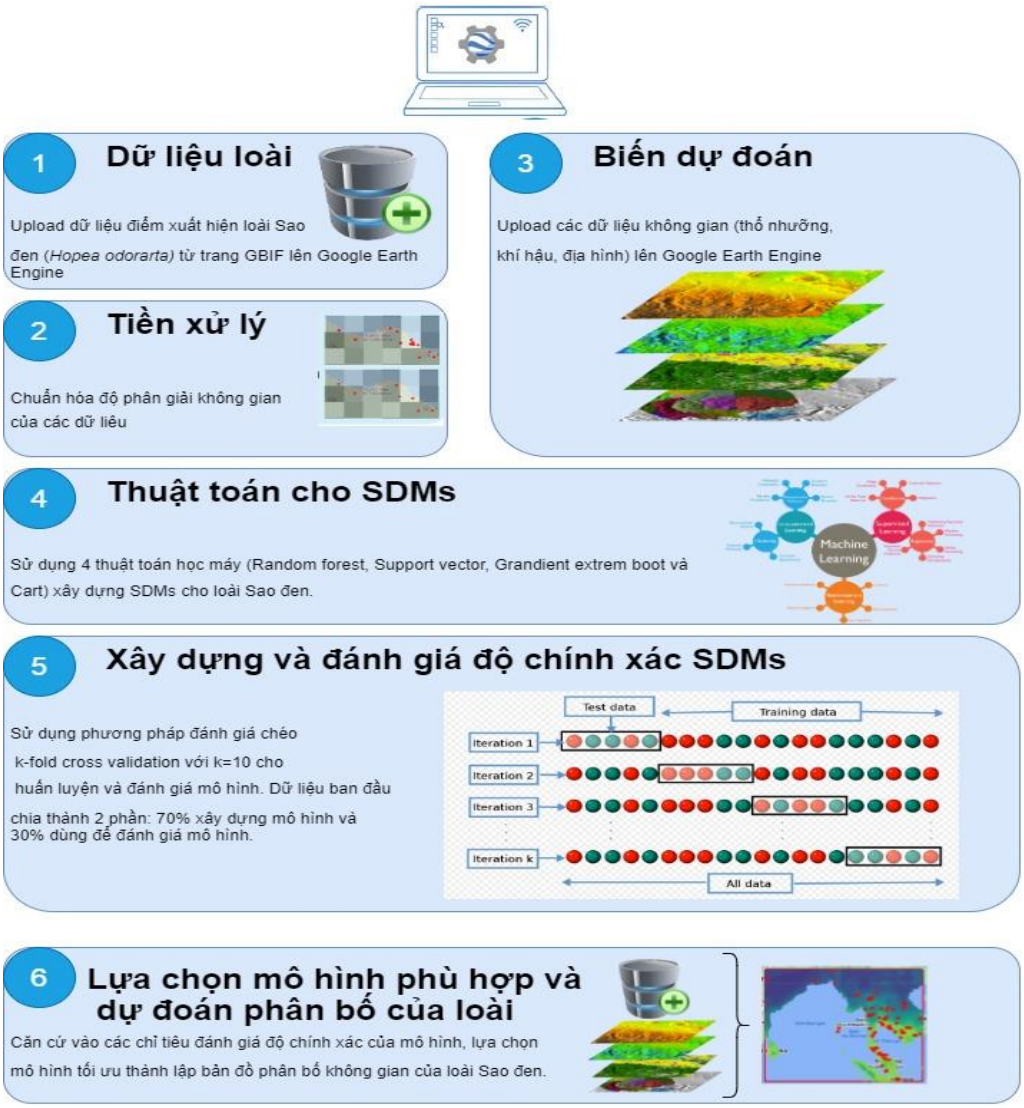
Độ đặc hiệu (Specificity-TNR): Là tỉ lệ số điểm true negative trong số những điểm được phân loại là negative [20].

AUC-ROC (Diện tích dưới đường cong - Area Under The Curve): Là xác suất một mẫu dương tính được lấy ngẫu nhiên sẽ được xếp hạng cao hơn một mẫu âm tính. Chỉ số AUC-ROC càng cao thì mô hình càng chính xác trong việc phân loại các lớp, giá trị nằm trong khoảng từ 0-1[20]. Trong đó, ROC (Receiver operating characteristic) là Đường cong hoạt động của bộ thu nhận.

