TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Dự đoán khả năng sạt lở đất ở Việt Nam bằng các thuật toán học máy

Phạm Trọng Huynh[1]*

[1] Trường Đại học Tài nguyên và Môi trường thành phố Hồ Chí Minh;
pthuynh@hcmunre.edu.vn

*Tác giả liên hệ: pthuynh@hcmunre.edu.vn; Tel.: +84–977003834

Ban Biên tập nhận bài: 5/6/2023; Ngày phản biện xong: 12/7/2023; Ngày đăng bài: 25/7/2023

Tóm tắt: Việt Nam là quốc gia có địa hình đồi núi dốc và nằm trong vùng mưa nhiệt đới gió mùa, vì vậy hiện tượng sạt lở đất diễn ra khá phổ biến. Nghiên cứu này tập trung vào việc dự đoán khả năng sạt lở đất ở Việt Nam bằng các thuật toán hồi quy, *Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor regression (KNN), Linear Support Vector Regressor (SVR),* và *Linear Regression (LR)*. Các biến đặc trưng có liên quan đến sạt lở đất được sử dụng, bao gồm độ ẩm đất, địa chấn động đất, lượng mưa, độ cao và độ dốc. Các thuật toán được huấn luyện trên tập dữ liệu mẫu để đánh giá hiệu suất của chúng. Kết quả nghiên cứu cho thấy thuật toán Random Forest (RF) có thể dự đoán tốt khả năng sạt lở đất. Kết quả dự đoán từ tập huấn luyện và tập kiểm tra, với hệ số xác định R^2 có giá trị cao nhất 0,85, thể hiện khả năng giải thích biến động dữ liệu tốt. Bên cạnh đó các giá trị (MSE) và (RMSE) thấp nhất lần lượt là 150,21 và 12,25. Các thuật toán khác cũng cho kết quả tương đối tốt, nhưng (RF) vượt trội hơn. Điều đó cho thấy cần kết hợp năm thuật toán này lại với nhau để xử lý một lượng lớn các dữ liệu có độ phức tạp cao, nhằm tạo ra một mô hình dự đoán sạt lở đất ở Việt Nam bằng các thuật toán Học máy có tính ổn định, chính xác.

Từ khóa: Sạt lở đất; Học máy; Hồi quy; Rừng ngẫu nhiên; K- láng giềng.

1. Giới thiệu

Thảm họa sạt lở đất là một hiện tượng địa chất tiêu cực phổ biến và gây hủy hoại cao. Sạt lở đất xuất hiện do sự tương tác của nhiều yếu tố tự nhiên như địa chất, khí tượng, thủy văn, động đất, núi lửa,... và các yếu tố hoạt động của con người. Tại Việt Nam, hiện tượng sạt lở đất tại Đồng bằng sông Cửu Long (ĐBSCL) đã trở nên ngày càng nghiêm trọng, với số điểm sạt lở tăng từ dưới 100 điểm lên trên 600 điểm hiện nay. Hậu quả của sạt lở đất không chỉ gây thiệt hại về tài sản, tính mạng người dân, mà còn ảnh hưởng đến kinh tế xã hội, làm phá vỡ sự cân bằng tự nhiên của môi trường. Vì vậy, việc dự báo và đưa ra các biện pháp đề phòng sạt lở đất là một nhiệm vụ cấp bách hiện nay.

Phòng ngừa và kiểm soát các vụ sạt lở đất là một trong những vấn đề quan trọng của công tác phòng chống thiên tai và giảm nhẹ thiệt hại. Hiện nay, đã có nhiều phương pháp nghiên cứu với các bộ tiêu chí đánh giá khác nhau [1–3]. Hệ thống thông tin địa lý (GIS) với khả năng xử lý dữ liệu mạnh, đã được áp dụng rộng rãi vào việc xây dựng bản đồ cảnh báo sạt lở [4]. Bên cạnh đó các mô hình dự đoán độ nhạy của sạt lở đất như phân tích định tính, định lượng và trí tuệ nhân tạo [5–6] cũng được sử dụng. Phân tích định tính phụ thuộc nhiều vào kiến thức và tính chủ quan của các nhà nghiên cứu, dẫn đến sự khác biệt lớn về hiệu quả [7], chẳng hạn như phương pháp phân tích quy trình phân cấp (AHP) và phương pháp trọng số Entropy [8]. Các mô hình thống kê và mô hình kết hợp cũng được áp dụng rộng rãi, bao

gồm mô hình tỷ lệ tần số, mô hình giá trị thông tin và mô hình trọng số bằng chứng [9–11]. Tuy nhiên, các mô hình này chưa đạt được kết quả như mong đợi.

Trong những năm gần đây, các thuật toán Học máy đã được áp dụng để phân tích, trích xuất các đặc trưng quan trọng, hỗ trợ trong quyết định và dự đoán. Học máy đã chứng minh tính hữu ích của mình trong việc giải quyết các nhiệm vụ khó khăn trong nhiều lĩnh vực khác nhau. Trong lĩnh vực phòng chống sạt lở đất, Học máy đã được áp dụng để đưa ra các dự đoán tương đối chính xác với điều kiện phải có dữ liệu tốt. Theo các nghiên cứu gần đây [12–13], các dữ liệu trong lĩnh vực sạt lở đất ngày càng nhiều, nhờ vào sự phát triển của Internet of Things (IoT). Đây là nguồn dữ liệu quý để Học máy có thể phân tích, xử lý và đưa ra các dự đoán chính xác trong lĩnh vực sạt lở đất.

Trên thế giới đã có nhiều công trình nghiên cứu để xác định các yếu tố liên quan đến sạt lở đất. Tác giả [14] đã nghiên cứu tiềm năng của dữ liệu độ ẩm trong đất để phát hiện sạt lở khu vực. Kết quả nghiên cứu cho thấy rằng dữ liệu độ ẩm trong đất cung cấp thông tin đáng kể để phát hiện sạt lở sớm. Tác giả [15] đã xác định ảnh hưởng của sự đô thị hóa đến nguy cơ sạt lở do mưa và nhấn mạnh tầm quan trọng của việc xem xét quá trình đô thị hóa trong việc đánh giá nguy cơ sạt lở. Tác giả [16] đã thiết lập mối quan hệ giữa sạt lở, mưa và độ ẩm đất trước đó. Nghiên cứu cũng phát hiện rằng, ngay cả mưa không mạnh cũng có thể gây ra sạt lở khi độ ẩm đất cao. Tác giả [17] đã nhận thấy sự tăng cường hoạt động sạt lở sau các trận động đất nhỏ, dựa trên các hoạt động được ghi nhận đồng thời cùng khoảng thời gian của các năm trước. Tác giả [18] đã đề xuất rằng mưa, động đất và sử dụng đất là những yếu tố bên ngoài gây ra sự khởi đầu thực sự của sạt lở, trong khi địa chất, quá trình thời tiết, đất và địa hình đóng vai trò quan trọng trong việc tạo ra sự không ổn định của độ dốc. Tác giả [19] đã nhận thấy rằng kích thước của sạt lở đất tăng lên khi góc độ dốc tăng. Sạt lở chủ yếu xảy ra dọc theo các đường và trên các lỗi địa chất. Trước đây, đã có nhiều nghiên cứu được thực hiện để xác định khả năng sạt lở sử dụng các kỹ thuật Học máy. Tác giả [20] nhấn mạnh tầm quan trọng của việc sử dụng dữ liệu địa chất đa dạng và phức tạp để thu được các thông tin quan trọng và hữu ích liên quan đến nguy cơ địa chất thông qua các phương pháp Học máy. Tác giả [21] đã thảo luận về hiệu suất của một số phương pháp dựa trên dữ liệu, và kết quả cho thấy Random forest (RF) là phương pháp có hiệu suất dự đoán tốt nhất. Tác giả [22] đã nghiên cứu về khả năng sạt lở bằng cách sử dụng các mô hình Học máy khác nhau để tạo ra các bản đồ dự đoán khả năng sạt lở, nhằm hỗ trợ ra các quyết định cũng như chính sách. Tác giả [23] đã kết hợp việc sử dụng hình ảnh vệ tinh chất lượng cao để phân tích hình ảnh dựa trên đối tượng nhằm tạo ra các đặc trưng cho mô hình tập hợp Random forest (RF). Tuy nhiên, mức độ chính xác chỉ đạt khoảng 80%, vẫn còn khả năng cải thiện. Tác giả [24] đã sử dụng hình ảnh tương tự và mô hình mạng nơ-ron tích chập với pyramid pooling (FCN-PP) để trích xuất đặc trưng và xác định vị trí sạt lở trong hình ảnh sau thiên tai. Mặc dù phương pháp này có độ chính xác đáng tin cậy lên đến 95%, nhưng không thể áp dụng trực tiếp trong hệ thống cảnh báo sớm để ngăn chặn thảm họa quy mô lớn vì nó chỉ tập trung vào việc phát hiện vùng sạt lở sau khi sự cố đã xảy ra. Tác giả [28] đã sử dụng dữ liệu mưa kết hợp với mức nước ngầm và biến động của nó để xây dựng mô hình dự đoán sạt lở đất, kết quả RMSE là 0,144, cho thấy rằng dữ liệu mưa có mối liên hệ cao với sạt lở và có thể được sử dụng như một yếu tố dự đoán có khả năng dự báo tốt.

