ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY - CÂY QUYẾT ĐỊNH TRONG ĐÁNH GIÁ BIẾN ĐỘNG RỪNG NGẬP MẶN KHU VỰC XÃ ĐẤT MŨI

Nguyễn Thị Ngọc Ánh[(1)], Trần Đăng Hùng[(2)], Lê Phương Hà[(2)]
[(1)]Viện Chiến lược, Chính sách tài nguyên và môi trường (ISPONRE)
[(2)]Viện Khoa học Khí tượng thủy văn và Biến đổi khí hậu (IMHEN)

Ngày nhận bài: 04/11/2021; ngày chuyển phản biện: 05/11/2021; ngày chấp nhận đăng: 29/11/2021

Tóm tắt: *Phương pháp học máy - cây quyết định dùng để phục vụ các mục đích phân loại, tính hồi quy và các nhiệm vụ khác bằng cách xây dựng nhiều cây quyết định (Decision tree). Hiện nay cây quyết định là một phương pháp thông dụng trong khai thác dữ liệu. Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó [1]. Trong phạm vi bài báo này, nhóm nghiên tiến hành thử nghiệm một thuật toán của phương pháp học máy (Machine Learning) - cây quyết định trong phân loại các đối tượng sử dụng đất đặc biệt là rừng ngập mặn trên ảnh vệ tinh LANDSAT với khu vực thử nghiệm là xã Đất Mũi thuộc huyện Ngọc Hiển, tỉnh Cà Mau Cà Mau. Kết quả nghiên cứu đã phân loại thành công các lớp sử dụng đất giai đoạn 1995 - 2020 với độ chính xác tổng lần lượt cao là 88,8%, hệ số Kappa là 0,85 rất tốt đối với ảnh Landsat có độ phân giải trung bình.*

Từ khóa: *Viễn thám, rừng ngập mặn, cây quyết định.*

1. Giới thiệu

Từ trước đến nay, để chiết tách các thông tin ảnh viễn thám, việc ứng dụng các thuật toán có kiểm định như K-Nearest Neighbors (KNN) đã trở nên phổ biến. K-Nearest Neighbors phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp (Query point) và tất cả các đối tượng trong các bộ mẫu (Training Data). Tuy nhiên phương pháp này còn có 3 hạn chế là độ phức tạp tính toán do việc sử dụng tất cả các mẫu để phân loại, hiệu suất hoàn toàn phụ thuộc vào bộ mẫu giải đoán và không đánh giá được mức độ quan trọng giữa các mẫu. Vậy nên cần thiết phải xây dựng được một phương pháp phân loại mới, khắc phục được những hạn chế trên của các phương pháp cũ [2].

Hiện nay, các nhà nghiên cứu đã và đang phát triển nhiều thuật toán mới, phức tạp, mạnh mẽ và hiệu quả hơn. Một trong số đó là phương pháp học máy - cây quyết định. Phân loại cây quyết định như là phương pháp phân loại có giám sát khai thác dữ liệu không gian, phá vỡ các vấn đề và quy tắc phân loại trước đây cũng như luôn tận dụng được kiến thức sinh thái và viễn thám có tính chắc chắn và kết quả luôn liên quan chặt chẽ với kinh nghiệm và kiến thức chuyên môn. Nó có được các quy tắc phân loại bằng quy trình nghiên cứu quyết định và không cần phải thỏa mãn phân phối chuẩn. Nó có thể sử dụng kiến thức về Trái Đất trong cơ sở dữ liệu GIS để giúp phân loại và cải thiện độ chính xác của việc phân loại [3].

Phương pháp nghiên cứu cây quyết định là một trong những phương pháp khai phá dữ liệu để tìm ra các bài toán phân loại trong ứng dụng thực tế. Nó có thể phân loại các quy tắc của hình thức biểu thức cây quyết định. Ưu điểm tuyệt vời của cây quyết định là quá trình nghiên cứu không cần người dùng biết nhiều kiến thức nền tảng. Miễn là các ví dụ dữ liệu đầu vào có thể được thể hiện bằng "thuộc tính - kết quả" và sử dụng thuật toán này để học. Phân loại dữ liệu thu được bởi cây quyết định rất dễ thể hiện và áp dụng. Hiện nay, các học giả nước ngoài đã sử dụng cây quyết định để thu thập kiến thức và áp