AUC-PR: Giá trị nằm trong khoảng từ 0-1. So với chỉ số AUC-ROC thì chỉ số AUC-PR không bị ảnh hưởng bởi số lượng điểm mà không có loài xuất hiện [21].

Cuối cùng, theo K. Zhang và cộng sự (2018) mức độ thích hợp của loài được đánh giá thông qua xác suất xuất hiện loài ở 4 cấp: không thích hợp (0 - 0,2), mức thấp (0,2 - 0,4), trung bình (0,4 - 0,6), thích hợp cao (0,6 – 1) [22].

Toàn bộ sơ đồ xây dựng SDMs cho loài Sao đen được thể hiện ở Hình 1.



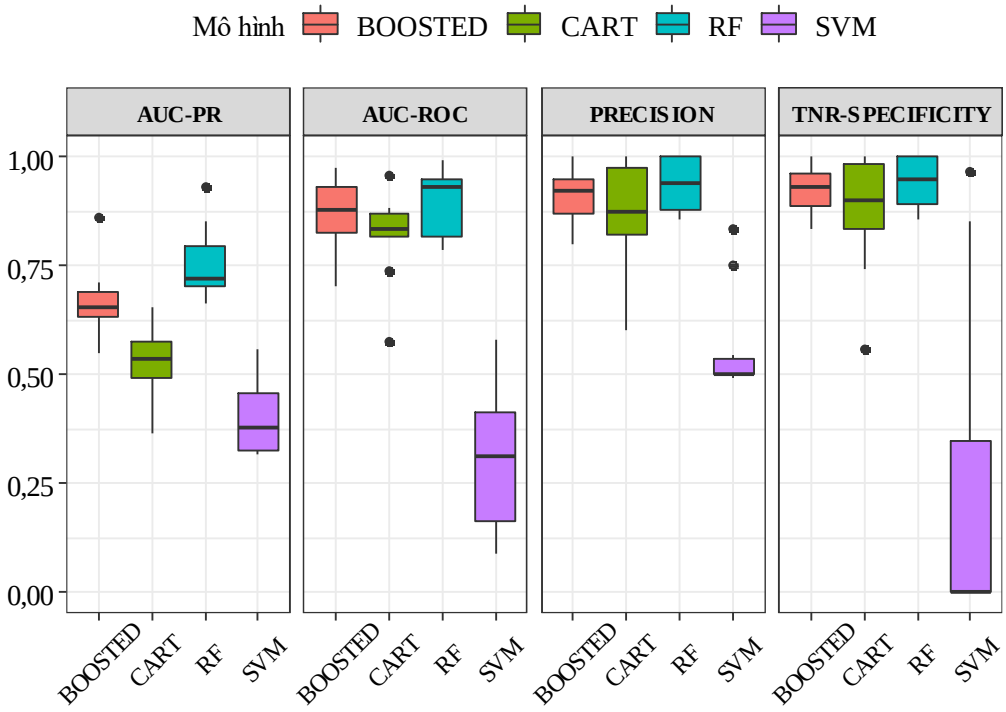
Hình 1. Sơ đồ xây dựng phân bố cho loài Sao đen trên Google Earth Engine

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Lựa chọn mô hình máy học mô phỏng phân bố Sao đen

Kết quả đánh giá độ chính xác mô hình phân bố loài của 4 thuật toán học máy cho phần dữ liệu kiểm chứng cho thấy thuật toán rừng ngẫu nhiên (RF) có độ chính xác cao nhất trong cả 4 chỉ tiêu đánh giá với giá trị AUC-PR = 0,75; AUC-ROC= 0,89; độ chính xác (precision) = 0,94 và giá trị TRN-specificity =0,94. Tiếp đến là thuật toán độ dốc tăng cường cấp cao (Boosted) với giá trị AUC-PR, AUC-ROC, độ chính xác và giá trị TRN-specificity lần lượt là 0,66; 0,87; 0,91 và 0,93. Trong khi đó, thuật toán CART là 0,52; 0,82; 0,87 và 0,87. Thấp nhất là mô hình SVM với 0,40; 0,30, 0,56 và 0,23 (Hình 2).

Tương tự với kết quả nghiên cứu này, thuật toán RF cũng thể hiện được sự vượt trội so với 8 thuật toán học máy khác (SVM, GARP, DT, RIPPER, KNN, Logistic, ANN và NativeBayes) khi xây dựng mô hình phân bố cho 35 loài thực vật ở khu vực Mỹ La tinh với độ chính xác AUC từ 0,82-0,96 [23]. Thuật toán RF được sử dụng rộng rãi trong các mô hình ứng dụng ở các lĩnh vực khác nhau và đào tạo mô hình cũng dễ dàng với việc thử nghiệm số lượng cây quyết định khác nhau. Mặt khác mô hình RF tránh được hiện tượng overfitting khi xây dựng mô hình [24].