Ở Việt Nam qua khảo sát, cho thấy các nghiên cứu về sạt lở đất chủ yếu diễn ra ở các tỉnh miền núi phía Bắc và các tỉnh miền Trung - Tây Nguyên. Về phương pháp nghiên cứu, phân tích thống kê được áp dụng rộng rãi để đánh giá nguy cơ sạt lở đất tại Việt Nam. Bên cạnh đó phương pháp phát hiện cũng được sử dụng trong các nghiên cứu, tuy nhiên phương pháp này không được phổ biến. Hiện nay học sâu (*DL-Deep Learning*) là một kỹ thuật bậc cao của lĩnh vực Trí tuệ nhân tạo, cũng đang được áp dụng trong nghiên cứu về sạt lở đất ở Việt Nam, tuy nhiên việc thu thập cơ sở dữ liệu về sạt lở còn thiếu, chưa đồng bộ dẫn đến thiếu thông tin về vị trí và thời điểm xảy ra sạt lở đất. Phần lớn các nghiên cứu hiện nay đang tập trung vào việc cải tiến mô hình tính để tăng độ chính xác, trong khi yếu tố đầu vào lại

chưa được đánh giá đúng mức. Các nghiên cứu thường so sánh các mô hình cho cùng một khu vực nghiên cứu để tìm ra mô hình phù hợp nhất. Ở đồng bằng sông Cửu Long các nghiên cứu về sạt lở đất sử dụng kỹ thuật Học máy chưa nhiều, chưa phổ biến, chưa mang lại hiệu quả.

Mặc dù đã có nhiều nỗ lực để phát triển hệ thống phát hiện sạt lở thông qua việc áp dụng các kỹ thuật Học máy, tuy nhiên vấn đề này vẫn gặp nhiều thách thức [25]. Một trong những rào cản chính là sự mất cân đối giữa các lớp dữ liệu. Do đó, khía cạnh này cần được giải quyết để tăng hiệu suất của mô hình. Ngoài ra nhiễu dữ liệu cũng là một thách thức đáng chú ý. Do dữ liệu quan sát từ nhiều nguồn cảm biến trên vệ tinh vì vậy thường chứa nhiều nguồn nhiễu và giá trị ngoại lệ. Điều này làm khó khăn cho các mô hình Học máy vì chúng có sự phụ thuộc mạnh vào dữ liệu đầu vào.

Mục tiêu chính của bài báo là đề xuất mô hình nhận dạng khả năng xảy ra sạt lở đất ở Việt Nam, bằng cách áp dụng năm thuật toán Học máy, bao gồm *Random Forest, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor regression (KNN), Linear Support Vector Regressor (SVR)* và *Linear regression*. Việc chọn các thuật toán này nhằm mục đích phủ sóng các phương pháp từ cây quyết định, đến phương pháp tuyến tính. Điều này giúp khám phá các mặt khác nhau của dữ liệu và tối ưu hóa khả năng dự đoán. Để đo lường hiệu suất của các thuật toán, bài báo sử dụng các chỉ số như sai số bình phương trung bình (MSE), sai số trung bình tuyệt đối (MAE) và sai số bình phương trung bình căn (RMSE). Phương pháp được xác thực bằng việc áp dụng bộ dữ liệu mẫu để nghiên cứu thử nghiệm khả năng dự báo chính xác của mô hình.

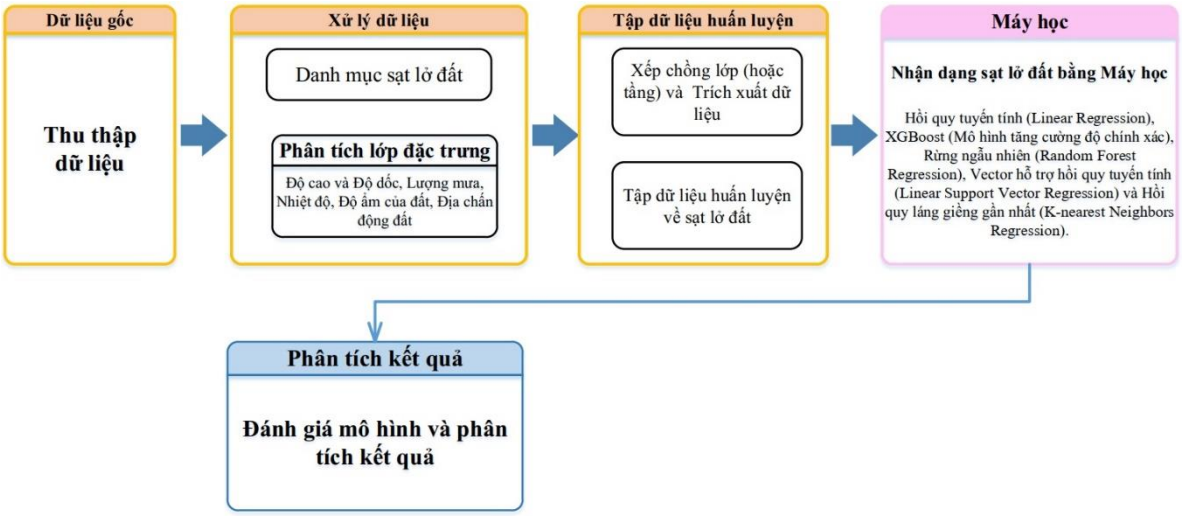
2. Phương pháp nghiên cứu và xử lý dữ liệu

2.1. Phương pháp nghiên cứu

Đồng bằng sông Cửu Long (ĐBSCL), còn được gọi là Đồng bằng sông Mekong, là một khu vực nằm ở hạ lưu sông Mekong trải dài qua lãnh thổ Việt Nam. Với diện tích khoảng 40.816,3 km^2, chiếm khoảng 12,3% diện tích của cả nước. Mỗi năm, dòng chảy sông Mekong mang đến cho vùng ĐBSCL một lượng nước lớn, các hạt phù sa mịn và cát sỏi. Ước tính có khoảng 160 triệu tấn phù sa mịn và 30 triệu tấn cát sỏi. Vùng ĐBSCL có tầm quan trọng đặc biệt trong khu vực và trên toàn cầu, với hệ sinh thái độc đáo và đa dạng sinh học. Ngoài ra, vùng ĐBSCL cũng đóng vai trò quan trọng trong sản xuất nông nghiệp và nuôi trồng thủy sản, đóng góp vào nền kinh tế của Việt Nam. Tuy nhiên, hiện nay đồng bằng sông Cửu Long (ĐBSCL) đang đối mặt với tình trạng sạt lở đất nghiêm trọng. Các điểm sạt lở được thể hiện trong Hình 1. Có rất nhiều nguyên nhân gây ra sạt lở, trong đó có các nguyên nhân chính như sự thay đổi dòng chảy của sông, khai thác cát trái phép và các tác động từ biến đổi khí hậu.