Liên hệ tác giả: Nguyễn Thị Ngọc Ánh
Email: ngocanhnguyen1985@gmail.com

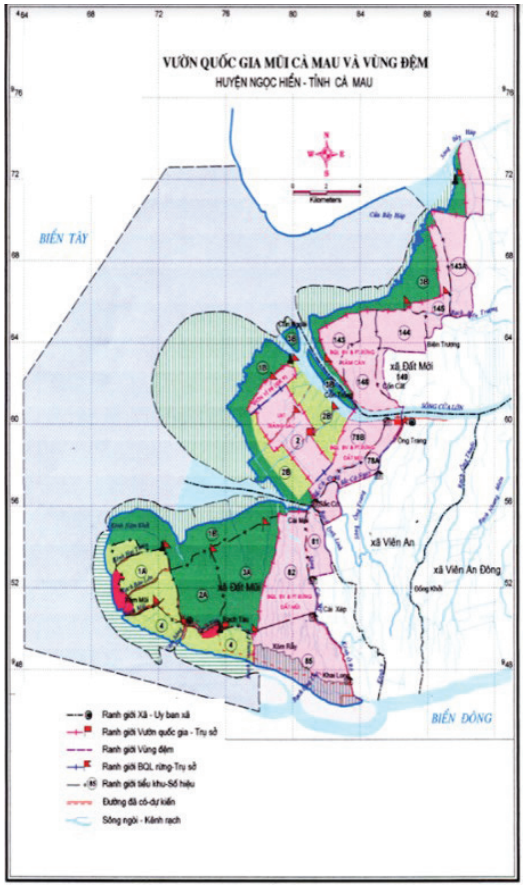
dụng trong quá trình nghiên cứu và phân tích không gian [6].

Thuật toán này cho phép con người xác định chính xác các thông tin phân loại và thống kê dựa vào các tập dữ liệu khổng lồ. Trong phạm vi bài báo này, nhóm nghiên cứu tiến hành thử nghiệm một thuật toán của phương pháp học máy (Machine Learning) - cây quyết định dùng ảnh vệ tinh Landsat có khả năng thành lập được các loại bản đồ biến động mục đích sử dụng đất tại từng thời điểm cụ thể; đảm bảo tính khách quan; tuy nhiên độ chính xác phụ thuộc vào nhiều yếu tố như chất lượng dữ liệu; kỹ năng sử dụng phần mềm; chọn mẫu.

2. Phương pháp nghiên cứu và tư liệu sử dụng

2.1. Khu vực nghiên cứu

Vườn quốc gia Mũi Cà Mau là một vườn quốc gia tại xã Đất Mũi, huyện Ngọc Hiển, tỉnh Cà Mau. Vị trí địa lý vườn quốc gia này có vị trí tại mũi đất cực Nam của lãnh thổ Việt Nam. Tọa độ từ 8º32’ đến 8º49’ vĩ Bắc và từ 104º40’ đến 104º55’ kinh Đông. Tổng diện tích tự nhiên 41.862 ha, trong đó diện tích đất liền 15.262 ha. Diện tích phần ven biển 26.600 ha. Vùng đệm của Vườn quốc gia Mũi Cà Mau có tổng diện tích 8.194 ha, nằm trên địa bàn các xã: Đất Mũi, Viên An và Đất Mới thuộc huyện Ngọc Hiển, tỉnh Cà Mau.



Hình 1. Khu vực nghiên cứu

2.2. Dữ liệu nghiên cứu

Nghiên cứu này sẽ kiểm tra khả năng nhận dạng và phân loại bằng thuật toán cây quyết định đối với sự thay đổi sử dụng đất đặc biệt là rừng ngập mặn của khu vực Vườn quốc gia Mũi Cà Mau. Hình ảnh vệ tinh quang học đa phổ cho thấy biến động rừng ngập mặn theo thời gian có thể được giám sát bằng cách sử dụng phương pháp phân tích biến động sau phân loại. Trong phương pháp này, trước tiên dữ liệu ảnh vệ tinh đa phổ khu vực nghiên cứu từng thời điểm được tiến hành phân loại độc lập. Sau đó sử

dụng phương pháp GIS để tiến hành phát hiện biến động bằng cách so sánh ảnh phân loại của cùng 1 vùng tại hai thời điểm khác nhau.