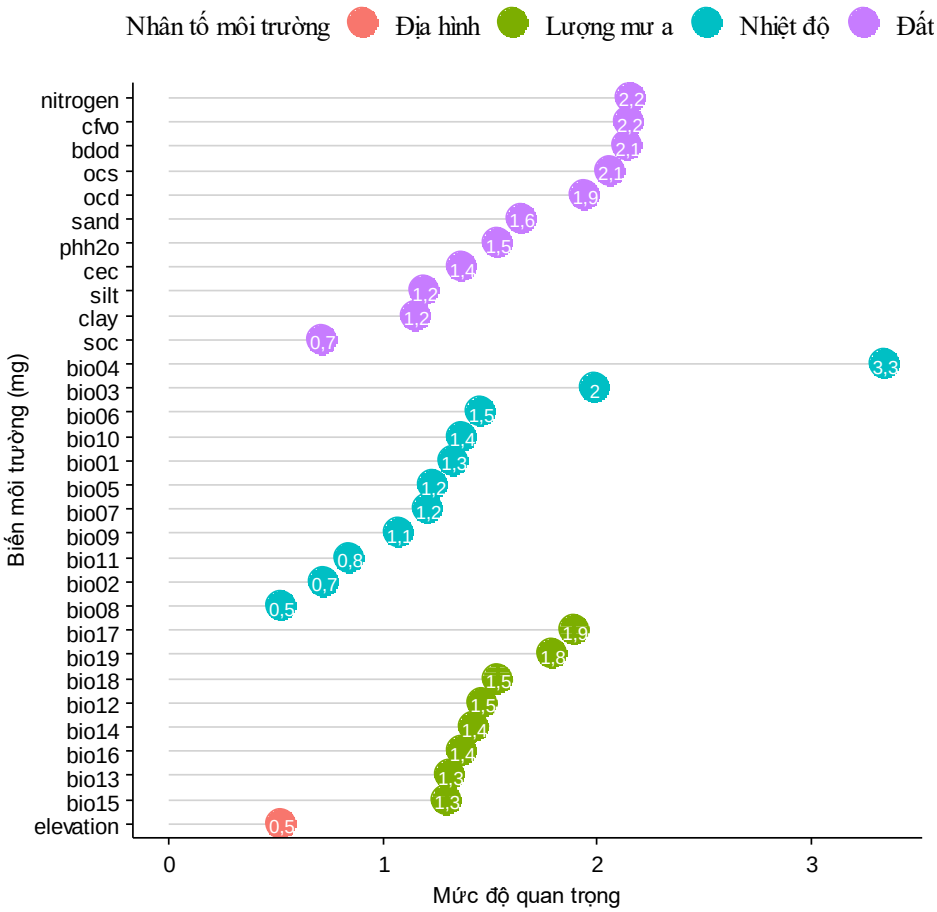
Hình 2. Độ chính xác của các thuật toán học máy