Hình 1. Khu vực nghiên cứu.

Nghiên cứu đề xuất mô hình cùng với năm thuật toán Học máy để dự đoán khả năng xảy ra sạt lở đất và có thể áp dụng ở Việt Nam. Sơ đồ hoạt động của mô hình được mô tả trong Hình 2.



Hình 2. Mô hình đề xuất dự đoán khả năng sạt lở đất bằng các thuật toán Học máy.

Dữ liệu mẫu được sử dụng để huấn luyện cho mô hình bao gồm độ cao và độ dốc (tính theo đơn vị feet), lượng mưa (theo mm/giờ), độ ẩm của đất (theo %), địa chấn động đất (theo joule). Hiện ở Việt Nam dữ liệu này chưa phổ biến rộng rãi, vì vậy bài báo lấy dữ liệu mẫu từ nguồn truy cập mở của Google Earth Pro, GPS Visualizer, Trung tâm Dữ liệu và Dịch vụ Khoa học Trái đất Goddard Earth Sciences, Dữ liệu Độ ẩm Đất Toàn cầu NASA-USDA, Chương trình Mối nguy hiểm động đất USGS, Dự báo Lở đất Toàn cầu, dữ liệu thu thập từ tháng 1 đến tháng 12 năm 2020 để thử nghiệm mô hình.

2.2. Xử lý dữ liệu

2.2.1. Các biến đặc trưng dùng để huấn luyện mô hình

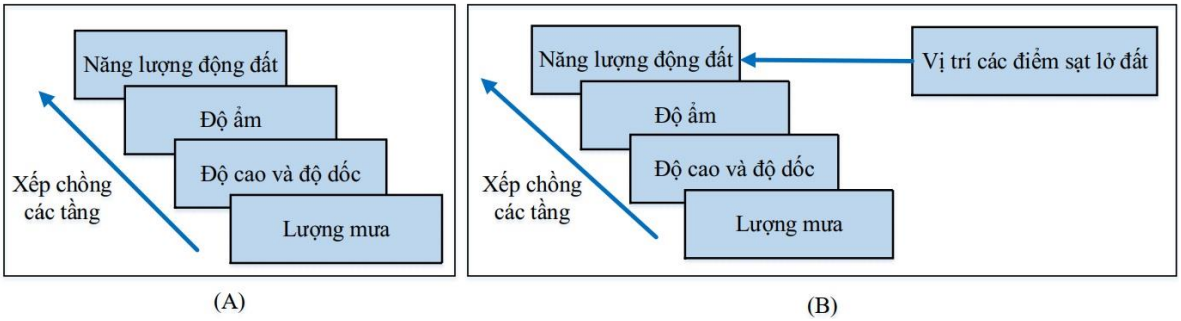
Các biến đặc trưng được lấy từ tập dữ liệu mẫu để dùng huấn luyện. Các biến này bao gồm độ ẩm đất, địa chấn động đất, lượng mưa, độ cao và độ dốc, ngoài ra nghiên cứu đã mở rộng thêm dữ liệu mẫu lưu lượng nước, chế độ phù sa bùn cát, địa chất, chế độ thủy lực, địa hình-hình thái sông, giao thông thủy, xây dựng cơ sở hạ tầng, khai thác cát để thử nghiệm mô hình đã được đề xuất [26].

Bảng 1. Các biến được chọn trong tập dữ liệu mẫu dựa trên tính phù hợp của mô hình

Biến phái sinh	Bộ dữ liệu	Thời gian
Độ cao và Độ dốc (theo feet)	Google Earth Pro, GPS Visualizer	Từ tháng 1 đến tháng 12 năm 2020
Lượng mưa (theo mm/giờ)	Trung tâm Dữ liệu và Dịch vụ Khoa học Trái đất Goddard Earth Sciences	Từ tháng 1 đến tháng 12 năm 2020
Độ ẩm đất (theo mm/giờ)	Dữ liệu Độ ẩm Đất Toàn cầu NASA-USDA	Từ tháng 1 đến tháng 12 năm 2020
Địa chấn động đất (theo joule)	Chương trình Mối nguy hiểm động đất USGS	Từ tháng 1 đến tháng 12 năm 2020
Xác suất lở đất	Dự báo Lở đất Toàn cầu	Từ tháng 1 đến tháng 12 năm 2020

Độ dốc là đo lường độ dốc bề mặt. Sự hình thành, phát triển của sạt lở đều bị ảnh hưởng đáng kể bởi độ dốc. Thông tin này được thu thập từ Google Earth Pro. Địa chấn động đất xác định khả năng của nó trong việc tạo ra sạt lở. Điều kiện độ ẩm đất đóng vai trò quan trọng trong việc khởi đầu sạt lở. Trong nghiên cứu này, dữ liệu độ ẩm đất được trích xuất từ bộ dữ

liệu độ ẩm đất NASA-USDA [26]. Lượng mưa là biến chủ yếu gây ra sạt lở, vì nó ảnh hưởng đến sự ổn định của độ dốc khi thấm qua đất và đá làm cho độ dốc trở nên yếu, không ổn định gây ra sạt lở. Biến này được lấy từ Trung tâm Dữ liệu và Dịch vụ Khoa học Trái đất Goddard. Ngoài ra, lưu lượng nước chảy qua khu vực và có thể gây ảnh hưởng đến tính ổn định của đất. Lưu lượng phù sa là lượng chất thải, bùn đất được kéo theo trong quá trình mưa và có thể tạo ra tác động xói mòn và sạt lở. Lưu lượng đất cát bị mất có khả năng làm tổn thương nền đất tạo ra sạt lở. Các lớp dữ liệu được xếp chồng lên nhau để tạo thành bộ dữ liệu cuối cùng. Quá trình xếp chồng các lớp dữ liệu được thể hiện trong Hình 3.

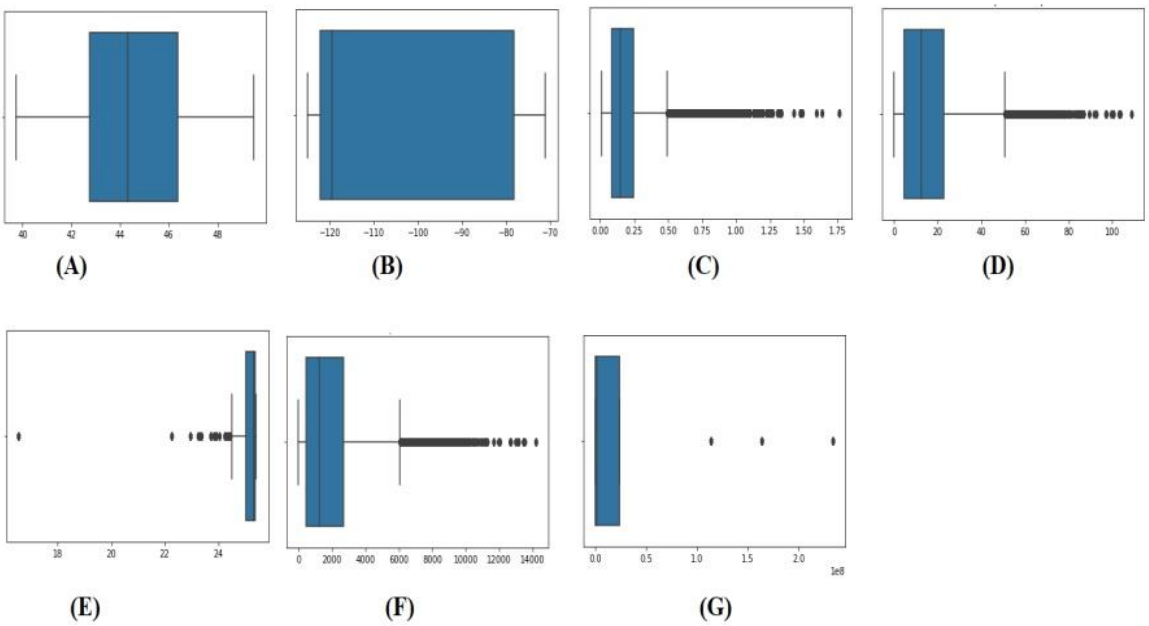


Hình 3. Các tầng dữ liệu xếp chồng: (A) Các tầng trong tập huấn luyện; (B) Các tầng trong tập kiểm tra.