Vệ tinh LANDSAT có đặc tính kỹ thuật thu nhận trên nhiều kênh phổ khác nhau nên thể hiện tương đối đầy đủ các đặc trưng nổi bật và khái quát của các đối tượng trên bề mặt Trái Đất. Nhưng vấn đề cốt lõi để có thể giải đoán, chiết suất các thông tin hữu ích từ ảnh viễn thám đòi hỏi phải có kiến thức chuyên gia và bề dày kinh nghiệm về giải đoán ảnh, xử lý ảnh. Sử dụng ảnh viễn thám Landsat 5 và Landsat 8 để giải đoán và thành lập các bản đồ hiện trạng sử dụng đất năm 1993, năm 2020 sau đó tính toán để đánh giá sự biến động diện tích RNM trong thời kì 1993 - 2020 diễn ra như thế nào. Để giảm thiểu ảnh hưởng của mây, chúng tôi ưu tiên sử dụng ảnh được chụp vào mùa khô (từ tháng 11 - tháng 4), nhưng do số lượng ảnh hạn chế nên việc sử dụng ảnh được chụp vào cuối mùa mưa là hoàn toàn chấp nhận được. Thông tin về ảnh vệ tinh được thể hiện trong Bảng 1:

Bảng 1. Bảng thống kê dữ liệu sử dụng trong nghiên cứu

Mã ảnh	Chất lượng ảnh	Độ phân giải	Ngày chụp
LANDSAT/LT05/C01/T1_SR/LT05_126054_19951226	7	30 m	26/12/1995
Image LANDSAT/LC08/C01/T1_SR/LC08_126054_20200317 (12 bands)	9	30 m	17/3/2020

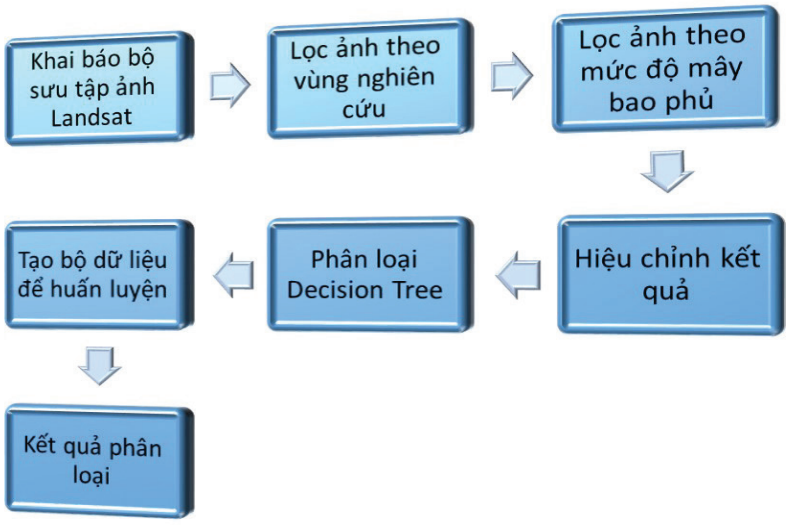
Nguồn: http://earthexplorer.usgs.gov

Hình 2. Tổ hợp màu tự nhiên ảnh năm 1995 và 2020

2.3. Phương pháp thực hiện

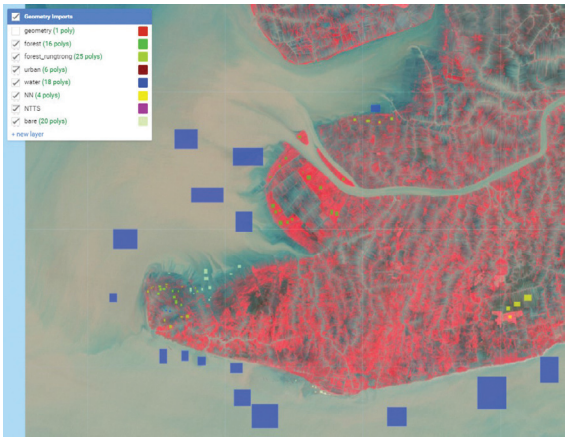
Cây quyết định là một phương pháp có thể học quy nạp bằng cách đào tạo các mẫu và xây dựng cây quyết định hoặc quy tắc quyết định và sau đó sử dụng cây quyết định hoặc quy tắc quyết định để phân loại dữ liệu. Cây quyết định là một công trình cây. Nó được cấu tạo bởi một nút gốc, một loạt các nút bên trong và các nút lá. Mỗi nút chỉ có thể có một nút chính và hai hoặc nhiều nút phụ. Các nút được kết nối với nhau bằng các nhánh. [4] Mỗi nút bên trong tương ứng với một thuộc tính hoặc nhóm thuộc tính thử nghiệm và mọi bên tương ứng với mọi giá trị có thể có của thuộc tính. Nút tương ứng với một giá trị thuộc tính của lớp và nút khác nhau có thể tương ứng với cùng một giá trị thuộc tính của lớp. Cây quyết định không chỉ có thể được thể hiện bằng cây, mà còn là một nhóm các quy tắc sản IF-THEN [5]. Mỗi đường từ gốc đến lá tương ứng với một quy tắc và điều kiện của quy tắc là tùy chọn tất cả các giá trị thuộc tính của các nút, kết quả của quy tắc là thuộc tính lớp của nút lá trên đường. So với các thuộc tính quyết định, các quy tắc đơn giản và thuận tiện hơn để hiểu, sử dụng và sửa chữa và có thể tạo nên cơ sở của hệ thống chuyên gia. Vì vậy quy tắc được sử dụng ngày càng nhiều trong ứng dụng thực tế.

