



Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trang website: <https://jstt.vn/index.php/vn>



Article info

Type of article:

Original research paper

DOI:

<https://doi.org/10.58845/jstt.utt.2025.vn.5.4.110-122>

*Corresponding author:

Email address:

damnd@utt.edu.vn

Received: 23/10/2025

Received in Revised Form:
11/12/2025

Accepted: 12/12/2025

Application of the LGBM model for predicting the maximum dry density of soil used in roadbed compaction

Vu Trong Hieu, Nguyen Duc Dam*, Vu Quang Dung, Be Ngoc Son, Ngo Quoc Trinh, Phung Tang Nghi
University of Transport Technology, 54 Trieu Khuc Street, Thanh Liet Ward, Hanoi, Vietnam

Abstract: In civil engineering, accurately determining the maximum dry density (MDD) of embankment soil plays a crucial role in ensuring construction quality. This study proposes the application of the Light Gradient Boosting Machine (LGBM) algorithm to predict MDD based on the geotechnical properties of soil. The dataset consists of 214 soil samples collected from the Van Don – Mong Cai expressway construction project (Vietnam) with input parameters including gravel content (G), coarse sand content (CS), fine sand content (FS), silt-clay content (SC), organic content (O), liquid limit (LL), plastic limit (PL), plasticity index (PI), and MDD as the output parameter. Quantitative evaluation metrics such as R^2 , RMSE, and MAE were used to assess the model's performance. The results indicate that the LGBM achieves good predictive performance with the following metrics on the training dataset: $R^2 = 0.941$, RMSE = 0.028, MAE = 0.020; and on the testing dataset: $R^2 = 0.771$, RMSE = 0.059, MAE = 0.048. Additionally, model interpretation methods such as SHAP and partial dependence plots (PDP) show that factors like gravel content and coarse sand content contribute to increasing MDD, while other factors tend to decrease it. The findings highlight the significance of adopting a LGBM-based approach for rapid and accurate MDD prediction to support road design and calculation.

Keywords: LGBM, soil compaction, machine learning, SHAP, geotechnical engineering.



Thông tin bài viết
Dạng bài viết:
Bài báo nghiên cứu

DOI:
<https://doi.org/10.58845/jstt.utt.2025.vn.5.4.110-122>

*Tác giả liên hệ:
Địa chỉ Email:
damnd@utt.edu.vn

Ngày nộp bài: 23/10/2025
Ngày nộp bài sửa: 11/12/2025
Ngày chấp nhận: 12/12/2025

Nghiên cứu áp dụng mô hình LGBM trong dự báo khối lượng thể tích khô lớn nhất của đất sử dụng trong đầm nén nền đường

Vũ Trọng Hiếu, Nguyễn Đức Đàm*, Vũ Quang Dũng, Bế Ngọc Sơn, Ngô Quốc Trinh, Phùng Tăng Nghị
Trường Đại học Công nghệ Giao thông vận tải, 54 Triều Khúc, Thanh Liệt, Hà Nội, Việt Nam

Tóm tắt: Trong xây dựng nền đường, xác định chính xác khối lượng thể tích khô lớn nhất (MDD) của đất đắp đóng vai trò then chốt nhằm đảm bảo chất lượng công trình. Nghiên cứu đề xuất ứng dụng thuật toán Light Gradient Boosting Machine (LGBM) để dự báo MDD dựa trên các thông số cơ lý của đất. Dữ liệu sử dụng gồm 214 mẫu đất thu thập từ dự án xây dựng đường cao tốc Vân Đồn – Móng Cái, Việt Nam với các tham số đầu vào bao gồm: hàm lượng sỏi (G), hàm lượng cát thô (CS), hàm lượng cát mịn (FS), hàm lượng bụi sét (SC), hàm lượng hữu cơ (O), giới hạn chảy (LL), giới hạn dẻo (PL), chỉ số dẻo (PI) và MDD là tham số đầu ra. Các chỉ số đánh giá định lượng như R^2 , RMSE, MAE được sử dụng để đánh giá hiệu suất của mô hình. Kết quả đánh giá cho thấy mô hình đạt hiệu suất dự báo tốt với các chỉ số trên tập huấn luyện gồm: $R^2 = 0.941$, RMSE = 0.028, MAE = 0.020 và trên tập kiểm chứng các chỉ số tương ứng là: $R^2 = 0.771$, RMSE = 0.059, MAE = 0.048. Đồng thời, các phương pháp giải thích mô hình như SHAP và biểu đồ phụ thuộc từng phần (PDP) cho thấy các yếu tố như hàm lượng sỏi và cát thô góp phần làm tăng MDD, trong khi các yếu tố khác lại làm giảm giá trị. Kết quả nghiên cứu có ý nghĩa trong việc tiếp cận phương pháp hiệu quả dựa trên LGBM trong dự báo nhanh, chính xác giá trị MDD phục vụ tính toán và thiết kế nền đường.

Từ khóa: LGBM, đầm nén đất, học máy, SHAP, địa kỹ thuật.

1. Đặt vấn đề

Khối lượng thể tích khô lớn nhất của đất (Maximum Dry Density - MDD) là đại lượng đặc trưng cho khả năng đạt độ chặt của đất khi được đầm nén trong điều kiện tiêu chuẩn, đồng thời là cơ sở để kiểm soát chất lượng công tác đầm nén tại hiện trường [1]. Theo các tiêu chuẩn hiện hành TCVN 4447:2012 và 22TCN 333-06, độ chặt thực tế của nền đất sau khi đầm nén thường được yêu cầu đạt từ 95% đến 98% tùy thuộc vào loại công trình (đường cấp cao, sân bay, nền móng công

trình trọng yếu) và điều kiện nền đất. Việc đạt được độ chặt theo yêu cầu không chỉ đảm bảo yêu cầu kỹ thuật mà còn mang lại nhiều lợi ích về mặt cơ học cho công trình bao gồm: nâng cao khả năng chịu tải và độ ổn định tổng thể của nền đường, góp phần giảm thiểu hiện tượng lún sụt dưới tác động của tải trọng; giảm tính thấm của đất, hạn chế sự xâm nhập và tích tụ nước gây suy yếu kết cấu nền; tăng khả năng chống chịu với các biến động môi trường như dao động nhiệt độ hoặc chu trình khô – ẩm; đồng thời làm giảm độ rỗng và tăng mức độ

liên kết giữa các hạt đất, qua đó hạn chế co ngót, trương nở và lún thứ cấp sau thi công.

Để đảm bảo kiểm soát và đáp ứng các yêu cầu kỹ thuật nêu trên được thực hiện trên cơ sở xác định giá trị MDD trong giai đoạn thiết kế và thi công nền đường. Thông thường, thông số này được xác định thông qua thí nghiệm đầm nén trong phòng, phổ biến nhất là thí nghiệm Proctor tiêu chuẩn hoặc Proctor cải tiến, với số lượng đầm xác định cụ thể [2]. Trong quá trình thí nghiệm, các mẫu đất được đầm nén ở các mức độ ẩm khác nhau để xác định khối lượng thể tích khô lớn nhất tương ứng, từ đó thiết lập đường cong biểu diễn mối quan hệ giữa độ ẩm và khối lượng thể tích khô lớn nhất. Mặc dù phương pháp thí nghiệm này cho kết quả tin cậy, song quá trình thực hiện thường tốn nhiều thời gian, công sức và chi phí. Xuất phát từ nhu cầu đơn giản hóa và tăng hiệu quả trong công tác đánh giá đặc tính đầm nén của đất, nhiều phương pháp thay thế đã được nghiên cứu, trong đó đáng chú ý là các mô hình thực nghiệm, mô hình vật lý và đặc biệt là các phương pháp dự báo dựa trên trí tuệ nhân tạo. Mục tiêu chung của các phương pháp này là giảm bớt số lượng thí nghiệm Proctor phải thực hiện bằng cách ước tính nhanh khối lượng riêng khô lớn nhất từ các chỉ tiêu cơ lý cơ bản của đất.