Các giá trị này được trích xuất từ bản đồ khả năng xảy ra sạt lở toàn cầu [27]. Các giá trị xác suất này nằm trong khoảng từ 0 đến 1. Độ cao được đo bằng đơn vị feet, độ ẩm đất và lượng mưa được đo bằng đơn vị mm/giờ và độ dốc được đo bằng đơn vị độ. Xác suất lở đất tính toán trong nghiên cứu này là 0,033.

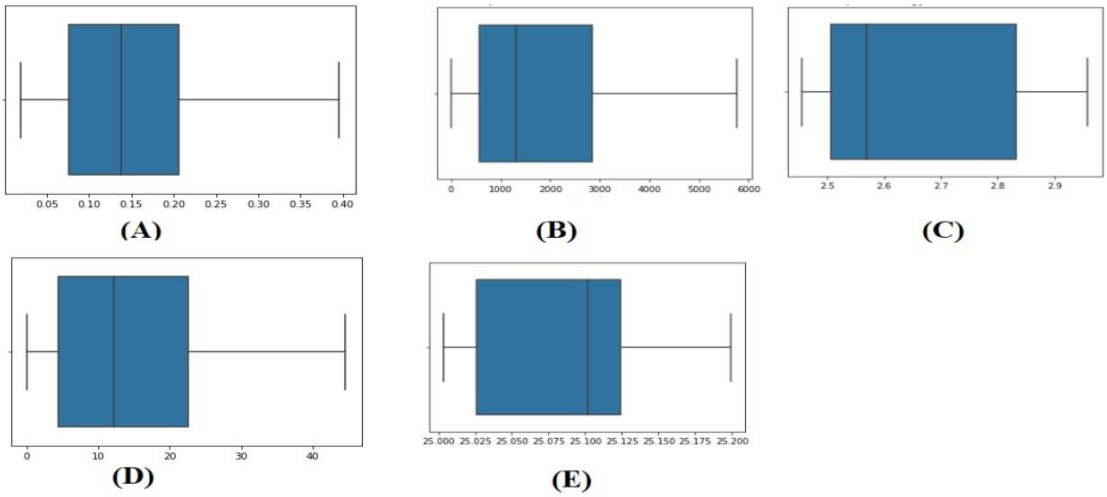
2.2.2. Chuẩn bị dữ liệu

Bước tiếp theo là tạo tập huấn luyện. Việc trích xuất dữ liệu, xếp chồng thành các tầng. Chỉ những biến cần thiết được giữ lại trong quá trình trích xuất dữ liệu. Ví dụ như vĩ độ, kinh độ, ngày và lượng mưa (theo mm/giờ) được giữ lại từ bộ dữ liệu mưa. Sau đó, mỗi tầng dữ liệu thu được được xếp chồng như được hiển thị trong Hình 3. Mỗi tầng trong hình biểu thị một biến dự báo, được kết hợp để để xây dựng một bộ dữ liệu huấn luyện nhiều chiều. Các tầng được xếp chồng lên nhau dựa trên các giá trị.



Hình 4. Với các ngoại lệ: A) Biểu đồ hộp của vĩ độ; B) Biểu đồ hộp của kinh độ; C) Biểu đồ hộp của lượng mưa; D) Biểu đồ hộp của độ dốc; E) Biểu đồ hộp của độ ẩm đất; F) Biểu đồ hộp của độ cao; G) Biểu đồ hộp của địa chấn động đất.

Tiền xử lý dữ liệu: Dữ liệu không cân bằng gây ảnh hưởng đến hiệu suất của các thuật toán hồi quy. Trên thực tế, việc xử lý vấn đề mất cân bằng trong bài toán hồi quy là khó khăn do giá trị mục tiêu là liên tục và có thể có vô số giá trị. Trong nghiên cứu này, giá trị mục tiêu là xác suất sạt lở đất. Vì vậy nghiên cứu đã tạo ra các hạng mục lớp bằng cách phân phối các giá trị xác suất thành ba khoảng khác nhau: Thấp, Trung bình và Cao. Sau đó sử dụng thuật toán SMOGN [28] để đảm bảo rằng các mẫu được tạo ra đáp ứng được yêu cầu. Các giá trị ngoại lệ là các giá trị bất thường có thể làm sai lệch kết quả. Để thực hiện được điều đó bài báo đã kết hợp hai kỹ thuật Winsorization [29] và Boxplot [30], do đó dữ liệu có giá trị null (giá trị bị thiếu) đã được loại bỏ. Trước khi loại bỏ các giá trị ngoại lệ, nghiên cứu đã phân tích các biến đặc trưng được thể hiện trong Hình 4. Sau khi áp dụng phương pháp Winsorization và boxplot để loại bỏ các giá trị ngoại lệ, được mô tả như trong Hình 5.



Hình 5. Không có ngoại lệ: A) Biểu đồ hộp của lượng mưa; B) Biểu đồ hộp của độ cao; C) Biểu đồ hộp địa chấn động đất; D) Biểu đồ hộp của độ dốc; E) Biểu đồ hộp của độ ẩm đất.

2.2.3. Phân chia dữ liệu

Bộ dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 70-30. Giá trị đặc trưng đã được tỷ lệ bằng phương pháp chuẩn hóa tiêu chuẩn [31]. Việc tỷ lệ chuẩn được tiến hành sau khi tập dữ liệu được chia để ngăn chặn việc kiểm tra dữ liệu [30]. Tập huấn luyện có 10.165.554 mẫu, và tập kiểm tra có 1.457.808 mẫu, các dữ liệu trong tập này được lấy từ tập dữ liệu mẫu để kiểm tra mô hình.

2.3. Các thuật toán dùng cho mô hình đề xuất

Sau khi thiết lập tập dữ liệu huấn luyện, bước tiếp theo là huấn luyện các mô hình Học máy và thực hiện dự đoán sạt lở đất. Năm thuật toán Học máy được chọn để đánh giá khả năng dự đoán nguy cơ sạt lở đất.

Thuật toán Hồi quy tuyến tính: Mô hình hồi quy tuyến tính được sử dụng để phân tích mối quan hệ tuyến tính giữa biến phụ thuộc (xác suất sạt lở đất) và các biến độc lập (các biến dự đoán khác). Trong nghiên cứu này, sử dụng hồi quy tuyến tính đa biến vì một biến dự đoán đơn lẻ không đủ để giải thích xác suất lở đất. Phương trình 1 biểu diễn hàm tuyến tính.

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + \cdots + b_n \times X_n \quad (1)$$

Trong đó Y là biến phụ thuộc (xác suất sạt lở đất trong trường hợp này), X_i đại diện cho các biến độc lập (các chỉ báo khác nhau), a là hằng số và b_i là hệ số hồi quy của biến X_i.