3.2. Ảnh hưởng của một số nhân tố môi trường đến phân bố Sao đen

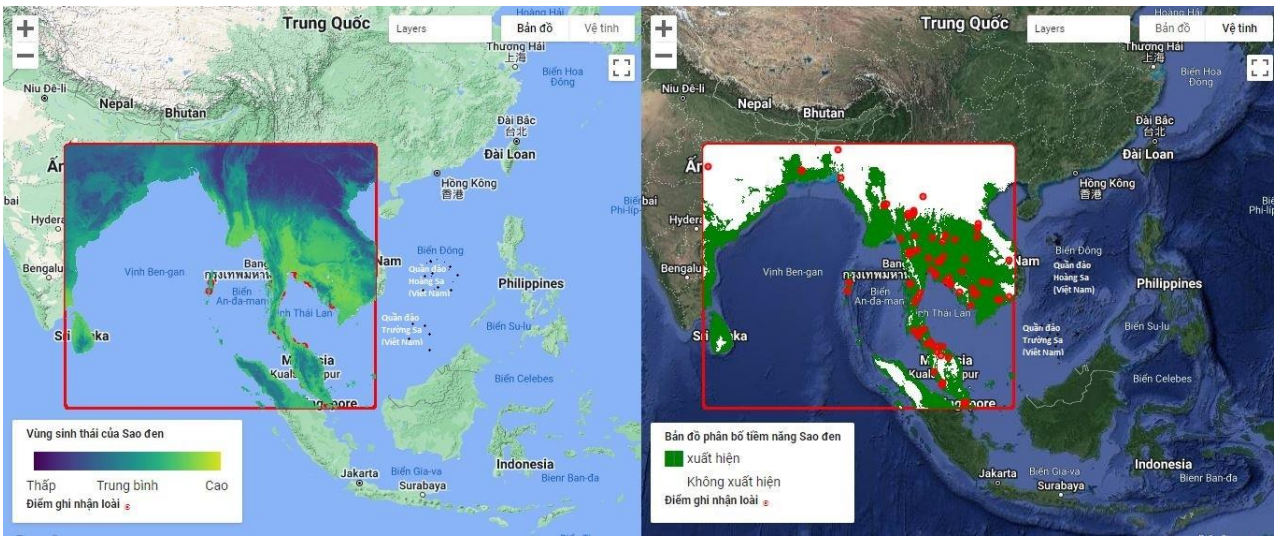
Sự ảnh hưởng các yếu tố môi trường đến phân bố địa lý của Sao đen theo mô hình RF được sắp xếp theo mức độ đóng góp của các biến đến kết quả mô hình từ cao đến thấp được thể hiện ở Hình 3. Trong số 4 nhóm nhân tố sinh thái ảnh hưởng đến phân bố của Sao đen, đất đai có mức độ quan trọng cao nhất (18,06), tiếp theo là nhiệt độ (15,1), lượng mưa (12,09) và cuối cùng là địa hình (0,53). Nếu xét từng nhân tố sinh thái ảnh hưởng đến mô hình phân bố của Sao đen thì nhiệt độ mùa (Bio4) ảnh hưởng lớn nhất với mức độ quan trọng (3,3), tiếp đến là các đặc điểm của đất đai như hàm lượng đạm (nitrogen), trữ lượng carbon hữu cơ (ocs), mảnh thô lẫn trong đất (cfvo) và dung trọng (bdod) với mức độ quan trọng lớn hơn 2,0. Ngược lại, độ cao có mức độ ảnh hưởng thấp nhất đến mô hình phân bố của Sao đen với mức độ quan trọng bằng 0,5.

Kết quả này phù hợp với các nghiên cứu trước đây cho thấy đặc điểm khí hậu như lượng mưa và nhiệt độ đóng một vai trò hiệu quả trong việc phân định khu vực địa lý của các loài thực vật đặc biệt là họ Dầu ở Đông Nam Á (Laos, Vietnam, Cambodia, Thailand, Myanmar and Indo-Malayan) và Nam Á (India, Pakistan and Sri Lanka) [25]. Tương tự, nhiệt độ mùa là nhân tố quan trọng nhất chi phối phân bố không gian của loài *P. dielsiana* tại khu vực cận nhiệt đới Trung Quốc [26].

Các thông số đất, đặc biệt là mảnh thô lẫn trong đất, ảnh hưởng đến sự phân bố không gian của các khu rừng lá rộng rụng lá và cận nhiệt đới [27-29]. Ở quy mô không gian lớn, các chất dinh dưỡng cung cấp cho thực vật bị ảnh hưởng do lượng nước trong đất bị thay đổi . Trong khi đó, mảnh thô lẫn trong đất có vai trò quan trọng ảnh hưởng đến khả năng giữ nước của đất . Theo nghĩa này, các tính chất của đất có thể định hình phân bố không gian của Sao đen do chi phối hàm lượng nước trong đất. Các nghiên cứu trước đây cũng chỉ ra rằng mối quan hệ giữa đất (hàm lượng đạm, mảnh thô lẫn trong đất và carbon hữu cơ) và sự phân bố không gian của thực vật [30-32].



