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

NovaLog AI: A machine learning system for travel time prediction and logistics routing support

Trong Dung Nguyen*, Van Hai Nguyen, Van Anh Trieu, Thi Hanh Do
Research group on Industry 4.0 in Transportation (I4T group), University of Transport Technology, 54 Trieu Khuc Street, Thanh Liet Ward, Hanoi, Vietnam

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.vn.6.4.26-48

***Corresponding author:**
Email address:
dungnt88@utt.edu.vn

Received: 09/03/2026
Received in Revised Form: 10/04/2026
Accepted: 14/04/2026

Abstract: Logistics costs in Vietnam currently account for approximately 16.8% of the gross domestic product (GDP), significantly higher than the global average due to infrastructure limitations and the fragmentation and lack of synchronization in data. To address this issue, the study proposes the development of the NovaLog AI system to predict travel time and optimize management processes at freight transit stations. The research utilizes a multidimensional empirical dataset impacting travel time, including spatial, temporal, and climatic features. Eight popular machine learning algorithms were implemented and their performance compared using a cross-validation method combined with 30 Monte Carlo simulation runs to verify stability. The final evaluation results based on the R^2 and RMSE metrics indicate that XGBoost is the most accurate forecasting model, with R^2 reaching 0.9879 (RMSE = 1.1557) on the training set and R^2 reaching 0.9375 (RMSE = 2.7585) on the testing set. Simultaneously, through ranking variable importance, the study identified the four factors with the strongest impact on prediction performance: distance, wind direction, departure longitude, and departure latitude. The NovaLog AI system provides a practical scientific basis for enterprises in applying artificial intelligence to support decision-making in selecting optimal routes, improving supply chain operational efficiency, and reducing logistics costs in Vietnam.

Keywords: Logistics, travel time prediction, machine learning, XGBoost, Monte Carlo.

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

NovaLog AI: Hệ thống học máy dự báo thời gian di chuyển và hỗ trợ định tuyến Logistic

Nguyễn Trọng Dũng*, Nguyễn Văn Hải, Triệu Vân Anh, Đỗ Thị Hạnh
Nhóm nghiên cứu Ứng dụng công nghệ 4.0 trong Giao thông vận tải (I4T), Trường Đại học Công nghệ Giao thông vận tải, 54 Triều Khúc, Thanh Liệt, Hà Nội, Việt Nam

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

DOI:
https://doi.org/10.58845/jstt.utt.2026.vn.6.4.26-48

***Tác giả liên hệ:**
Địa chỉ Email:
dungnt88@utt.edu.vn

Ngày nộp bài: 09/03/2026
Ngày nộp bài sửa: 10/04/2026
Ngày chấp nhận: 14/04/2025

Tóm tắt: Chi phí logistics tại Việt Nam hiện chiếm khoảng 16.8% tổng sản phẩm quốc nội (GDP), cao hơn đáng kể so với mức trung bình toàn cầu do hạn chế về hạ tầng và sự phân mảnh, thiếu đồng bộ của dữ liệu. Để giải quyết bài toán này, nghiên cứu đề xuất xây dựng hệ thống NovaLog AI nhằm dự báo thời gian di chuyển và tối ưu hóa quy trình quản lý tại các trạm trung chuyển hàng hóa. Nghiên cứu sử dụng một bộ dữ liệu thực chứng đa chiều tác động đến thời gian di chuyển, bao gồm các đặc trưng về không gian, thời gian và khí hậu. Tám thuật toán học máy phổ biến đã được triển khai và so sánh hiệu suất thông qua phương pháp xác thực chéo kết hợp với 30 lần chạy mô phỏng Monte Carlo để kiểm định tính ổn định. Kết quả đánh giá cuối cùng dựa trên các chỉ số R^2 và RMSE chỉ ra rằng XGBoost là mô hình dự báo chính xác nhất, với R^2 đạt 0.9879 (RMSE = 1.1557) trên tập huấn luyện và R^2 đạt 0.9375 (RMSE = 2.7585) trên tập kiểm tra. Đồng thời, thông qua việc xếp hạng độ quan trọng của các biến, nghiên cứu xác định được 4 yếu tố có ảnh hưởng mạnh mẽ nhất đến hiệu suất dự báo là: quãng đường, hướng gió, kinh độ đi, và vĩ độ đi. Hệ thống NovaLog AI cung cấp cơ sở khoa học thiết thực cho các doanh nghiệp trong việc ứng dụng trí tuệ nhân tạo để hỗ trợ việc ra quyết định lựa chọn lộ trình tối ưu, cải thiện hiệu quả vận hành chuỗi cung ứng và cắt giảm chi phí logistics tại Việt Nam.

Từ khóa: Logistics, dự báo thời gian, học máy, XGBoost, Monte Carlo.

1. Đặt vấn đề

Trong bối cảnh toàn cầu hóa và sự bùng nổ của thương mại điện tử sau đại dịch, nhu cầu vận tải hàng hóa gia tăng đột biến đang đặt ra thách thức chưa từng có về quy hoạch hạ tầng và nguồn lực logistics [1]. Trên bình diện quốc tế, các tập đoàn bán lẻ hàng đầu đã chứng minh được hiệu quả vượt trội của trí tuệ nhân tạo. Điển hình như Walmart đã ứng dụng trí tuệ nhân tạo (AI) vào tự động hóa chuỗi cung ứng và đàm phán, giúp cải thiện mức chi phí đơn vị trung bình lên tới 20% [2]. Tuy nhiên, tại Việt Nam, bài toán tối ưu hóa vận hành đang trở nên vô cùng bức thiết khi chi phí logistics ước tính chiếm từ 16.8% đến 18% tổng sản phẩm quốc nội (GDP), vượt xa mức trung bình của thế giới và các nước phát triển [3,4].