Thuật toán Hồi quy Rừng Ngẫu Nhiên - Random Forest Regression (RFR): đây là một thuật toán được sử dụng trong Học máy, bằng cách kết hợp nhiều cây quyết định để đưa ra dự đoán tổng quát. Các cây trong RFR được huấn luyện thông qua việc sử dụng các tập con

được tạo từ tập dữ liệu huấn luyện chính thông qua phương pháp Bootstrap. Trong nhiệm vụ hồi quy, RFR tính toán dự đoán trung bình từ K cây hồi quy, được tính bằng phương trình 2.

$$\mathrm{RFR}_{\mathrm{prediction}} = \frac{1}{K}\sum_{k=1}^{K} h_k(x) \quad (2)$$

Thuật toán Hồi quy XGBoost: *Extreme Gradient Boosting*, hay còn gọi là XGBoost, là một thuật toán Học máy sử dụng một tập hợp cây quyết định để dự đoán. Mỗi cây trong tập hợp được huấn luyện để học hàm quyết định bằng cách giảm thiểu hàm mất mát sử dụng phương pháp Gradient Descent. Mỗi cây nhằm mục tiêu sửa chữa những sai sót của cây trước đó trong quá trình học. Việc huấn luyện tập hợp K cây được mô tả trong phương trình 3.

$$O_{bj} = \sum_{i=1}^{n} \mathrm{loss}(y_i, \hat{y}) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

Dự đoán cuối cùng của tập hợp được tính toán bằng phương trình 4.

$$Y_{xgb}(x) = \sum_{k=1}^{K} \mathrm{tree}_k(x), \mathrm{tree}_k \in T \quad (4)$$

Thuật toán Hồi quy Vector Hỗ trợ tuyến tính (SVR): là một thuật toán Học máy được sử dụng để dự đoán một giá trị số. Nó hoạt động trên cả đầu vào có giá trị rời rạc và giá trị liên tục. Trong nghiên cứu này đã sử dụng phương pháp hồi quy đa biến để khai thác các biến đầu vào. SVR cố gắng tìm ra một hàm tuyến tính sao cho khoảng cách giữa các dữ liệu đầu vào và đường hồi quy là nhỏ nhất. Phương trình 5 được sử dụng để tính toán đầu ra của SVR.

$$Y_{srv}(x) = \sum_{i=1}^{n} \beta_i K(x; x_i) + b \quad (5)$$

Ở đây β_i và x_i lần lượt là trọng số và vị trí của mỗi SVs. Ngoài ra, n là số lượng SVs, b là sai số, và $K(x; x_i)$ là hàm Kernel tương ứng với x_i.

Thuật toán Hồi quy K láng giềng gần nhất -K-Nearest Neighbors (KNN): là một phân loại của các thuật toán gom cụm mà mục tiêu là nhóm các mẫu có các giá trị đặc trưng tương tự vào các "khu vực láng giềng" để tìm mối tương quan giữa các đặc trưng và giá trị nhãn. Khoảng cách giữa mỗi mẫu được quyết định dựa trên khoảng cách Euclidean của các đặc trưng.

$$Y_{KNN} = \frac{\sum_{i=1}^{K} N_i}{K}; N = X \text{ được sắp xếp theo khoảng cách Euclidean} \quad (6)$$

Để thực hiện dự đoán, thuật toán KNN sẽ tìm K điểm gần nhất với giá trị tham số đầu vào và đưa ra trung bình của nhãn của chúng như được thể hiện trong phương trình 6.

2.4. Phương pháp lựa chọn đặc trưng

Các phương pháp chọn đặc trưng được áp dụng để loại bỏ những đặc trưng không quan trọng. Trong trường hợp này, chú trọng đến những đặc trưng góp phần quan trọng nhất vào biến mục tiêu. Điều này nhằm tiết kiệm chi phí trong quá trình mô hình hóa và cải thiện hiệu suất của mô hình. Nghiên cứu này đã sử dụng phương pháp SelectKBest kết hợp với lựa chọn đặc trưng theo hệ số tương quan [32] và thông tin chung (*Mutual Information*) [33] để tìm ra những đặc trưng tốt nhất từ bộ dữ liệu. Hàm SelectKBest sử dụng các phương pháp này như một hàm tính điểm để xác định mức độ tương quan giữa mỗi đặc trưng và biến mục tiêu. Ở đây bài báo quy định nếu điểm số thấp cho thấy đặc trưng đó không phụ thuộc vào biến mục tiêu. Ngược lại, giá trị điểm số cao thì đặc trưng đó có liên quan đến biến mục tiêu.

2.5. Phương pháp đánh giá hiệu suất của mô hình

Trong nghiên cứu này, sử dụng các độ đo thống kê tiêu chuẩn để đánh giá độ chính xác của mô hình, bao gồm: Sai số trung bình bình phương (MSE), sai số trung bình bình phương căn (RMSE) và sai số trung bình tuyệt đối (MAE). RMSE được tính bằng công thức 7, MSE thể hiện trong công thức 8 và MAE được tính bằng công thức 9.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i-1}^{n}(Y_i - Yi)^2}{n}} \quad (7)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (8)$$

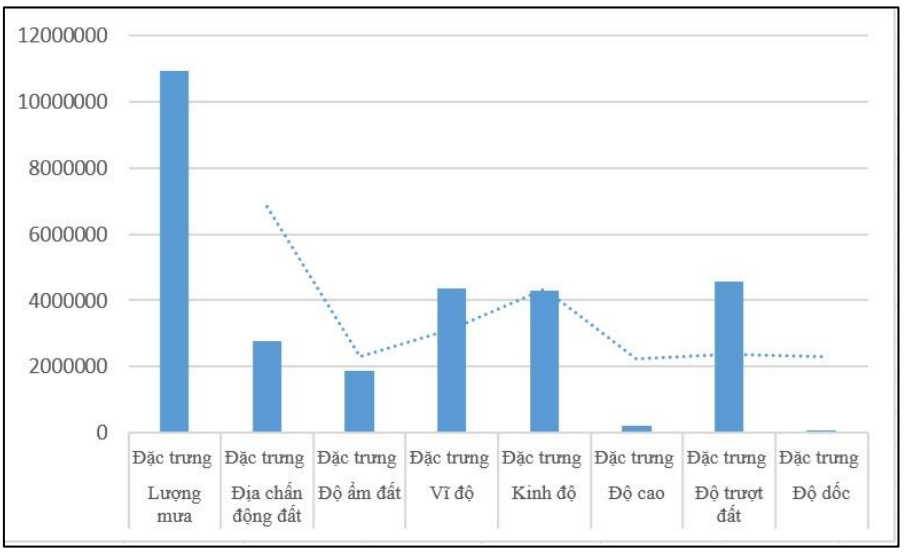
$$MAE = \frac{1}{n}\sum_{i=0}^{n-1}(|Y_i - \hat{Y}_i|) \quad (9)$$

Trong đó Y_i biểu thị giá trị kỳ vọng và $\hat{Y}_i$ là giá trị dự đoán.

3. Kết quả và thảo luận

3.1. Phương pháp thực hiện

Để tìm ra các tham số tối ưu cho mô hình, bài báo đã sử dụng phương pháp GridSearchCV [34]. Quá trình này bao gồm việc xác định một lưới các giá trị tham số khác nhau để thử nghiệm. Mô hình được đào tạo và đánh giá sử dụng từng cấu hình tham số trong lưới. Mục tiêu của quá trình là tìm ra cấu hình tham số tối ưu mang lại hiệu suất tốt nhất trên dữ liệu đánh giá. GridSearchCV là một phương pháp tìm kiếm thông qua lưới các giá trị tham số để tìm ra giá trị tối ưu. Bước này giúp thử nghiệm các cấu hình tham số khác nhau của mô hình và đánh giá hiệu suất của từng cấu hình trên dữ liệu đánh giá. Việc lựa chọn ba lượt chia dữ liệu sau khi phân tích thử nghiệm nhằm đảm bảo tính chính xác và đáng tin cậy của kết quả. Sau khi có các tham số tối ưu cho mô hình, nghiên cứu tiến hành đánh giá tác động của việc lựa chọn các biến đặc trưng đến hiệu suất của các thuật toán máy học trong mô hình. Bằng cách so sánh hiệu suất giữa việc sử dụng và không sử dụng các biến đặc trưng, qua đó có thể đánh giá mức độ ảnh hưởng của từng biến đặc trưng đến kết quả dự đoán. Điều này giúp xác định độ quan trọng của các biến đặc trưng và có thể giúp tối ưu hóa mô hình và cải thiện hiệu suất dự đoán. Kết quả của quá trình này mang lại một mô hình tối ưu với các tham số được điều chỉnh và cung cấp hiểu biết về tác động của các biến đặc trưng đến hiệu suất của mô hình. Việc này đóng góp vào quá trình xây dựng mô hình có hiệu suất cao trong việc dự đoán. Kết quả thu được bằng cách sử dụng phương pháp chọn đặc trưng dựa vào thông tin chung, được mô tả trong Hình 6.