Gần đây, các mô hình trí tuệ nhân tạo (AI) như mạng nơron nhân tạo (ANN), hệ mờ (Fuzzy Logic), thuật toán di truyền (GA) và các mô hình học máy hiện đại như rừng ngẫu nhiên (RF), Gradient Boosting, LGBM đã được ứng dụng rộng rãi trên thế giới để dự đoán các chỉ tiêu cơ lý của đất, bao gồm cả MDD [3-10]. Cụ thể, Günaydin [11] sử dụng mạng nơron nhân tạo (ANN) để dự báo MDD từ thành phần hạt và các chỉ tiêu giới hạn Atterberg, đạt hệ số xác định $R^2 = 0.95$ cao hơn đáng kể so với hồi quy tuyến tính ($R^2 = 0.70$), cho thấy khả năng mô phỏng phi tuyến của ANN. Tương tự, Trọng Kiến Dương và cộng sự [12] đã áp dụng mô hình di truyền GEP và MEP cho đất sét trương nở, không chỉ đạt độ chính xác cao ($R^2 > 0.90$) mà còn cung cấp các biểu thức dự báo rõ ràng, trong đó giới hạn dẻo được xác định là yếu

tố ảnh hưởng chính đến MDD. Li và cộng sự [13] đã thực hiện một so sánh toàn diện giữa các mô hình học máy (SVM, ANN, RF, XGBoost) và kết quả cho thấy XGBoost có hiệu suất cao nhất trong dự báo MDD với $R^2 = 0.91$, vượt trội so với SVM ($R^2 = 0.90$), ANN ($R^2 = 0.85$) và RF ($R^2 = 0.85$). Mô hình rừng ngẫu nhiên tuy cho kết quả tốt với độ ẩm tối ưu (OMC) nhưng chưa thực sự tối ưu với MDD. Những kết quả này cho thấy mô hình AI hoàn toàn có thể ứng dụng để thay thế hoặc hỗ trợ thí nghiệm Proctor trong giai đoạn thiết kế nền đường, giúp tiết kiệm thời gian, chi phí và tăng độ linh hoạt trong lựa chọn phương án thi công.

Nhìn chung, các nghiên cứu trên đã cho thấy xu hướng kết hợp dữ liệu thực nghiệm với các mô hình AI đang mang lại hiệu quả cao trong lĩnh vực địa kỹ thuật. Việc áp dụng các mô hình học máy hiện đại không chỉ giúp rút ngắn thời gian và chi phí thí nghiệm, mà còn góp phần nâng cao tính chính xác và khả năng phân tích mối quan hệ giữa các biến đầu vào. Trong bối cảnh đó, mô hình LGBM là một thuật toán tăng cường cây quyết định với ưu điểm về tốc độ huấn luyện, khả năng xử lý dữ liệu lớn và độ chính xác cao được lựa chọn trong nghiên cứu này nhằm xây dựng mô hình dự báo MDD từ các chỉ tiêu cơ lý cơ bản của đất. Đây là hướng tiếp cận phù hợp với xu thế hiện đại trong địa kỹ thuật, đồng thời có tiềm năng ứng dụng thực tiễn cao trong thiết kế và thi công nền đường tại Việt Nam.

2. Dữ liệu và phương pháp nghiên cứu

2.1. Dữ liệu nghiên cứu

Khu vực nghiên cứu thuộc dự án xây dựng tuyến đường cao tốc Vân Đồn – Móng Cái, thuộc tỉnh Quảng Ninh (Việt Nam). Tuyến đường có tổng chiều dài khoảng 80 km, được thiết kế theo quy mô 4 làn xe, bề rộng nền đường 25 m và vận tốc thiết kế 120 km/h. Tuyến cao tốc nối liền từ khu vực Vân Đồn đến cửa khẩu quốc tế Móng Cái, đóng vai trò đặc biệt quan trọng trong việc kết nối các trung tâm kinh tế – du lịch ven biển như Hạ Long, Cẩm Phả, Vân Đồn với khu vực biên giới phía Đông Bắc, đồng thời là một phần của hành lang giao thông trọng điểm trong chiến lược phát triển kinh tế và

hợp tác thương mại giữa Việt Nam và Trung Quốc.

Về đặc điểm tự nhiên, tuyến đi qua nhiều dạng địa hình đa dạng, chủ yếu là các dải đồi núi thấp xen kẽ với thung lũng hẹp và vùng đất thấp ven biển. Trong đó, nhiều đoạn tuyến nằm trên các khu vực có nền đất yếu, điển hình là các loại đất bùn, đất sét mềm và trầm tích bồi tích ven biển vốn có đặc điểm cơ lý bất lợi như sức chịu tải thấp, độ lún lớn và dễ mất ổn định khi chịu tác động của tải trọng công trình và ảnh hưởng của nước ngầm. Do đó, việc nghiên cứu đặc điểm địa kỹ thuật của đất nền tại khu vực này có vai trò quan trọng, không chỉ nhằm đảm bảo an toàn và ổn định công trình, mà còn góp phần tối ưu hóa giải pháp thiết kế, lựa chọn vật liệu nền phù hợp và giảm thiểu rủi ro trong quá trình thi công.

Dữ liệu nghiên cứu bao gồm 214 mẫu đất được lấy trực tiếp từ hiện trường và mô vật liệu với các mẫu được thu thập tại các vị trí khác nhau trên toàn tuyến. Các mẫu được tiến hành thí nghiệm tại phòng thí nghiệm cơ học đất... tuân thủ theo các tiêu chuẩn hiện hành của Việt Nam như TCVN 4200:2012 và TCVN 5747:1993... Trong quá trình thí nghiệm, các chỉ tiêu được xác định bao gồm: hàm lượng sỏi (G), hàm lượng cát thô (CS), hàm lượng cát mịn (FS), hàm lượng bụi sét (SC), hàm lượng hữu cơ (O), giới hạn chảy (LL), giới hạn dẻo

(PL), chỉ số dẻo (PI) và MDD. Đây là những thông số cơ bản nhưng đóng vai trò quan trọng trong việc đánh giá đặc trưng cơ lý của đất, phục vụ trực tiếp cho quá trình xây dựng mô hình dự báo.

Các chỉ tiêu như hàm lượng sỏi (G), cát thô (CS), cát mịn (FS) và bụi sét (SC) phản ánh thành phần hạt của đất, từ đó ảnh hưởng đến độ chặt, tính thấm và khả năng chịu tải. Đất có hàm lượng sỏi và cát thô cao thường hình thành khung hạt ổn định, khả năng kháng cắt lớn, ít lún và thoát nước tốt. Ngược lại, đất chứa nhiều cát mịn hoặc hạt bụi dễ bị hóa lỏng và suy giảm cường độ khi bão hòa. Hàm lượng hữu cơ (O) cao làm giảm khối lượng thể tích khô, tăng độ rỗng và độ nén lún, khiến đất yếu và mất ổn định, đặc biệt trong điều kiện ẩm ướt. Bên cạnh đó, các chỉ tiêu giới hạn chảy (LL), giới hạn dẻo (PL) và chỉ số dẻo (PI) thể hiện khả năng giữ nước và biến dạng của đất; trong đó, chỉ số PI cao thường đi kèm với nguy cơ trương nở, co ngót và suy giảm sức chịu tải khi độ ẩm thay đổi. Tất cả các chỉ tiêu nêu trên được xem là các biến đầu vào của mô hình dự báo, đóng vai trò giải thích đặc trưng cơ lý của đất và tác động đến khả năng đầm chặt. Trong khi đó, khối lượng thể tích khô lớn nhất (MDD) là biến đầu ra của mô hình phản ánh khả năng đạt được độ chặt tối ưu của đất trong điều kiện đầm nén tiêu chuẩn.