Hình 3. Mức độ quan trọng các nhân tố sinh thái tới phân bố Sao đen



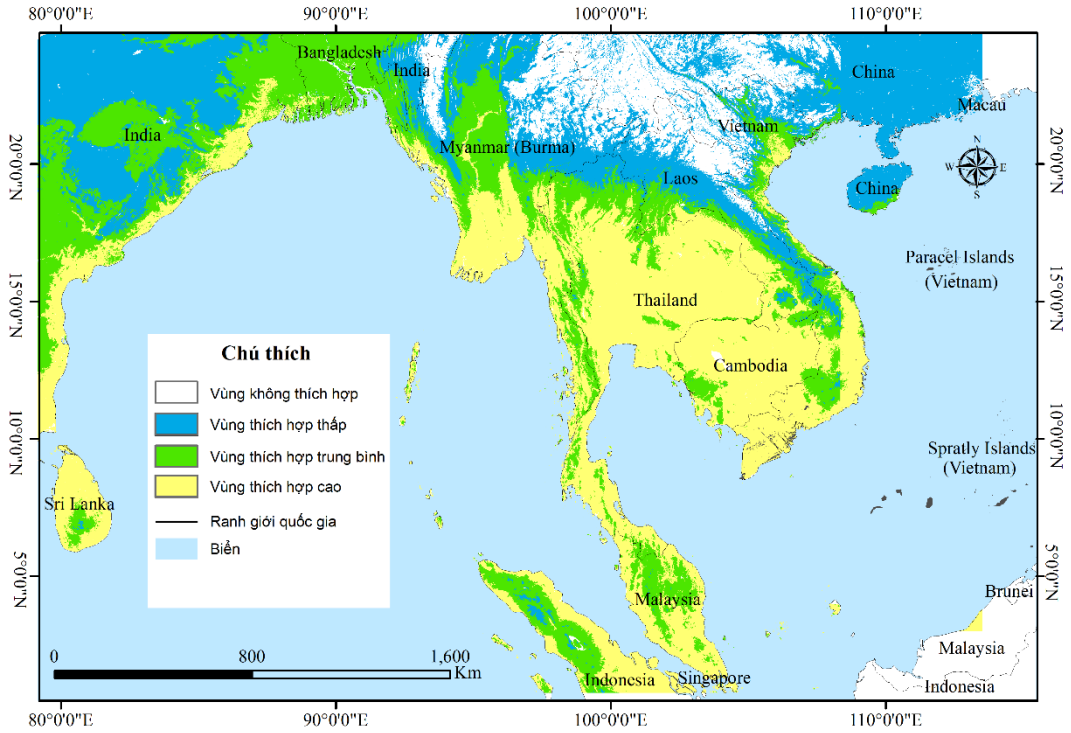
Hình 4. Phân bố tự nhiên của loài Sao đen

3.3. Phân bố tiềm năng của Sao đen

Vùng phân bố địa lý của Sao đen dựa vào các biến môi trường và điểm ghi nhận loài thông qua thuật toán RF được thể hiện ở hình 4. Kết quả hình 4b cho thấy, Sao đen phân bố tự nhiên ở Việt Nam (Nam bộ và Tây Nguyên), Lào, Campuchia, Thái Lan, Myanmar, Malaysia, Indonesia, Bangladesh, Srilanka và Ấn Độ. Theo Foxworthy (1946) Sao đen xuất hiện rộng rãi ở các trạng thái rừng lá rộng thường xanh ở hầu hết các quốc gia Nam Á như Đảo Andaman, Đảo Borneo và quần đảo Philippine [33]. Ở bán đảo Mã Lai, Sao đen được tìm thấy tự nhiên ở Langkawi, Perlis, Kedah và Bắc Perak, Kelantan

và Trengganu [34]. *H.odorata* được tìm thấy tự nhiên ở hầu hết các tỉnh Trung và Nam Lào, từ tỉnh Xayaboury và Viêng Chăn. Sao đen thường mọc thành quần thể với mật độ cao ở rừng nửa rụng lá và rừng thường xanh khô [25].

Kết quả mô phỏng vùng phân bố sinh thái thích hợp cho loài Sao đen bằng thuật toán RF cho thấy diện tích vùng thích hợp thấp và thích hợp trung bình và thích hợp cao theo thứ tự là 859.181 km^2, 988.577 km^2 và 1.227.788 km^2. Vùng sinh thái phù hợp cho loài thuộc 13 quốc gia: Bangladesh, Myanmar, Campuchia, Sri Lanka, Trung Quốc, Indonesia, Ấn độ, Lào, Malaysia, Singapore, Thái Lan và Việt Nam (Hình 5). Trong đó, vùng sinh thái thích hợp cao của loài nhiều nhất ở Campuchia 166.225 km^2, Myanma 160.743,9 km^2, Ấn độ 123.717,5 km^2. Tại Việt Nam, vùng phân bố sinh thái thích hợp cao của loài khá rộng với diện tích là 94.469,8 km^2 thuộc khu vực Nam bộ, Tây Nguyên và rải rác ở ven biển các tỉnh từ Ninh Bình trở vào. Sao đen là loài có biên độ sinh thái rộng, ở nước ta Sao đen được người Pháp di thực từ miền Nam trồng ở các tỉnh miền Bắc từ những năm đầu của thế kỷ 20 [25].