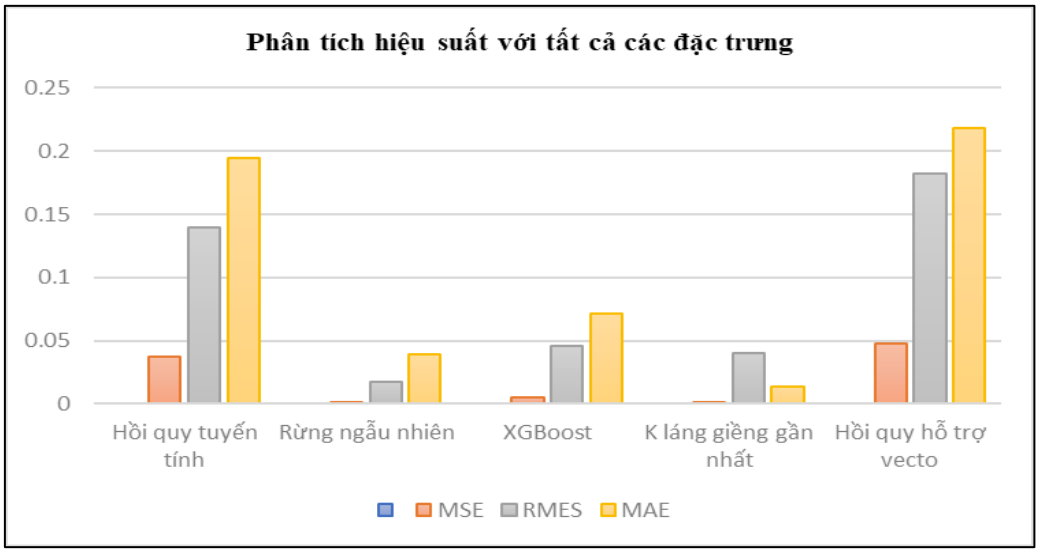
Hình 6. Giá trị điểm số của các yếu tố được thu được bằng phương pháp lựa chọn dựa trên thông tin chung.

3.2. Đánh giá hiệu suất của các thuật toán Học máy sử dụng trong mô hình được đề xuất

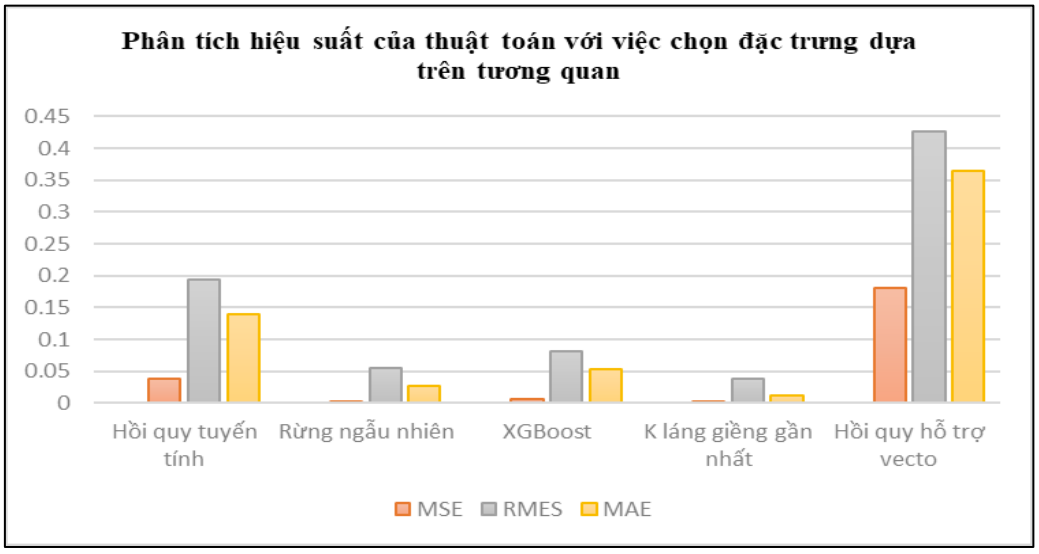
Để đánh giá hiệu suất của các thuật toán Học máy sử dụng trong mô hình đề xuất, bài báo đã áp dụng phương pháp lựa chọn đặc trưng dựa trên tương quan và phương pháp lựa chọn đặc trưng dựa trên thông tin chung. Quá trình lựa chọn đặc trưng này dựa trên việc xác định mức độ tương quan giữa các đặc trưng và mục tiêu dự đoán, cũng như độ quan trọng

của các đặc trưng trong mô hình. Trong nghiên cứu này, đã loại bỏ các đặc trưng có điểm số tương quan thấp nhất. Các mô hình được chạy với cả tập dữ liệu chứa tất cả các đặc trưng và tập dữ liệu chỉ chứa các đặc trưng đã được lựa chọn. Điều này nhằm đánh giá tác động của việc lựa chọn đặc trưng đến khả năng dự đoán sạt lở đất.

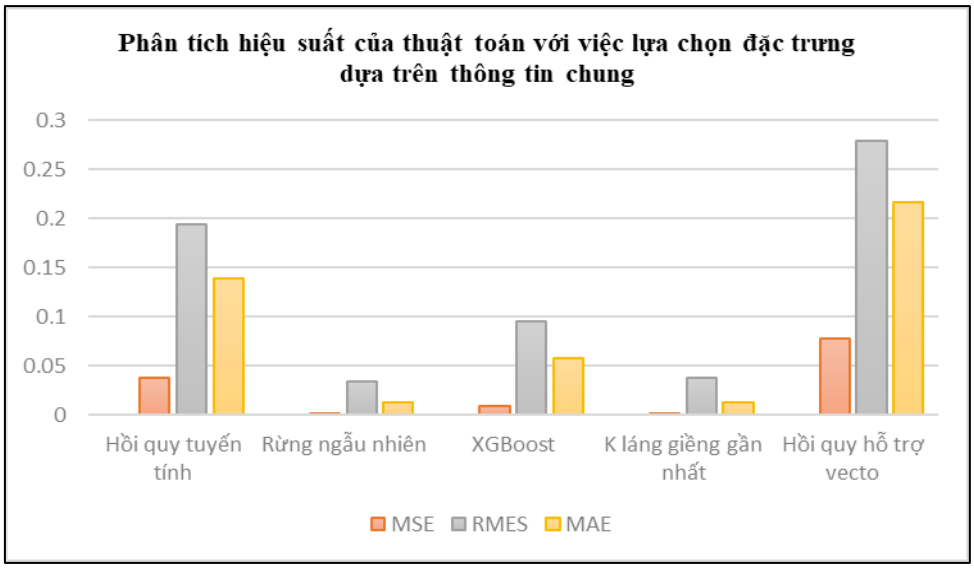
Trong quá trình đánh giá hiệu suất, bài báo đã sử dụng năm thuật toán Học máy khác nhau. Thuật toán 1 đại diện cho Hồi quy tuyến tính - Linear Regression (LR), thuật toán 2 đại diện cho thuật toán Rừng ngẫu nhiên - Random Forest (RF), thuật toán 3 biểu thị cho thuật toán XGBoost, thuật toán 4 đại diện cho thuật toán K-láng giềng gần nhất-KNN Regression và thuật toán 5 đại diện cho thuật toán Hồi quy Vector hỗ trợ - Linear SVR. Hiệu suất của các thuật toán Học máy đã được đánh giá trên tập dữ liệu mẫu được sử dụng để kiểm tra mô hình. Các kết quả về hiệu suất của các thuật toán được mô tả và trình bày trong các hình từ Hình 7 đến Hình 9. Thông qua việc so sánh kết quả, bài báo có thể xác định hiệu suất của từng thuật toán trong việc dự đoán khả năng sạt lở đất và đánh giá đóng góp của việc lựa chọn đặc trưng đối với hiệu suất của mô hình. Việc này nhằm mục đích có được cái nhìn tổng quan về hiệu suất của các thuật toán Học máy và tác động của việc lựa chọn đặc trưng đến kết quả dự đoán. Các kết quả này cung cấp thông tin quan trọng và là cơ sở để cải thiện mô hình và tối ưu hóa hiệu suất trong tương lai.