Bảng 1. Bảng thống kê dữ liệu sử dụng

STT	Tham số	Ký hiệu	Đơn vị	Giá trị trung bình	Độ lệch chuẩn (Std)	Giá trị nhỏ nhất	Giá trị lớn nhất
1	Hàm lượng sỏi	G	%	22.057	13.295	0.000	51.40
2	Hàm lượng cát thô	CS	%	24.101	7.017	3.000	46.30
3	Hàm lượng cát mịn	FS	%	9.035	6.468	2.500	41.50
4	Hàm lượng bụi sét	SC	%	44.807	10.447	17.870	88.70
5	Hàm lượng hữu cơ	O	%	1.509	0.373	0.120	2.94
6	Giới hạn chảy	LL	%	39.515	6.173	2.080	48.45
7	Giới hạn dẻo	PL	%	20.318	3.068	1.170	28.49
8	Chỉ số dẻo	PI	%	19.198	4.078	0.910	27.48
9	Khối lượng thể tích khô lớn nhất	MDD	g/cm ³	1.880	0.118	1.672	2.14

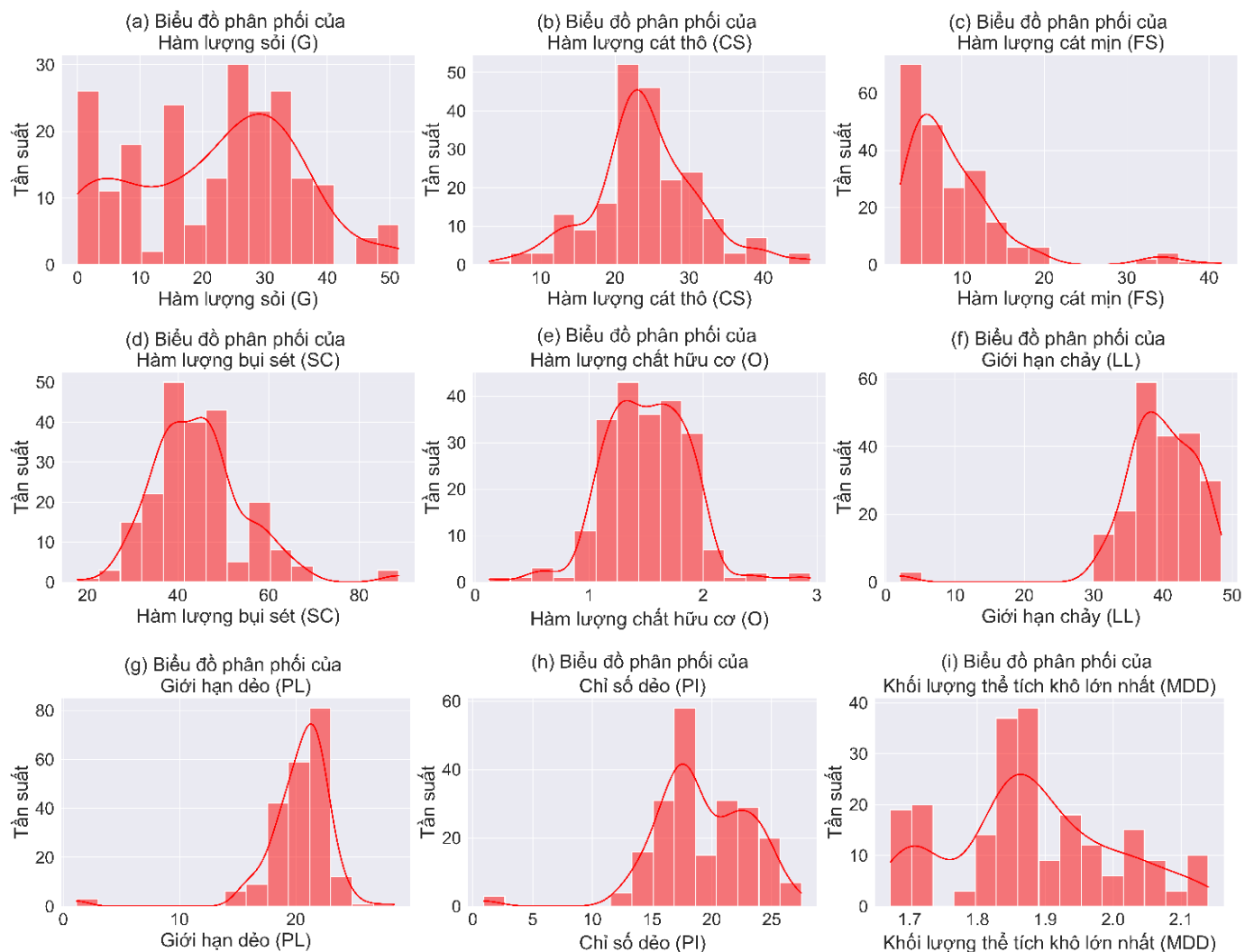
Phân tích bảng thống kê trong Bảng 1 cho thấy, hàm lượng bụi sét chiếm tỷ lệ trung bình cao 44.81 % và biến động lớn giữa các mẫu, phản ánh

đặc điểm đất chủ yếu thuộc loại hạt mịn. Trong khi đó, các thành phần hạt thô như sỏi và cát thô có giá trị trung bình thấp hơn 22.06% và 24.10%

nhưng thể hiện sự hiện diện đáng kể trong thành phần đất, cho thấy nhiều mẫu đất có kết cấu phối hợp giữa hạt lớn và hạt mịn. Hàm lượng hữu cơ ở mức thấp trung bình 1.51 %. Các chỉ tiêu tính dẻo như giới hạn chảy, giới hạn dẻo và chỉ số dẻo có mức trung bình từ trung bình đến cao cho thấy đất có độ dẻo nhất định. Với số liệu thống kê cho thấy sự đa dạng về thành phần hạt và đặc tính vật lý của đất trên toàn tuyến.

Các biểu đồ phân phối trong Hình 1 cho thấy, hầu hết các biến trong tập dữ liệu đều có sự phân bố rõ ràng và hợp lý, phản ánh tính đa dạng của các mẫu đất được thu thập. Nhìn chung, các đặc

trung đầu vào có xu hướng phân bố lệch nhưng vẫn duy trì được dạng phân bố nhất quán, không có sự xuất hiện bất thường hay giá trị ngoại lai quá lớn. Các thành phần hạt như sỏi, cát thô và bụi sét phân bố trải rộng, phản ánh sự khác biệt đáng kể về kết cấu đất giữa các mẫu. Trong khi đó, các biến liên quan đến tính dẻo như giới hạn chảy, giới hạn dẻo và chỉ số dẻo có phân phối tập trung hơn, cho thấy tính ổn định về mặt cơ lý trong nhóm mẫu khảo sát. Đặc biệt, biến mục tiêu MDD phân bố tương đối đều, tập trung chủ yếu từ 1.8 đến 2.0 g/cm³, với giá trị cao nhất lên tới 2.14 cho thấy độ biến thiên phù hợp.