Hình 5. Phân vùng môi trường sống phù hợp của loài Sao đen

4. KẾT LUẬN

Kết quả kiểm tra mức độ phù hợp của 4 thuật toán máy học (rừng ngẫu nhiên, vecto hỗ trợ, độ dốc tăng cường cấp cao và CART) cho thấy rừng ngẫu nhiên có độ chính xác cao nhất trong mô phỏng phân bố của Sao đen.

Sự ảnh hưởng của các yếu tố môi trường đến phân bố địa lý của loài được sắp xếp theo mức độ đóng góp của các biến đến kết quả mô hình từ cao đến thấp: nhiệt độ mùa (bio4), hàm lượng đạm (nitrogen), trữ lượng carbon hữu cơ (ocs), mảnh thô lẫn trong đất (cfvo), dung trọng (bdod), cuối cùng là các tham số còn lại.

Thông qua mô hình rừng ngẫu nhiên nghiên cứu đã xác định vùng sinh thái phù hợp cho loài Sao đen khoảng 1.227.788 km^2 thuộc 13 quốc gia: Bangladesh, Myanmar, Campuchia, Sri Lanka, Trung Quốc, Indonesia, Ấn độ, Lào, Malaysia, Singapore, Thái Lan và Việt Nam. Ở nước ta, vùng phân bố địa lý tự nhiên của loài thuộc khu vực Tây Nguyên và Nam bộ. Trong khi đó, vùng sinh thái thích hợp để bảo tồn và gây trồng loài khá rộng, từ Nam bộ, Tây Nguyên và rải rác ở ven biển các tỉnh từ Ninh Bình trở vào phía Nam.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Phân hiệu

trường Đại học Lâm nghiệp tại tỉnh Đồng Nai trong Đề tài nghiên cứu khoa học cấp cơ sở mã số 910-QĐ-PHĐHLN: "Ứng dụng mô hình Niche xây dựng bản đồ phân bố loài Vù hương (*Cinnamomum balanse* Lecomte) tại Việt Nam".

TÀI LIỆU THAM KHẢO

[1]. Crego RD, Stabach JA & Connette G (2022). Implementation of species distribution models in Google Earth Engine. Diversity and Distributions. 28(5): 904-16; DOI: https://doi.org/10.1111/ddi.13491.

[2]. Guisan A, Thuiller W & Zimmermann NE (2017). Habitat suitability and distribution models: with applications in R. Cambridge University Press. DOI: https://doi.org/10.1017/9781139028271.

[3]. Pecchi M, Marchi M, Burton V & Giannetti F, Moriondo M, Bernetti I, Bindi M, Chirici G (2019). Species distribution modelling to support forest management. A literature review. Ecological Modelling. 411(1): 1-12. DOI: https://doi.org/10.1016/j.ecolmodel.2019.108817.

[4]. Zimmermann NE, Edwards Jr TC, Graham CH, Pearman PB & Svenning JC (2010). New trends in species distribution modelling. Ecography. 33(6):985-9. DOI: https://doi.org/10.1111/j.1600-0587.2010.06953.x.

[5]. Araújo MB, Anderson RP, Márcia Barbosa A, Beale CM, Dormann CF, Early R, Garcia RA, Guisan A, Maiorano L, Naimi B & O'Hara RB (2019). Standards for distribution models in biodiversity assessments. Science Advances. 5(1):eaat4858. DOI: https:// doi/10.1126/sciadv.aat4858.

[6]. Kass JM, Vilela B, Aiello-Lammens ME, Muscarella R, Merow C & Anderson RP (2018). Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods in Ecology and Evolution. 9(4):1151-6. DOI: https://doi.org/10.1111/2041-210X.12945.

[7]. Phillips SJ, Anderson RP & Schapire RE (2006). Maximum entropy modeling of species geographic distributions. Ecological modelling. 190(3-4):231-59 . DOI: https://doi.org/10.1016/j.ecolmodel.2005.03.026.

[8]. Thuiller W, Lafourcade B, Engler R & Araújo MB (2009). BIOMOD–a platform for ensemble forecasting of species distributions. Ecography. 32(3):369-73. DOI: https://doi.org/10.1111/j.1600-0587.2008.05742.x.