Bài báo sử dụng phần mềm Google Earth Engine (GEE). Google Earth Engine làm việc thông qua Giao diện Trực tuyến của Ứng Dụng JavaScript (API) được gọi là Code Editor. Trên giao diện này, người dùng có thể viết và chạy các tập lệnh/script để chia sẻ và lặp lại các quy trình phân tích cũng như xử lý dữ liệu không gian địa lý. Code Editor giúp người dùng thực hiện toàn bộ các chức năng có trong Earth Engine. Quy trình xây dựng phương pháp phân tích thảm phủ bao gồm các đối tượng rừng ngập mặn già, rừng ngập mặn mới trồng, nông thủy hải sản, đất trống và đất dân cư cho ảnh LANDSAT được thể hiện chi tiết ở Hình 3.

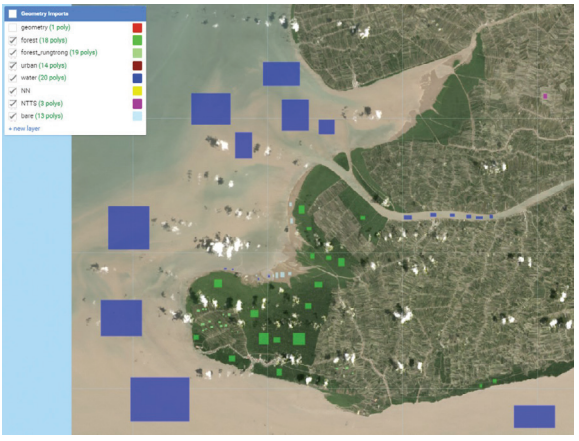


Hình 3. Phương pháp xây dựng phân loại cây quyết định trên GEE

Đầu tiên tiến hành khai báo bộ dữ liệu LANDSAT là bộ dữ liệu đầu vào để phân tích. Tiếp theo tiến hành lọc ảnh theo khu vực nghiên cứu cũng như là tiến hành lọc các cảnh ảnh ít mây. Sau khi lọc ảnh tiến hành tạo bộ dữ liệu để huấn luyện phân loại các lớp đối tượng sử dụng đất. Tiến hành lấy mẫu thật chi tiết, chính xác cũng như bộ mẫu càng nhiều thì kết quả đầu ra càng tốt.



Hình 4. Số lượng và vị trí điểm lấy mẫu phân loại cho từng đối tượng cho ảnh Landsat năm 1995 trên giao diện GEE

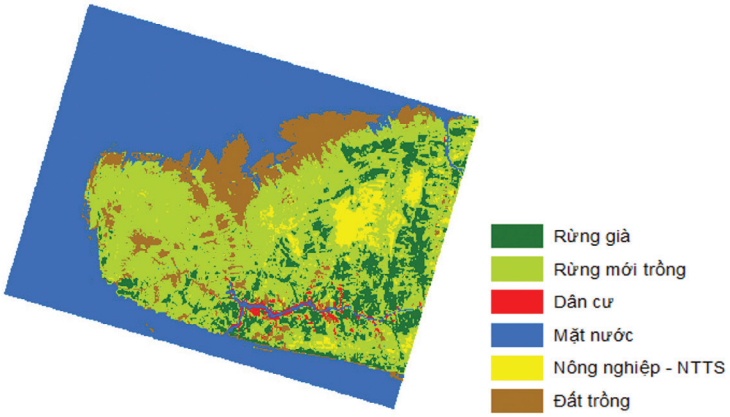


Hình 5. Số lượng và vị trí điểm lấy mẫu phân loại cho từng đối tượng cho ảnh Landsat năm 2020 trên giao diện GEE