Hình 1. Phân phối dữ liệu của các biến được sử dụng trong mô hình

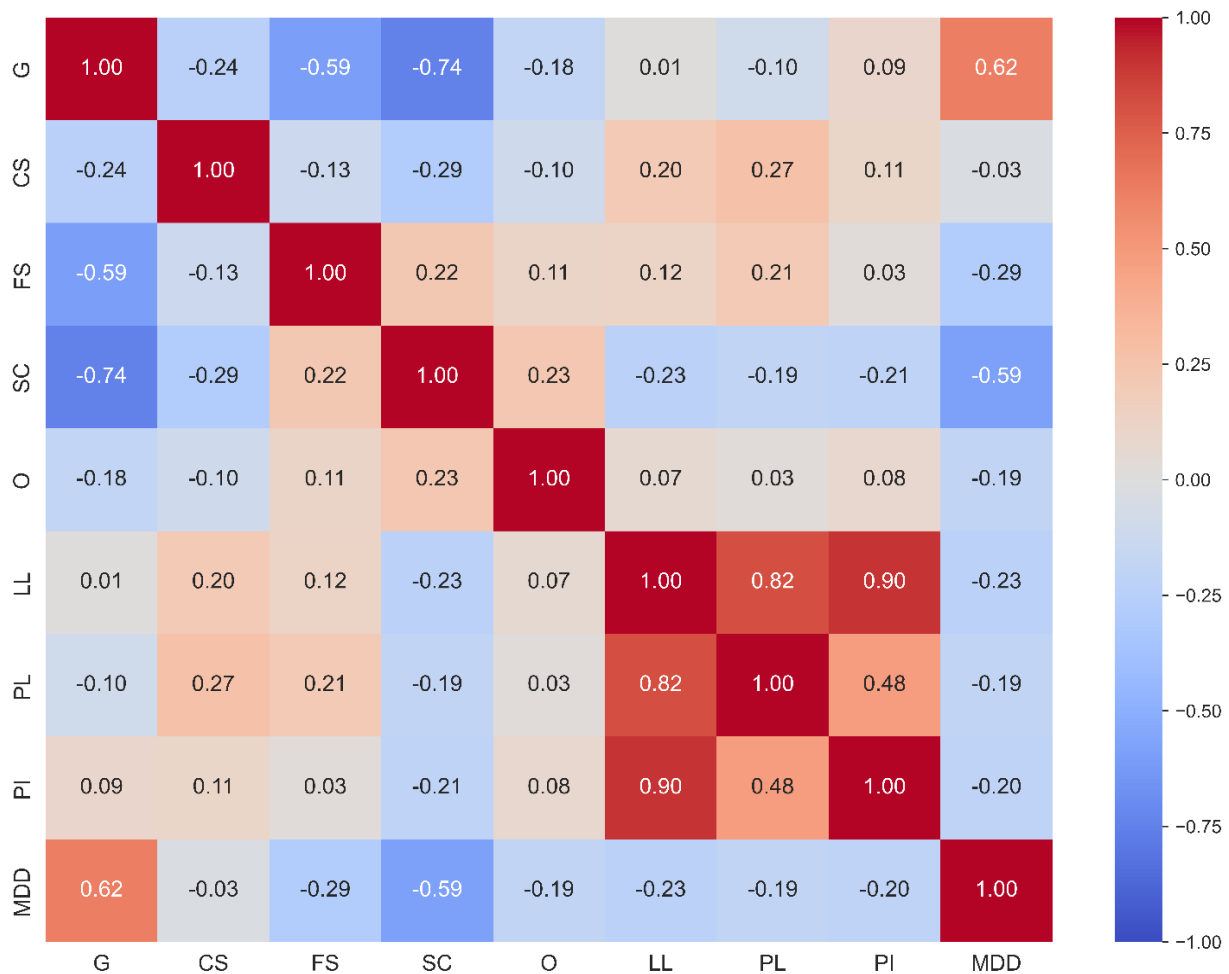
Phân tích ma trận tương quan trong Hình 2 cho thấy, các đặc trưng của đất có mối quan hệ và ảnh hưởng khác nhau đến giá trị MDD. Trong đó, hàm lượng sỏi là yếu tố có ảnh hưởng tích cực rõ

rệt nhất đến MDD, với hệ số tương quan đạt 0.62. Điều này phản ánh vai trò quan trọng của thành phần hạt lớn trong việc tăng khả năng đầm nén và giảm thể tích lỗ rỗng, từ đó làm tăng giá trị MDD.

Ngược lại, hàm lượng bụi sét có hệ số tương quan âm mạnh nhất với MDD (-0.59), cho thấy sự hiện diện nhiều của các hạt mịn trong đất làm tăng khả năng giữ nước và tính dẻo, từ đó cản trở việc đạt độ chặt tối ưu. Ngoài ra, các chỉ tiêu thể hiện tính dẻo của đất như giới hạn chảy, giới hạn dẻo và chỉ số dẻo cũng có mối tương quan âm nhẹ với MDD (dao động từ -0.20 đến -0.23), phản ánh đất càng

đều thì càng khó đầm chặt. Trong khi đó, các yếu tố như hàm lượng cát thô và hàm lượng hữu cơ có hệ số tương quan rất thấp, cho thấy ảnh hưởng của chúng đến MDD là không rõ rệt trong bộ dữ liệu này. Bên cạnh đó, các chỉ tiêu tính dẻo có tương quan cao với nhau, cho thấy mối liên hệ nội tại giữa các đặc trưng cơ lý của đất.

2.2. Phương pháp nghiên cứu



Hình 2. Ma trận tương quan của các biến được sử dụng trong mô hình

2.2.1. Light Gradient Boosting Machine (LGBM)

Thuật toán LGBM được Microsoft giới thiệu vào năm 2017, là một phương pháp học máy tiên tiến thuộc nhóm mô hình cây quyết định tăng cường độ dốc [14]. Về nguyên lý, LGBM huấn luyện các cây quyết định theo thứ tự tuần tự, trong đó mỗi cây mới được xây dựng nhằm giảm sai số còn lại từ các cây trước. Khác với các thuật toán boosting truyền thống, thuật toán chia nhánh theo lá (leaf-wise) thay vì chia theo mức (level-wise), cho phép tập trung vào các vùng dữ liệu có sai số

lớn, từ đó tăng độ chính xác của mô hình. Bên cạnh đó, LGBM sử dụng kỹ thuật lượng hóa dữ liệu liên tục thành các phân lớp (histogram binning), giúp giảm khối lượng tính toán và tăng tốc quá trình huấn luyện. Nhờ các cải tiến này, LGBM cho kết quả dự báo nhanh, chính xác và hiệu quả hơn so với các mô hình boosting truyền thống như XGBoost, đặc biệt khi xử lý dữ liệu lớn hoặc có số lượng đặc trưng cao. Ngoài tốc độ huấn luyện vượt trội, LGBM còn thể hiện khả năng tổng quát hóa tốt trên dữ liệu chưa từng thấy, giảm thiểu nguy cơ

quá khớp nhờ giới hạn độ sâu cây, tỷ lệ mẫu và tốc độ học. Trong thực tiễn, LGBM đã được chứng minh là phù hợp với nhiều bài toán như phân loại nhị phân, hồi quy phi tuyến với độ chính xác cao.

Công thức tổng quát mô tả nguyên lý hoạt động của mô hình LGBM trong bài toán hồi quy được mô tả như sau:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \quad (1)$$

Trong đó: $F_m(x)$: dự đoán của mô hình tại vòng lặp thứ m ; $F_{m-1}(x)$: dự đoán từ mô hình sau vòng lặp $m-1$; η : hệ số học (learning rate); $h_m(x)$: cây quyết định mới tại vòng lặp thứ m , được huấn luyện để tối thiểu hóa phần dư (residual).

2.2.2. Đánh giá độ chính xác mô hình

Đối với bài toán hồi quy, hiệu quả của mô hình được đánh giá thông qua các chỉ số phổ biến như hệ số xác định (R^2), sai số bình phương trung bình (RMSE) và sai số tuyệt đối trung bình (MAE). Trong đó, hệ số xác định R^2 phản ánh mức độ mà mô hình có thể giải thích được phương sai của dữ liệu thực tế, cho thấy mức độ phù hợp giữa giá trị dự báo và giá trị quan sát; giá trị này càng tiến gần đến 1 càng thể hiện mô hình có năng lực dự báo tốt. Sai số bình phương trung bình (RMSE) được sử dụng để đo lường độ lệch trung bình giữa giá trị thực và giá trị mô hình dự đoán, có ưu điểm thể hiện rõ mức độ sai số lớn nhờ việc làm nổi bật các điểm sai lệch lớn thông qua bình phương sai số. Đối với sai số tuyệt đối trung bình (MAE) thể hiện mức sai lệch trung bình tuyệt đối giữa giá trị thực tế và giá trị dự báo. Cả hai chỉ số RMSE và MAE đều càng nhỏ thì càng chứng tỏ mô hình hoạt động chính xác và ổn định. Việc kết hợp các chỉ số này giúp kiểm tra toàn diện tính chính xác, độ tin cậy và khả năng tổng quát hóa của mô hình trên cả dữ liệu huấn luyện và kiểm chứng.