Hình 7. Biểu đồ mô tả giá trị lỗi đạt được bởi các thuật toán sử dụng tất cả các đặc trưng.



Hình 8. Biểu đồ mô tả giá trị lỗi đạt được bởi các thuật toán sau khi áp dụng phương pháp lựa chọn đặc trưng dựa trên tương quan.



Hình 9. Biểu đồ mô tả giá trị lỗi đạt được bởi các thuật toán sau khi áp dụng phương pháp lựa chọn đặc trưng dựa trên thông tin chung.

Sau khi phân tích các chỉ số, có thể khẳng định rằng thuật toán Rừng ngẫu nhiên -Random Forest (RF) vượt trội hơn các thuật toán khác khi sử dụng phương pháp lựa chọn đặc trưng dựa trên thông tin chung. Phương pháp lựa chọn đặc trưng dựa trên thông tin chung là phương pháp tập trung vào đánh giá mức độ quan trọng của các đặc trưng, sử dụng khái niệm "thông tin chung" để đo lường mức độ liên kết giữa các đặc trưng và mục tiêu. Các đặc trưng có độ thông tin chung cao được coi là quan trọng và được chọn để xây dựng mô hình. Ngược lại thuật toán Hồi quy K láng giềng gần nhất (KNN) lại có hiệu suất tốt khi sử dụng phương pháp lựa chọn đặc trưng dựa trên tương quan. Phương pháp lựa chọn đặc trưng dựa trên tương quan là quá trình xác định độ tương quan giữa các đặc trưng và mục tiêu dự đoán. Các đặc trưng có độ tương quan cao được coi là quan trọng và được chọn để xây dựng mô hình. Điều đó cho thấy chúng ta cần kết hợp các thuật toán Học máy này lại với nhau để đạt được kết quả dự đoán tốt nhất.

4. Kết luận

Xác định và dự báo khả năng sạt lở đất ở nước ta đóng một vai trò then chốt trong việc đánh giá rủi ro. Việc sử dụng mô hình Máy học để dự báo khả năng sạt lở đất tại các vị trí cơ sở hạ tầng quan trọng có thể hỗ trợ trong việc giám sát và giảm thiểu nguy cơ. Nghiên cứu này đã xây dựng một mô hình áp dụng năm thuật toán hồi quy phổ biến của Học máy gồm: Rừng ngẫu nhiên, Hồi quy tuyến tính, Hồi quy XGBoost, Vector Hỗ trợ tuyến tính, K-Láng giềng gần nhất để dự báo khả năng lở đất dựa trên các biến đặc trưng của tập dữ liệu mẫu. Sau khi phân tích nghiên cứu đã chia tỷ lệ 70/30 cho tập huấn luyện và tập kiểm tra. Mục đích chính là đánh giá hiệu suất của các thuật toán trong hai trường hợp: sử dụng phương pháp lựa chọn đặc trưng dựa trên tương quan và phương pháp lựa chọn đặc trưng dựa trên thông tin chung. Kết quả cho thấy hiệu suất của tất cả các thuật toán Học máy đều đạt được các điểm MSE hợp lý. Trong đó thuật toán (RF) đã vượt trội hơn, khi sử dụng phương pháp lựa chọn đặc trưng dựa trên thông tin chung. Thuật toán K-Láng giềng gần nhất lại cho thất hiệu suất tốt hơn, khi sử dụng phương pháp lựa chọn đặc trưng dựa trên tương quan. Vì vậy để xây dựng được một mô hình dự đoán ổn định, có độ tin cây cao cần phải kết hợp năm thuật toán này lại với nhau.

Nghiên cứu này đã chứng minh tiềm năng của việc kết hợp các thuật toán Học máy trong dự báo khả năng sạt lở đất. Hướng phát triển trong tương lai sẽ tiếp tục tăng kích thước và đa dạng hóa bộ dữ liệu nhằm tăng cường độ chính xác của các thuật toán. Xây dựng mô hình

Máy học dự đoán sạt lở đất ở đồng bằng sông Cửu Long với bộ dữ liệu của khu vực này. Nâng cao tính khả diễn giải của các thuật Học máy, để hiểu rõ hơn về các yếu tố quan trọng trong dự đoán khả năng sạt lở đất. Đồng thời, việc phát triển một bảng điều khiển thông tin để hiển thị các dự báo cho từng cơ sở hạ tầng quan trọng cũng là một hướng phát triển trong tương lai.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: P.T.H.; Xử lý số liệu, Viết bản thảo bài báo, Chỉnh sửa bài báo: P.T.H.

Lời cảm ơn: Bài báo hoàn thành nhờ vào kết quả của việc: "Nghiên cứu đánh giá các thuật toán Học Máy và mô hình phân tầng dữ liệu dựa vào các biến đặc trưng để đưa ra dự đoán sạt lở đất". Xin cảm ơn các tác giả đã có những công trình nghiên cứu liên quan mà tôi đã tham khảo.

Lời cam đoan: Chúng tôi xin cam đoan bài báo này là công trình nghiên cứu của chúng tôi, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong bài báo này.

Tài liệu tham khảo

1. Fadhillah, M.F.; Hakim, W.L.; Panahi, M.; Rezaie, M.; Lee, C.W.; Lee, S. Mapping of landslide potential in Pyeongchang-gun, South Korea, using machine learning meta-based optimization algorithms. *Egypt. J. Remote Sens. Space Sci.* **2022**, *25*, 463–472. https://doi.org/10.1016/j.ejrs.2022.03.008.
2. Liu, Y.; Xu, C.; Huang, B.; Ren, X.; Liu, C.; Hu, B.; Chen, Z. Landslide displacement prediction based on multi-source data fusion and sensitivity states. *Eng. Geol.* **2020**, *271*, 105608. https://doi.org/10.1016/j.enggeo.2020.105608.
3. Wang, J.; Xiao, L.; Ward, S.N. Tsunami Squares modeling of landslide tsunami generation considering the 'Push Ahead' effects in slide/water interactions: Theory, experimental validation, and sensitivity analyses. *Eng. Geol.* **2021**, *288*, 106141. https://doi.org/10.1016/j.enggeo.2021.106141.
4. Bragagnolo, L.; da Silva, R.V.; Grzybowski, J.M.V. Landslide susceptibility mapping with landslide: A free open-source GIS-integrated tool based on Artificial Neural Networks. *Environ. Model. Softw.* **2020**, *123*, 104565. https://doi.org/10.1016/j.envsoft.2019.104565.
5. Huang, F.; Cao, Z.; Guo, J.; Jiang, S. H.; Li, S.; Guo, Z. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. *Catena* **2020**, *191*, 104580. https://doi.org/10.1016/j.catena.2020.104580.
6. Małka, A. Landslide susceptibility mapping of Gdynia using geographic information system-based statistical models. *Nat. Hazards* **2021**, *107*, 639–674. https://doi.org/10.1007/s11069-021-04599-8.
7. Das, S.; Sarkar, S.; Kanungo, D.P. GIS-based landslide susceptibility zonation mapping using the analytic hierarchy process (AHP) method in parts of Kalimpong Region of Darjeeling Himalaya. *Environ. Monit. Assess.* **2022**, *194*, 234. https://doi.org/10.1007/s10661-022-09851-7.
8. Ba, Q.; Chen, Y.; Deng, S.; Wu, Q.; Yang, J.; Zhang, J. An Improved Information Value Model Based on Gray Clustering for Landslide Susceptibility Mapping. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 18. https://doi.org/10.3390/ijgi6010018.
9. Ilia, I.; Tsangaratos, P. Applying weight of evidence method and sensitivity analysis to produce a landslide susceptibility map. *Landslides* **2016**, *13*, 379–397. https://doi.org/10.1007/s10346-015-0576-3.
10. Lambie, S.M.; Awatera, S.; Daigneault, A.; Kirschbaum, M.U.F.; Marden, M.; Soliman, T.; Spiekermann, R.I.; Walsh, P.J. Trade-offs between environmental and

economic factors in conversion from exotic pine production to natural regeneration on erosion prone land. *N. Z. J. For. Sci.* **2021**, *51*, 14.