[9]. Oeser J, Heurich M, Senf C, Pflugmacher D, Belotti E & Kuemmerle T (2020). Habitat metrics based on multi-temporal Landsat imagery for mapping large mammal habitat. Remote Sensing in Ecology and Conservation. 6(1):52-69. DOI: https://doi.org/10.1002/rse2.122.

[10]. Pettorelli N, Laurance WF, O'Brien TG, Wegmann M, Nagendra H & Turner W (2014). Satellite remote sensing for applied ecologists: opportunities and challenges. Journal of Applied Ecology. 51(4):839-48. DOI: https://doi.org/10.1111/1365-2664.12261.

[11]. Xiong J, Thenkabail PS, Gumma MK, Teluguntla P, Poehnelt J, Congalton RG, Yadav K & Thau D (2017). Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. ISPRS Journal of Photogrammetry and Remote Sensing. 126:225-44. DOI: https://doi.org/10.1016/j.isprsjprs.2017.01.019.

[12]. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D & Moore R (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment. 202:18-27. DOI: https://doi.org/10.1016/j.rse.2017.06.031.

[13]. Tamiminia H, Salehi B, Mahdianpari M, Quackenbush L, Adeli S & Brisco B (2020). Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS Journal of Photogrammetry and Remote Sensing. 164:152-70. DOI: https://doi.org/10.1016/j.isprsjprs.2020.04.001.

[14]. Hossain MM, Kabir MS, Chowdhury TA, Hasanat A & Chakrabarty N (2015). Anthelmintic effects of different extracts of Hopea odorata leaves on Tubifex tubifex worm using in vitro method and their condensed tannin content. Br J Pharm Res. 8(3):1-7. DOI: https://doi.org/10.9734/BJPR/2015/19064.

[15]. Ho TK. Random decision forests (1995). InProceedings of 3rd international conference on document analysis and recognition. IEEE. (1): 278-282 . DOI: https://doi.org/10.1109/ICDAR.1995.598994.

[16]. Nguyen TT, Mac Duy H, Duong TN & Nghiem TD (2020). Forecast of Hourly Tropospheric Ozone Concentration in Quang Ninh using MLP and SVM. VNU Journal of Science: Earth and Environmental Sciences. 36(3): 46-54. DOI: https://doi.org/10.25073/2588-1094/vnuees.4604.

[17]. Chen G & Hay GJ (2011). A support vector regression approach to estimate forest biophysical parameters at the object level using airborne lidar transects and quickbird data. Photogrammetric Engineering & Remote Sensing. 77(7):733-41. DOI: https://doi.org/10.14358/PERS.77.7.733.

[18]. Friedman JH (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics. (1): 1189-232. DOI: https://doi.org/10.1214/aos/1013203451.

[19]. Đặng Thanh Tùng, Nguyễn Thanh Tùng, Hoàng Thị Thủy, Tăng Thị Thanh Nhàn, Đặng Thu Hằng, Võ Ngọc Hải & Nguyễn Dũng Dương (2021). Khai thác trực tuyến cơ sở dữ liệu ảnh vệ tinh, so sánh thuật toán học máy về phân loại lớp phủ trên nền google earth engine. Kỷ yếu hội thảo khoa học quốc gia giải pháp kết nối và chia sẻ hệ thống cơ sở dữ liệu phục vụ công tác đào tạo, quản lý lĩnh vực tài nguyên và môi trường. 1-11.

[20]. Fielding AH & Bell JF (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental conservation. 24(1):38-49. DOI: https://doi.org/10.1017/S0376892997000088.

[21]. Sofaer HR, Hoeting JA & Jarnevich CS (2019). The area under the precision-recall curve as a

performance metric for rare binary events. Methods in Ecology and Evolution. 10(4):565-77. DOI: https://doi.org/10.1111/2041-210X.13140.

[22[. Zhang K, Yao L, Meng J & Tao J (2018). Maxent modeling for predicting the potential geographical distribution of two peony species under climate change. Science of the Total Environment. 634:1326-34. DOI: https://doi.org/10.1016/j.scitotenv.2018.04.112.

[23]. Lorena AC, Jacintho LF, Siqueira MF, De Giovanni R, Lohmann LG, De Carvalho AC & Yamamoto M (2011). Comparing machine learning classifiers in potential distribution modelling. Expert Systems with Applications. 38(5):5268-75. DOI: https://doi.org/10.1016/j.eswa.2010.10.031.