Công thức tổng quát của các chỉ số như sau [15-18]:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

Trong đó, y_i là giá trị thực tế, $\hat{y}_i$ là giá trị dự báo, $\bar{y}$ là giá trị trung bình của dữ liệu thực tế.

2.2.3. Giải thích mô hình Shapley (SHAP)

SHAP được xây dựng dựa trên lý thuyết giá trị Shapley trong lý thuyết trò chơi, do Lloyd Shapley phát triển [19]. Nguyên lý cốt lõi của SHAP là ước lượng mức độ đóng góp của từng đặc trưng đầu vào bằng cách tính giá trị trung bình của các đóng góp cận của đặc trưng đó qua tất cả các tổ hợp đặc trưng có thể xảy ra. Nói cách khác, đối với mỗi dự báo, giá trị SHAP của một đặc trưng phản ánh mức độ mà đặc trưng đó làm thay đổi kết quả dự báo so với giá trị trung bình toàn bộ mô hình khi đặc trưng đó được đưa vào hoặc rút ra khỏi tập hợp đặc trưng.

Với nguyên lý tính toán nhất quán, SHAP mang lại nhiều ưu điểm nổi bật trong việc giải thích mô hình. Phương pháp này cung cấp cách tiếp cận trực quan và định lượng rõ ràng để xác định mức độ ảnh hưởng của từng đặc trưng đầu vào đối với kết quả dự báo. SHAP có thể áp dụng hiệu quả cho cả mô hình tuyến tính và phi tuyến, nhờ đó SHAP góp phần nâng cao tính minh bạch và khả năng kiểm chứng của mô hình, đồng thời hỗ trợ tốt trong việc đánh giá tầm quan trọng của các đặc trưng đầu vào.

Công thức tổng quát của SHAP được biểu diễn như sau:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (5)$$

Trong đó: $f(x)$ là giá trị dự báo của mô hình đối với mẫu x ; ϕ_0 là giá trị trung bình của mô hình khi không có đặc trưng nào; ϕ_i là giá trị SHAP của đặc trưng i , biểu thị mức độ đóng góp của đặc trưng này vào kết quả dự báo; M là số lượng đặc trưng trong mô hình.

2.2.4. Biểu đồ phụ thuộc từng phần (PDP)

PDP là một kỹ thuật trực quan hóa được đề xuất bởi Friedman (2001) nhằm hỗ trợ giải thích và phân tích tác động của một hoặc một nhóm biến

đầu vào đến kết quả dự báo của các mô hình học máy, đặc biệt là các mô hình phi tuyến hoặc khó giải thích như mô hình “hộp đen” [20]. Nguyên lý hoạt động của PDP dựa trên việc tính toán giá trị dự báo trung bình của mô hình ứng với từng giá trị cụ thể của biến đầu vào đang xét, bằng cách giữ nguyên các biến khác và thay thế giá trị của biến đó trên toàn bộ tập dữ liệu. Trung bình các giá trị dự báo thu được sẽ phản ánh mối quan hệ giữa biến đầu vào và đầu ra mục tiêu, từ đó giúp làm sáng tỏ mức độ ảnh hưởng của đặc trưng đến kết quả dự báo của mô hình.

Công thức tổng quát của (PDP) đối với một đặc trưng x_s được biểu diễn như sau:

$$\hat{f}_{x_s}(x_s) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_s, \mathbf{x}_i^C) \quad (6)$$

Trong đó: $\hat{f}_{x_s}(x_s)$ là giá trị phụ thuộc từng phần (partial dependence) tại giá trị x_s , $\hat{f}$ là mô hình học máy đã huấn luyện, $\mathbf{x}_i^C$ là các giá trị của tất cả các đặc trưng còn lại (ngoại trừ x_s) trong mẫu thứ i , n là số lượng mẫu trong tập dữ liệu.

2.2.5. Quy trình thực hiện

Trong quá trình xây dựng mô hình dự báo cho bài toán hồi quy, việc tuân thủ một quy trình thực hiện rõ ràng và khoa học là điều kiện tiên quyết nhằm đảm bảo độ tin cậy và hiệu quả của kết quả đầu ra. Quy trình này thường bao gồm 4 bước cơ bản sau:

Bước 1: Thu thập dữ liệu: Đảm bảo tính đại diện và số lượng mẫu nghiên cứu.

Bước 2: Tiền xử lý và phân chia dữ liệu: Dữ liệu được chuẩn hóa bằng Min-Max Scaler và phân tích mối quan hệ tương quan giữa các yếu tố ảnh hưởng. Dữ liệu được chia thành dữ liệu huấn luyện (70%) và dữ liệu kiểm chứng (30%) để đánh giá hiệu suất mô hình.

Bước 3: Lựa chọn và huấn luyện mô hình: Mô hình LGBM kết hợp với phương pháp GridSearch được sử dụng nhằm tối ưu hóa siêu tham số và nâng cao hiệu suất dự báo.

Bước 4: Đánh giá mô hình: Sử dụng các chỉ số đánh giá MAE, RMSE, R^2 để đo lường độ chính xác.

3. Kết quả và thảo luận

Trong nghiên cứu này, quá trình huấn luyện mô hình được thực hiện kết hợp với phương pháp tìm kiếm lưới nhằm xác định tổ hợp siêu tham số tối ưu. Sau quá trình tối ưu, các siêu tham số tốt nhất thu được bao gồm: số lượng cây ($n_estimators$) = 1500, tốc độ học ($learning_rate$) = 0.02, độ sâu tối đa của cây (max_depth) = 9 cùng với hai hệ số là reg_alpha (L1 regularization) = 0.1 và reg_lambda (L2 regularization) = 0.1. Việc áp dụng đồng thời Grid Search trong quá trình huấn luyện giúp đảm bảo rằng mô hình không những đạt được hiệu suất cao trên tập huấn luyện, mà còn có khả năng dự báo tốt trên tập kiểm chứng.

Bảng 2. Kết quả của mô hình LGBM

Tham số	R^2	MAE	RMSE	Std
Huấn luyện	0.941	0.02	0.028	0.103
Kiểm chứng	0.771	0.048	0.059	0.090

Kết quả đánh giá hiệu suất của mô hình LGBM trên tập huấn luyện và tập kiểm chứng được tổng hợp trong Bảng 2. Trên tập huấn luyện, mô hình đạt hiệu suất tốt với giá trị $R^2 = 0.941$, MAE = 0.020, RMSE = 0.028 cho thấy khả năng học tốt các mối quan hệ trong dữ liệu. Độ lệch chuẩn trên tập dữ liệu này ở mức thấp 0.103. Trong khi đó, trên tập kiểm chứng các chỉ số lần lượt là $R^2 = 0.771$, MAE = 0.048, RMSE = 0.059 cho thấy giá trị R^2 giảm và sai số tăng so với tập huấn luyện