11. Wang, G.; Chen, X.; Chen, W. Spatial Prediction of Landslide Susceptibility Based on GIS and Discriminant Functions. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 144. https://doi.org/10.3390/ijgi9030144.
12. Tehrani, F.S.; Calvello, M.; Liu, Z.; Zhang, L.; Lacasse, S. Machine learning and landslide studies: recent advances and applications. *Nat. Hazards.* **2022**, *114*, 1197–1245. https://doi.org/10.1007/s11069-022-05423-7.
13. Bergen, K.J.; Johnson, P.A.; de Hoop, M.V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **2019**, *363(6433)*, 1–11.
14. Wicki, A.; Lehmann, P.; Hauck, C.; Seneviratne, S.I.; Waldner, P.; Stähli, M. Assessing the potential of soil moisture measurements for regional landslide early warning. *Landslides* **2020**, *17*, 1881–1896. https://doi.org/10.1007/s10346-020-01400-y.
15. Johnston, E.C.; Davenport, F.V.; Wang, L.; Caers, J.K.; Muthukrishnan, S.; Burke, M.; Diffenbaugh, N.S. Quantifying the Effect of Precipitation on Landslide Hazard in Urbanized and Non-Urbanized Areas. *Geophys. Res. Lett.* **2021**, *48*, e2021GL094038. https://doi.org/10.1029/2021GL094038.
16. Abraham, M.T.; Satyam, N.; Pradhan, B.; Alamri, A.M. Forecasting of Landslides Using Rainfall Severity and Soil Wetness: A Probabilistic Approach for Darjeeling Himalayas. *Water* **2020**, *12*, 804. https://doi.org/10.3390/w12030804.
17. Martino, S.; Fiorucci, M.; Marmoni, G.M.; Casaburi, L.; Antonielli, B.; Mazzanti, P. Increase in landslide activity after a low-magnitude earthquake as inferred from DInSAR interferometry. *Sci. Rep.* **2022**, *12*, 2686. https://doi.org/10.1038/s41598-022-06508-w.
18. Nakileza, B.R.; Nedala, S. Topographic influence on landslides characteristics and implication for risk management in upper Manafwa catchment, Mt Elgon Uganda. *Geoenvironmental Disasters* **2020**, *7*, 27. https://doi.org/10.1186/s40677-020-00160-0.
19. Hosseini, S.A.; Lotfi, R.; Lotfalian, M.; Kavian, A.; Parsakhoo, A. The effect of terrain factors on landslide features along forest road. *Afr. J. Biotechnol.* **2011**, *10*, 14108–14115. https://doi.org/10.4314/ajb.v10i64.
20. Bergen, K.J.; Johnson, P.A.; de Hoop, M.V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **2019**, *363*, 1-10. https://doi.org/10.1126/science.aau0323.
21. Goetz, J.N.; Brenning, A.; Petschko, H.; Leopold, P. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci.* **2015**, *81*, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007.
22. Chen, T.; Zhu, L.; Niu, R.; Trinder, C.J.; Peng, L.; Lei, T. Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models. *J. Mt. Sci.* **2020**, *17*, 670–685. https://doi.org/10.1007/s11629-019-5839-3.
23. Stumpf, A.; Kerle, N. Combining Random Forests and object-oriented analysis for landslide mapping from very high resolution imagery. *Procedia Environ. Sci.* **2011**, *3*, 123–129. https://doi.org/10.1016/j.proenv.2011.02.022.
24. Lei, T.; Zhang, Y.; Lv, Z.; Li, S.; Liu, S.; Nandi, A.K. Landslide Inventory Mapping From Bitemporal Images Using Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 982–986. https://doi.org/10.1109/LGRS.2018.2889307.
25. Ma, Z.; Mei, G.; Piccialli, F. Machine learning for landslides prevention: a survey. *Neural Comput. Appl.* **2021**, *33*, 10881–10907. https://doi.org/10.1007/s00521-020-05529-8.

26. Mohr, K. NASA-USDA Global Soil Moisture Data. Earth, Online avaliable: https://earth.gsfc.nasa.gov/hydro/data/nasa-usda-global-soil-moisture-data.
27. GES DISC Dataset: Global Landslide Nowcasts from LHASA L4 1 day 1 km x 1 km version 1.1 (Global_Landslide_Nowcast) at GES DISC (Global_Landslide_Nowcast 1.1). https://disc.gsfc.nasa.gov/datasets/Global_Landslide_Nowcast_1.1/summary
28. Branco, P.; Torgo, L.; Ribeiro, R.P. SMOGN: a Pre-processing Approach for Imbalanced Regression. Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, 2017, pp. 36–50.
29. Hamadani, A.; Ganai, N.A.; Raja, T.; Alam, S.; Andrabi, S.M.; Hussain, I.; Ahmad, H.A. Outlier Removal in Sheep Farm Datasets Using Winsorization. Bhartiya Krishi Anusandhan Patrika, 2022.
30. Online avaliable: http://www.ss-pub.org/wp content/uploads/2019/09/JMSS18122402.pdf.
31. Ahsan, M.M.; Mahmud, M.A.P.; Saha, P.K.; Gupta, K.D.; Siddique, Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies* **2021**, *9*, 52. https://doi.org/10.3390/technologies9030052.
32. Mielniczuk, J.; Teisseyre, P. Model Selection in Logistic Regression Using p-Values and Greedy Search. In: Bouvry, P. Kłopotek, M.A. Leprévost, F. Marciniak, M. Mykowiecka, A. and Rybiński, H. (Eds.) Security and Intelligent Information Systems. Springer, Berlin, Heidelberg, 2012, pp. 128–141.
33. Sulaiman, M.A.; Labadin, J. Feature selection based on mutual information. In: 2015 9th International Conference on IT in Asia (CITA), 2015, pp. 1–6.
34. Sulaiman, M.A.; Labadin, J. Feature selection with mutual information for regression problems. Proceeding of the 9th International Conference on IT in Asia (CITA), 2015, pp. 1–6.

Landslide likelihood prediction in Vietnam using machine learning algorithms

Pham Trong Huynh[1]*

[1] University of Natural Resources and Environment, Ho Chi Minh City;
pthuynh@hcmunre.edu.vn

Abstract: Vietnam is a country with hilly and sloping terrain, located in a region with tropical monsoon climate, making landslides quite common. This study focuses on predicting the likelihood of landslides in Vietnam using regression algorithms, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor regression (KNN), Linear Support Vector Regressor (SVR), and Linear Regression (LR). Relevant landslide-related features are used, including soil moisture, seismic activity, rainfall, elevation, and slope. The algorithms are trained on a sample dataset to evaluate their performance. The research results show that the Random Forest (RF) algorithm can predict landslide susceptibility effectively. The prediction results from the training and testing datasets, with the highest coefficient of determination (R^2) value of 0.85, demonstrate a good ability to explain data variations. Additionally, the lowest mean squared error (MSE) and root mean squared error (RMSE) values are 150.21 and 12.25, respectively. The other algorithms also yield relatively good results, but RF outperforms them. This indicates the need to combine these five algorithms to handle a large amount of complex data, to create a stable and accurate landslide prediction model in Vietnam using machine learning algorithms.

Keywords: Landslides; Machine Learning; Regression; Random Forest; K-Nearest Neighbor.