[24]. Breiman L (2001). Random forests. Machine learning. 45(1):5-32. DOI: https://doi.org/10.1023/A:1010933404324.

[25]. Phothisat S (1998). Some Ecological Studies on *Hopea Odorata* Roxb. *and Hopea Pubescens* Ridl. Seedlings. Doctoral dissertation. Universiti Putra Malaysia.

[26]. Zhu H, Yi X, Li Y, Duan Y, Wang X & Zhang L (2021). Limiting climatic factors in shaping the distribution pattern and niche differentiation of Prunus dielsiana in subtropical China. Journal of Forestry Research. 32(4):1467-77. DOI: https://doi.org/10.1007/s11676-020-01194-8.

[27]. Černý T, Doležal J, Janeček Š, Šrůtek M, Valachovič M, Petřík P, Altman J, Bartoš M & Song JS (2013). Environmental correlates of plant diversity in Korean temperate forests. Acta Oecologica. 47:37-45. DOI: https://doi.org/10.1016/j.actao.2012.12.001.

[28]. Kooch Y, Samadzadeh B & Hosseini SM (2017). The effects of broad-leaved tree species on litter quality and soil properties in a plain forest stand. Catena. 150:223-9.
DOI: https://doi.org/10.1016/j.catena.2016.11.023

[29]. Wan JZ, Yu JH, Yin GJ, Song ZM, Wei DX & Wang CJ (2019). Effects of soil properties on the spatial distribution of forest vegetation across China. Global Ecology and Conservation. 18:e00635. DOI: https://doi.org/10.1016/j.gecco.2019.e00635

[30]. Sollins P (1998). Factors influencing species composition in tropical lowland rain forest: does soil matter?. Ecology. 79(1):23-30. DOI: https://doi.org/10.1890/0012-9658(1998)079[0023:FISCIT]2.0.CO;2

[31]. Saha SK, Nair PK, Nair VD & Kumar BM (2009). Soil carbon stock in relation to plant diversity of homegardens in Kerala, India. Agroforestry systems. 76(1):53-65. DOI: https://doi.org/10.1007/s10457-009-9228-8.

[32]. Zhang, Keliang (2018). Maxent modeling for predicting the potential geographical distribution of two peony species under climate change. Science of the Total Environment, 2018, 634: 1326-1334.

[33]. Foxworthy FW (1946). Distribution of the Dipterocarpaceae. Journal of the Arnold Arboretum. 27(4):347-54. DOI: https://www.jstor.org/stable/43781212

[34]. Symington CF (1941). Foresters' manual of dipterocarps. Malayan forest records. (16):1-244.

APPLIED MACHINE LEARNING ALGORITHMS FOR THE DISTRIBUTION MODEL OF *Hopea odorata* IN GOOGLE EARTH ENGINE

Nguyen Thanh Tuan[1*], Phung Van Phe[2]

[1]*Vietnam National University of Forestry - Dong Nai Campus*

[2]*Vietnam National University of Forestry*

ABSTRACT

Hopea odorata plays an important role both ecologically and economically in Vietnam natural forest. Forest degradation has threatened the survival of *H. odorata* in the natural world. For these reasons, it is necessary to devise strategies for effective conservation of this valuable species. This study analyzed potentially suitable areas for *H. odorata* using four machine learning algorithms (random forest-RF, gradient extreme boot-Boosted, support vector machine-SVM, classification and regression tree-CART). Modeling included 117 occurrence records along with 19 climate-related variables, soils, and topographies. The results indicated that the habitat suitability region for *H. odorata* covers approximately 1.227.788 km^2, and these locations were concentrated in 13 countries such as Bangladesh, Myanmar, Cambodia, Sri Lanka, China, Indonesia, India, Laos, Malaysia, Singapore, Thailand and Vietnam. Temperature seasonality, nitrogen, soil organic carbon, coarse fragments and bulk density were the key contributors to *H. odorata* distribution. Implementing the random forest algorithm in Google Earth Engine has shown outstanding advantages in obtaining habitat suitability maps for the restoration and conservation of *H.odorata* in the East Asia, South Asia and Southeast Asia region.

Keywords: bioclimatic variables, Dipterocarp, habitat suitability, random forest, soil properties.

Ngày nhận bài : 03/03/2023

Ngày phản biện : 06/04/2023

Ngày quyết định đăng : 25/04/2023