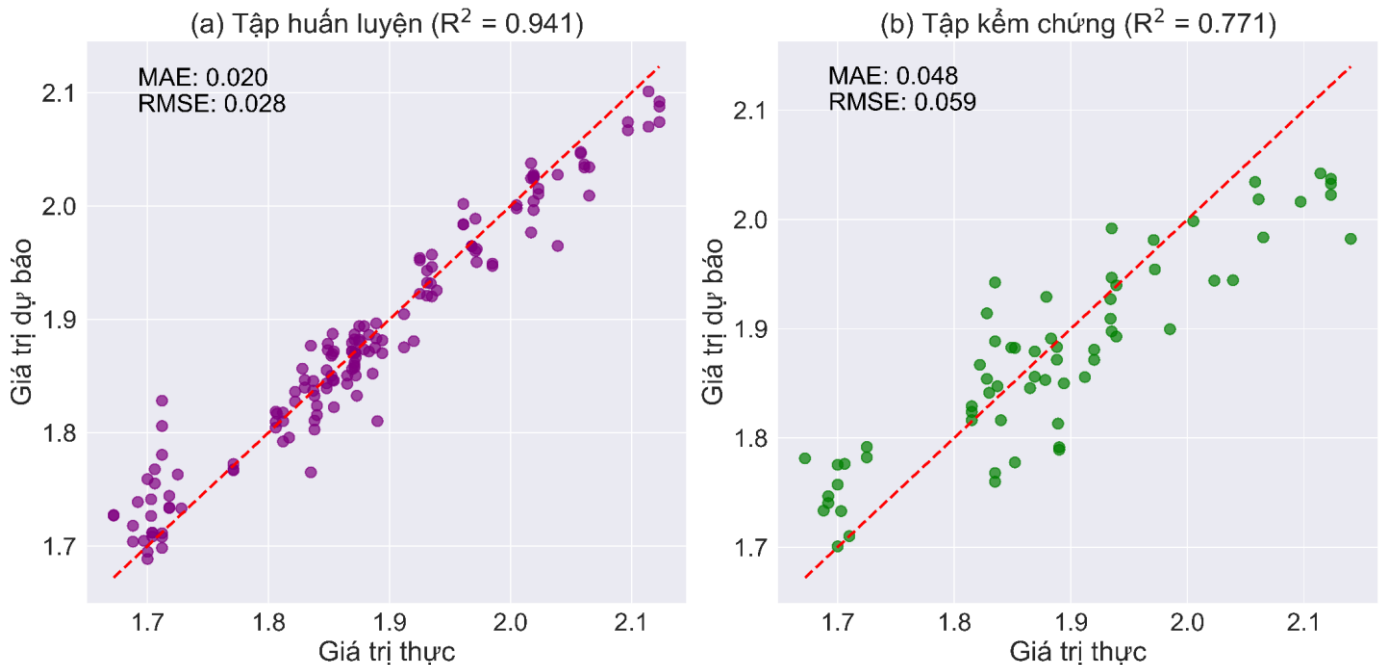
nhưng vẫn nằm trong khoảng giá trị chấp nhận được. Độ lệch chuẩn của sai số trên tập kiểm chứng là 0.090, sai khác không quá lớn so với tập huấn luyện cho thấy ít bị ảnh hưởng bởi sự biến động trong dữ liệu kiểm chứng.

Mối quan hệ giữa giá trị thực tế và giá trị dự báo của mô hình LGBM trên cả hai tập dữ liệu huấn luyện và kiểm chứng được thể hiện trực quan thông qua biểu đồ phân tán. Trong đó, các điểm dữ liệu càng nằm gần hoặc trùng với đường tham

chiếu (đường chéo $y = x$) thì mô hình càng thể hiện hiệu suất cao. Ở Hình 3(a), điểm dữ liệu từ tập huấn luyện phân bố tập trung dọc theo đường chéo, phản ánh khả năng học tốt của mô hình trong giai đoạn huấn luyện. Ngược lại, Hình 3(b) mô tả sự phân bố điểm trên tập kiểm chứng với độ phân tán lớn hơn, song phần lớn vẫn duy trì vị trí gần đường tham chiếu, cho thấy mức suy giảm hiệu suất không quá lớn.

Kết quả phân tích biểu đồ SHAP trong Hình 4 cho thấy, mức độ ảnh hưởng của từng đặc trưng đầu vào đến giá trị dự báo khối lượng thể tích khô lớn nhất (MDD) trong mô hình LGBM. Trục hoành biểu diễn giá trị SHAP – càng lệch phải chứng tỏ đặc trưng đó có xu hướng làm tăng MDD, trong khi lệch trái thể hiện xu hướng làm giảm MDD. Màu sắc từ xanh đến đỏ tương ứng với giá trị thấp đến cao của từng đặc trưng, giúp làm rõ hơn mối quan hệ giữa giá trị đầu vào và tác động lên đầu ra mô hình. Trong số các đặc trưng, hàm lượng sỏi (G) thể hiện ảnh hưởng rõ rệt và tích cực: khi G tăng (màu đỏ), giá trị SHAP cũng tăng theo, nghĩa là G

cao góp phần làm tăng MDD. Điều này phù hợp với nguyên lý cơ học đất, trong đó sỏi đóng vai trò khung liên kết, giúp giảm thể tích rỗng và tăng khả năng đầm nén. Đặc trưng CS cũng cho thấy ảnh hưởng tích cực trong một khoảng giá trị nhất định, tuy nhiên khi CS quá cao, ảnh hưởng này trở nên không rõ rệt hoặc có thể giảm. Ngược lại, SC và O có xu hướng làm giảm MDD khi giá trị tăng cao, phản ánh bản chất các hạt mịn và vật liệu hữu cơ thường giữ nước, làm giảm khả năng đầm chặt của đất. Các đặc trưng LL, PL, PI cũng cho thấy tác động tiêu cực rõ rệt đến MDD khi giá trị của chúng tăng, điều này phù hợp với thực tế rằng đất có tính dẻo cao thường giữ ẩm nhiều, dễ biến dạng và khó đạt được mật độ khô tối ưu. Nhìn chung, mô hình học máy không chỉ cung cấp dự báo chính xác mà còn phản ánh đúng bản chất cơ lý của từng đặc trưng, từ đó khẳng định tính hữu ích của SHAP trong việc giải thích mô hình và hỗ trợ phân tích, thiết kế nền móng công trình một cách khoa học và hiệu quả.



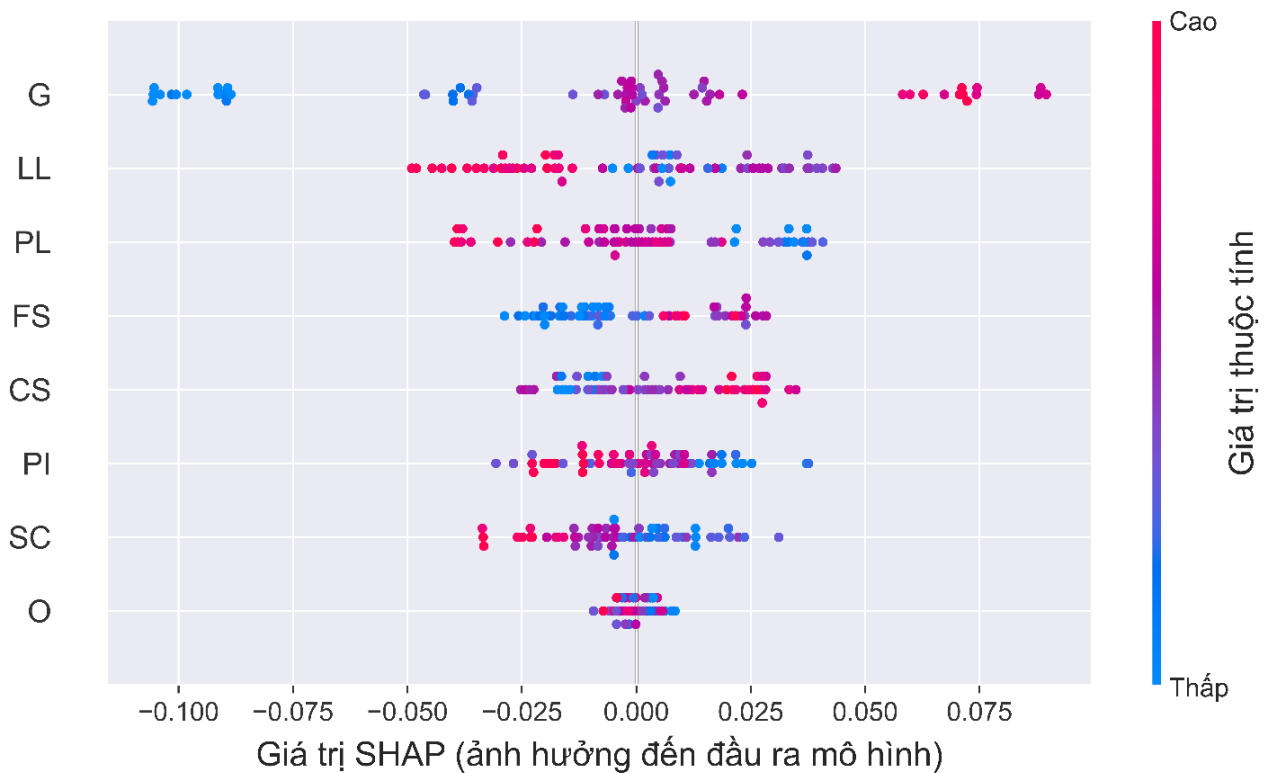
Hình 3. Biểu đồ phân tán giữa giá trị thực và giá trị dự báo