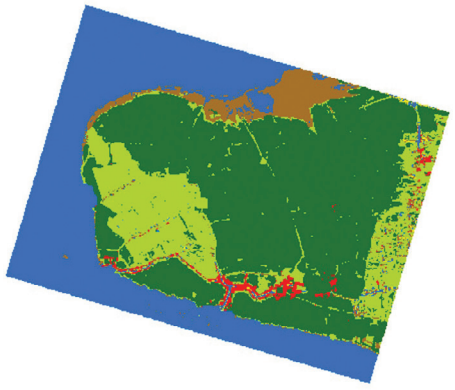
Sau khi tạo bộ dữ liệu huấn luyện xong tiến hành phân loại cây quyết định theo thuật toán Cart. Kết quả phân loại được hiệu chỉnh bằng cách lấy mẫu đi lấy mẫu lại cho đến khi đạt kết quả tốt nhất. Cuối cùng trích xuất kết quả phân loại rừng và hiệu chỉnh kết quả trên phần mềm Arcmap.

3. Kết quả và thảo luận

Kết quả phân loại cuối cùng cho các ảnh Landsat năm 1995 và 2020 được thể hiện chi tiết như Hình 6 và 7.



Hình 6. Kết quả phân loại ảnh Landsat năm 1995



Hình 7. Kết quả phân loại ảnh Landsat năm 2020

Để đánh giá độ chính xác của phương pháp phân loại do không có điều kiện thu thập dữ liệu trong quá khứ vậy nên bài báo đã sử dụng nguồn dữ liệu ảnh google earh để tham khảo và kiểm chứng. Một bộ sưu tập 150 điểm khảo sát ngẫu nhiên được tạo bằng phương pháp random point trên phần mềm Erdas với từng đối tượng phân loại để đánh giá và kiểm chứng. Kết quả cuối cùng được chi tiết trên Bảng 2.

Bảng 2. Bảng thống kê kết quả sau phân loại

Tên lớp	Số lượng mẫu tham chiếu	Số lượng mẫu chọn	Số lượng mẫu chính xác	Độ chính xác tham chiếu	Độ chính xác Thực tế
Thủy hệ	20	22	18	100%	90,91%
Dân cư	23	21	19	82,61%	90,48%
Nông lâm thủy hải sản	39	30	32	82,05%	94,12%
Rừng ngập mặn già	24	28	23	95,83%	82,14%
Rừng ngập mặn non	19	20	17	89,47%	85,00%
Đất trống	25	29	27	93,34%	91,2%
Độ chính xác phân loại tổng thể = 88,8% Số liệu thống kê Kappa tổng thể = 0,85					

Kết quả cho thấy độ chính xác tổng thể đạt 88.8%, số liệu thống kê Kappa tổng thể đạt 0.85. Đối với các lớp phân loại độ chính xác thực tế so với độ chính xác tham chiếu không có sự chênh lệch đáng kể. Lớp rừng ngập mặn già và rừng ngập mặn non độ chính xác đều đạt trên 80%.

Bảng 3 cho thấy, diện tích rừng ngập mặn khu vực rừng quốc gia Cà Mau có sự thay đổi tương đối lớn. Diện tích rừng già (rừng phòng hộ) năm 2020 tăng gần gấp 5 lần với diện tích rừng phòng hộ năm 1995. Trong khi đó diện tích rừng trồng mới năm 2020 giảm 2 lần so với diện tích rừng trồng mới năm 1995. Các loại đất như dân cư năm 2020 tăng gấp đôi so với năm 1995 kéo theo sự giảm đáng kể của các loại đất thủy hệ, đất nông nghiệp - nuôi trồng thuỷ sản, đất trống.