Kết quả phân tích biểu đồ phụ thuộc từng phần (PDP) trong Hình 5 giúp làm rõ ảnh hưởng của từng đặc trưng đất đến giá trị dự báo mật độ khô lớn nhất (MDD) trong mô hình LGBM. Nhìn chung, PDP cho thấy một số đặc trưng có tác động

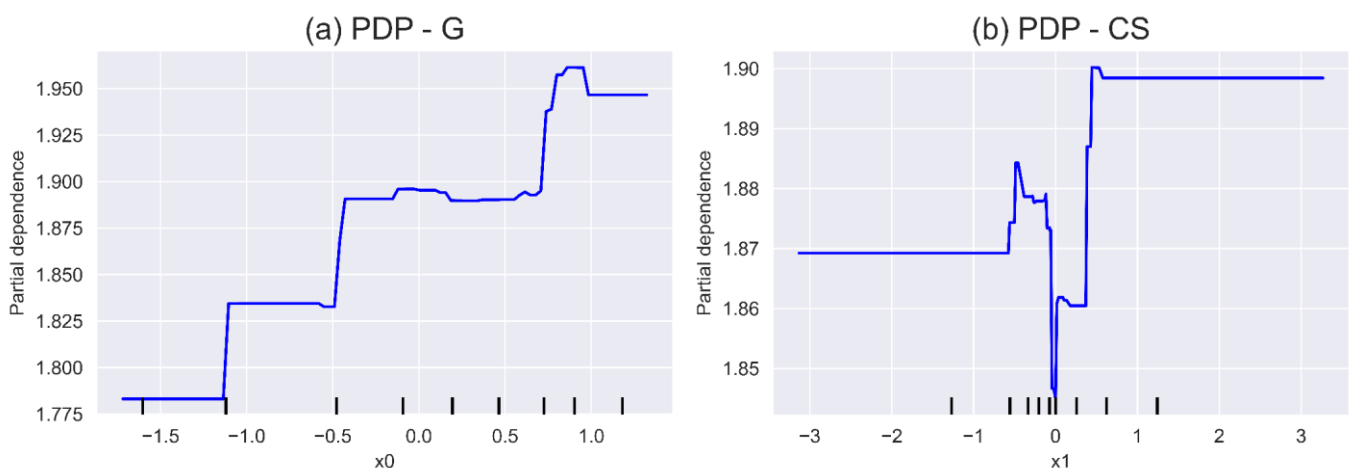
tích cực làm tăng MDD, trong khi các đặc trưng khác lại làm giảm giá trị này, phản ánh đúng quy luật của cơ học đất trong thực tế đầm nén. Cụ thể, hai yếu tố G và CS tỷ lệ thuận với MDD. Khi giá trị G tăng, MDD cũng tăng đáng kể do các hạt sỏi tạo

khung liên kết chắc chắn, giảm thể tích lỗ rỗng và tăng khả năng chịu nén. CS cũng giúp tăng MDD trong một khoảng giá trị tối ưu, tuy nhiên nếu vượt ngưỡng, hiệu quả sẽ bão hòa hoặc giảm do ảnh hưởng đến sự ổn định tổng thể của cấu trúc hạt. Ngược lại, các yếu tố như SC và O đều có ảnh hưởng tiêu cực đến MDD. Khi tăng SC làm giảm MDD do hạt mịn làm tăng khả năng giữ nước và giảm độ chặt. Giá trị O tăng cũng làm giảm nhẹ MDD do bản chất nhẹ và xốp của vật liệu hữu cơ, ảnh hưởng đến khả năng đầm chặt. Đối với các chỉ

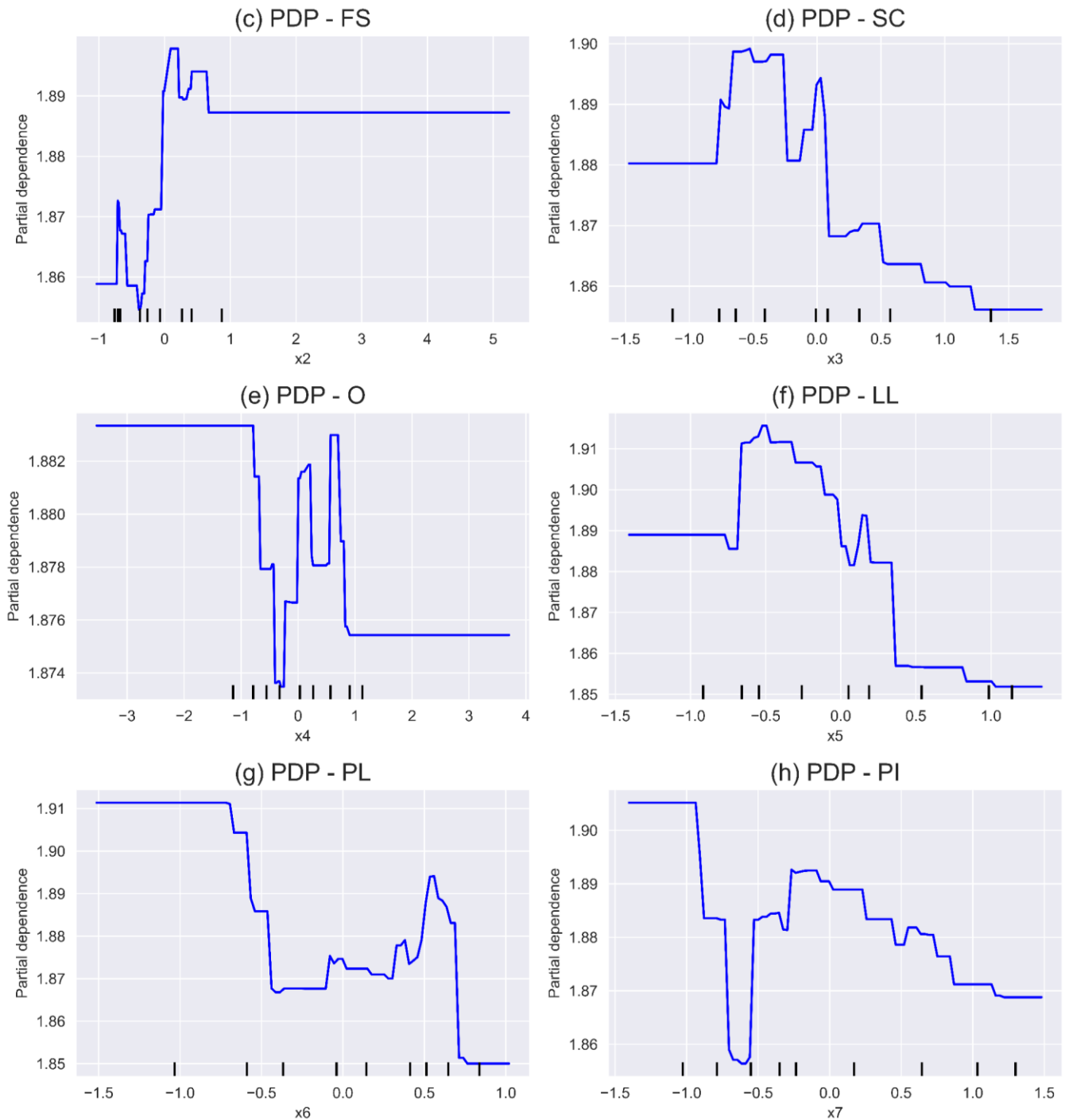
tiêu như LL, PL và PI, biểu đồ PDP cho thấy sự gia tăng các giá trị này đều làm giảm MDD. Điều này hoàn toàn hợp lý vì đất có tính dẻo cao thường giữ nước nhiều, dễ biến dạng và khó đạt được mật độ đầm chặt tối ưu. Từ những phân tích trên có thể thấy, mô hình học máy không chỉ cho kết quả dự báo chính xác mà còn phản ánh đúng bản chất cơ lý của các yếu tố ảnh hưởng đến MDD của đất. Điều này khẳng định tiềm năng của học máy như một công cụ hỗ trợ hiệu quả trong phân tích và thiết kế nền móng công trình.



Hình 4. Giải thích mô hình Shapley



Hình 5. Biểu đồ phụ thuộc từng phần



Hình 5. (tiếp)

4. Kết luận

Nghiên cứu áp dụng mô hình LGBM để dự báo MDD của đất tại tuyến cao tốc Vân Đồn – Móng Cái. Kết quả cho thấy mô hình dự báo tốt, với các chỉ số hiệu suất đạt $R^2 = 0.941$ và $RMSE = 0.028$ trên tập dữ liệu huấn luyện, cùng với $R^2 = 0.771$ và $RMSE = 0.059$ trên tập kiểm chứng, phản ánh độ chính xác và tính ổn định. Không chỉ dừng lại ở hiệu suất, LGBM cho thấy khả năng phân tích

tốt khi được kết hợp với các kỹ thuật giải thích như SHAP và biểu đồ phụ thuộc giúp nhận diện mức độ ảnh hưởng của từng yếu tố đặc trưng đầu vào. Với các đặc trưng như hàm lượng sỏi và cát thô đóng vai trò tích cực trong việc nâng cao MDD, trong khi các yếu tố khác lại góp phần làm giảm chỉ tiêu này, hoàn toàn phù hợp với cơ sở lý thuyết địa kỹ thuật. Nhờ đó, mô hình không chỉ dự báo chính xác mà còn hỗ trợ hiệu quả trong việc phân tích các yếu tố

ảnh hưởng, góp phần nâng cao chất lượng dự báo.

Tài liệu tham khảo

- [1] G.G. Tejani, B. Sadaghat, S. Kumar. (2023). Predict the maximum dry density of soil based on individual and hybrid methods of machine learning. *Advances in Engineering and Intelligence Systems*, 2(3), 95-106. <https://doi.org/10.22034/aeis.2023.414188.1129>
- [2] NG K.S., CHEW Y.M., OSMAN M.H., MOHAMAD GHAZALI S.K. (2015). Estimating maximum dry density and optimum moisture content of compacted soils. *International Conference on Advances in Civil and Environmental Engineering*, 2015.
- [3] A. Baghbani, T. Choudhury, S. Costa, J. Reiner. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. *Earth-Science Reviews*, 228, 103991. <https://doi.org/10.1016/j.earscirev.2022.103991>
- [4] M.N. Duc, A.H. Sy, T.N. Ngoc, T.L.H. Thi. (2022). An artificial intelligence approach based on multi-layer perceptron neural network and random forest for predicting maximum dry density and optimum moisture content of soil material in Quang Ninh Province, Vietnam. *CIGOS 2021, Emerging Technologies and Applications for Green Infrastructure. Lecture Notes in Civil Engineering*, vol 203. Springer, Singapore. https://doi.org/10.1007/978-981-16-7160-9_176
- [5] W.Z. Taffese, K.A. Abegaz. (2021). Artificial intelligence for prediction of physical and mechanical properties of stabilized soil for affordable housing. *Applied Sciences*, 11(16), 7503. <https://doi.org/10.3390/app11167503>
- [6] S.K. Sinha, M.C. Wang. (2008). Artificial neural network prediction models for soil compaction and permeability. *Geotechnical and Geological Engineering*, 26, 47-64. <https://doi.org/10.1007/s10706-007-9146-3>
- [7] B. Ghorbani, A. Arulrajah, G. Narsilio, S. Horpibulsuk, M.W. Bo. (2020). Development of genetic-based models for predicting the resilient modulus of cohesive pavement subgrade soils. *Soils and Foundations*, 60(2), 398-412. <https://doi.org/10.1016/j.sandf.2020.02.010>
- [8] M.N. Navidi, J. Seyedmohammadi, S.A.S. Jalali. (2022). Predicting soil water content using support vector machines improved by meta-heuristic algorithms and remotely sensed data. *Geomechanics and Geoengineering*, 17(3), 712-726. <https://doi.org/10.1080/17486025.2020.1864032>
- [9] B.T. Pham, M.D. Nguyen, N. Al-Ansari, Q.A. Tran, L.S. Ho, H.V. Le, I. Prakash. (2021). A comparative study of soft computing models for prediction of permeability coefficient of soil. *Mathematical Problems in Engineering*, 2021(1), 7631493. <https://doi.org/10.1155/2021/7631493>
- [10] S. Maroufpoor, E. Maroufpoor, O. Bozorg-Haddad, J. Shiri, Z.M. Yaseen. (2019). Soil moisture simulation using hybrid artificial intelligent model: Hybridization of adaptive neuro fuzzy inference system with grey wolf optimizer algorithm. *Journal of Hydrology*, 575, 544-556. <https://doi.org/10.1016/j.jhydrol.2019.05.045>
- [11] O. Günaydın. (2009). Estimation of soil compaction parameters by using statistical analyses and artificial neural networks. *Environmental Geology*, 57, 203-215. <https://doi.org/10.1007/s00254-008-1300-6>
- [12] D.K. Trong, B.T. Pham, F.E. Jalal, M. Iqbal, P.C. Roussis, A. Mamou, M. Ferentinou, D.Q. Vu, N.D. Dam, Q.A. Tran, P.G. Asteris. (2021). On random subspace optimization-based hybrid computing models predicting the california bearing ratio of soils. *Materials*, 14(21), 6516. <https://doi.org/10.3390/ma14216516>
- [13] B. Li, Z. You, K. Ni, Y. Wang. (2024). Prediction of Soil Compaction parameters using machine learning models. *Applied Sciences*, 14(7), 2716. <https://doi.org/10.3390/app14072716>
- [14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu. (2017). LightGBM: A

- highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3149-3157.
- [15] M.V. Chiến. (2024). Dự đoán sức kháng cắt của dầm FRP-RC bằng mô hình học máy CatBoost tối ưu hóa. *Tạp chí điện tử Khoa học và Công nghệ Giao thông*, 4(3), 13-27. <https://doi.org/10.58845/jstt.utt.2024.vn.4.3.13-27>
- [16] B.G. Linh, V.Đ. Phiên. (2024). Dự báo cường độ bê tông nhẹ bằng mô hình stacking và shap. *Tạp chí điện tử Khoa học và Công nghệ Giao thông*, 4(3), 48-60. <https://doi.org/10.58845/jstt.utt.2024.vn.4.3.48-60>
- [17] M.T.H. Vân, P.V. Huỳnh. (2025). Dự báo độ dẻo Marshall của hỗn hợp bê tông nhựa chặt bằng mô hình tăng cường độ dốc. *Tạp chí điện tử Khoa học và Công nghệ Giao thông*, 5(2), 56-68. <https://doi.org/10.58845/jstt.utt.2025.vn.5.2.56-68>
- [18] P.T. Nghị. (2025). Nghiên cứu dự đoán cường độ chịu nén của bê tông sử dụng tro bay bằng mô hình học máy XGBoost. *Tạp chí điện tử Khoa học và Công nghệ Giao thông*, 5(2), 99-106. <https://doi.org/10.58845/jstt.utt.2025.vn.5.2.99-106>
- [19] S.M. Lundberg, S.-I. Lee. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768-4777.
- [20] J.H. Friedman. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>