Bảng 3. Bảng so sánh diện tích sử dụng đất năm 1995 và năm 2020

Diện tích (ha)	1995	2020
Rừng già	818.518	3.791.948
Rừng trồng	2.814.599	1.321.205
Dân cư	63.096	100.869
Thủy hệ	3.631.993	2.743.868
Đất nông nghiệp -ntts	282.011	19.329
Đất trống	841.501	473.529

4. Kết luận

Kết quả nghiên cứu đã phân loại thành công được các lớp sử dụng đất cho khu vực vườn quốc gia Mũi Cà Mau và đều với độ chính xác cao, độ chính xác tổng đạt được tới 89%. Bài báo đã đưa ra kết quả sư thay đổi diện tích các loại rừng nói riêng và các loại đất sử dụng nói chung tại khu vực vườn Quốc gia Cà Mau. Cho thấy được tại khu vực nghiên cứu rừng được bảo tồn và trồng mới rất tốt trong giai đoạn 1995 - 2020 dưới tác động của biến đổi khí hậu. Sử dụng phương pháp học máy - cây quyết định đã giúp cải thiện được kết quả phân loại khá tốt. Điều đó cho thấy, việc sử dụng ảnh viễn thám Landsat và công nghệ AI trong đánh giá biến động diện tích rừng mang lại kết quả đáng tin cậy.

Kết quả thực hiện của nghiên cứu đã đạt được 2 điểm mới đó là: Ứng dụng thành công phương pháp học máy - cây quyết định trong phân loại ảnh viễn thám và khả năng của phương pháp phân loại được chi tiết theo loài thực vật của rừng ngập mặn tại khu vực thực nghiệm. Nghiên cứu được thực hiện và đã đạt được những kết quả nhất định, tuy nhiên nhóm nghiên cứu có một số kiến nghị cần nghiên cứu tiếp để có những kết quả có độ chính xác cao hơn. Đó là: 1) Sử dụng ảnh RADAR để có thể phản ảnh cấu trúc đứng của rừng ngập mặn; 2) Sử dụng các yếu tố kiến trúc ảnh (image texture) và các chỉ số hình dạng (shape index) trong phân loại kiểu rừng; 3) Thiết kế các điểm khảo sát, OTC để xác định loài, cấu trúc rừng, trữ lượng rừng để phân loại, kiểm chứng từ ảnh vệ tinh.

Tài liệu tham khảo

1. Sesnie, S.E. et al. (2018), *"Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments". Remote Sens. Environ*, 112, 2145-2159.
2. Rodriguez-Galiano et al. (2012), *"An assessment of the effectiveness of a random forest classifier for land-cover classification". ISPRS J. Photogramm. Remote Sens*, 67, 93-104.
3. Li, S., Ding, S. (2002), *"Decision Tree Classify Method and Application in Earth Coverage Classify", Remote Sensing Technology and Application* 17(1), 6-11.
4. Li, F., Li, M. (2003), *"Remote Sensing Image Auto Classify Study Based on Combination of Artificial Neural Networks and Decision Tree", Remote Sensing Information* 3, 3-25.
5. Jiang, Q., Liu, H. (2004), *"Use Texture Analysis to Extract TM Image Information", Remote Sensing Journal* 8(5), 458-464.
6. Friedl, M.A., Brodley, C.E., Strahler, A.H. (1999), *"Maximizing land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales", IEEE Transactions on Geoscience and Remote Sensing* 37(2), 969-977.

APPLYING THE METHOD OF MACHINE LEARNING - DECISION TREE IN ASSESSING THE MANGROVE FOREST CHANGES IN DAT MUI COMMUNE

Nguyen Thi Ngoc Anh[(1)], Tran Dang Hung[(2)], Le Phuong Ha[(2)]
[(1)]Institute of Strategy and Policy on Natural Resources and Environment
[(2)]Viet Nam Institute of Meteorology, Hydrology and Climate change

Received: 04/11/2021; Accepted: 29/11/2021

Abstract: *Method of machine learning - decision tree is used for classification, regression and other tasks by building many decision trees. Decision trees are now a popular method in data mining. The decision tree then describes a tree structure, where the leaves represent the categories and the branches represent the combinations of attributes that lead to that classification [1]. Within the scope of this paper, the research team tested an algorithm of machine learning method (Machine Learning) - decision tree in classifying land use objects, especially mangrove forests on LANDSAT satellite images with The test area is Dat Mui commune, Ngoc Hien district, Ca Mau province. The research results have successfully classified the land use classes for the period 1995 - 2020 with a high total accuracy of 88.8 %, respectively, and a Kappa coefficient of 0.85 which is very good for Landsat images with medium resolution.*

Keywords: *Remote sensing, mangrove forest, random forest.*