Mặc dù ngành logistics Việt Nam giữ vững tốc độ tăng trưởng 14-16% mỗi năm [5], năng lực cạnh tranh quốc gia lại có dấu hiệu chững lại. Cụ thể, chỉ số hiệu suất logistics của Việt Nam đã tụt 4 bậc trong năm 2023, rơi xuống vị trí thứ 43 toàn cầu do sự suy giảm về năng lực logistics, tính kịp

thời và khả năng truy xuất. Sự yếu kém này xuất phát từ cấu trúc ngành phân mảnh: hiện có hơn 30,000 doanh nghiệp logistics, nhưng hơn 97% trong số đó là các doanh nghiệp vi mô, vừa và nhỏ, chỉ nắm giữ khoảng 30% thị phần. Bên cạnh đó, các khảo sát thực chứng chỉ ra lỗ hổng nghiêm trọng về năng lực ứng dụng công nghệ của đội ngũ quản lý [6]. Cơ sở dữ liệu chuỗi cung ứng tại Việt Nam hiện nay phân tán, thiếu đồng bộ, có độ chính xác thấp và chủ yếu vẫn được xử lý thủ công qua các nền tảng cơ bản như Excel hay Google Drive. Hệ quả là tỷ lệ ứng dụng công nghệ AI trong ngành logistics Việt Nam còn rất hạn chế, chỉ đạt 16,9% trong quản lý đơn hàng và 10,8% trong logistics thương mại điện tử [5]. Bên cạnh áp lực về chi phí và năng lực vận hành, ngành vận tải đường bộ Việt Nam còn đang tạo ra gánh nặng lớn lên môi trường khi tiêu thụ 80% lượng nhiên liệu và chiếm 23% lượng khí thải CO2. Đối với phần lớn các doanh nghiệp vi mô, nhỏ và vừa, việc chuyển đổi xanh về mặt vật lý như đầu tư đồng loạt đội xe điện là bất khả thi do chi phí đầu tư ban đầu quá lớn và rào cản về trạm sạc [4]. Do đó, các nghiên cứu gần đây khẳng định rằng chuyển đổi số thông qua các hệ thống AI chính là động lực nội tại then chốt nhất để các doanh nghiệp vừa nâng cao hiệu suất vận hành, vừa thực hành logistics xanh thông qua việc tối ưu hóa quãng đường và giảm thiểu phát thải [7]. Việc ứng dụng AI không chỉ giúp dự báo thời gian di chuyển, cung cấp tham số đầu vào hỗ trợ việc ra quyết định lộ trình mà còn tác động trực tiếp đến tỷ suất sinh lời – yếu tố chiếm tỷ trọng áp đảo 35.48% trong cấu trúc đánh giá hiệu suất tài chính của doanh nghiệp [8].

Dù có tiềm năng lớn, các nghiên cứu về mô hình hóa nhu cầu vận tải tại Việt Nam vẫn tồn tại nhiều khoảng trống. Phần lớn các mô hình học máy hiện nay gặp khó khăn do thiếu hụt nguồn dữ liệu lịch sử trực tiếp, buộc phải phụ thuộc vào các chỉ số kinh tế, xã hội mang tính vĩ mô. Hơn nữa, trên bình diện học thuật, có một sự thiếu hụt các mô hình dự báo thời gian di chuyển ngắn hạn với độ phân giải không gian vi mô – vốn là yếu tố cốt lõi phục vụ cho quyết định định tuyến và vận hành của các nhà quản trị doanh nghiệp [1]. Hầu hết các ứng dụng công nghệ mới chỉ dừng lại ở mức độ nghiên cứu lý thuyết hoặc thử nghiệm các công cụ AI đơn lẻ mà chưa có một hệ thống tích hợp toàn diện [9].

Xuất phát từ những hạn chế nêu trên, nghiên cứu này đề xuất xây dựng hệ thống NovaLog AI – một nền tảng trí tuệ nhân tạo chuyên biệt nhằm dự báo thời gian di chuyển và cung cấp tham số đầu vào hỗ trợ việc ra quyết định lộ trình tại các trạm trung chuyển hàng hóa thương mại điện tử (như Shopee, Lazada). Điểm khác biệt mang tính đột phá của nghiên cứu là việc thiết lập thành công một bộ dữ liệu thực chứng vi mô toàn diện, tích hợp 13 đặc trưng đầu vào phức tạp bao gồm cả không gian, thời gian và điều kiện khí tượng học (lượng mưa, nhiệt độ, tốc độ gió). Thông qua việc đánh giá 8 thuật toán học máy tiên tiến và sử dụng kỹ thuật mô phỏng Monte Carlo 30 lần để kiểm định độ ổn định của các mô hình, nghiên cứu lựa chọn XGBoost làm mô hình dự báo. Kết quả của công trình không chỉ đóng góp một khung phương pháp luận vững chắc cho bài toán dự báo logistics ngắn hạn, mà còn cung cấp một giải pháp công nghệ giúp các doanh nghiệp logistics tại Việt Nam khỏa lấp khoảng trống năng lực số, tối ưu hóa chi phí vận hành và định hình chuỗi cung ứng xanh bền vững.

2. Cơ sở dữ liệu

Để phục vụ cho quá trình xây dựng và đánh giá các mô hình dự báo thời gian di chuyển, một bộ dữ liệu bao gồm thông tin về hành trình di chuyển thực tế kết hợp với các yếu tố khí tượng đã được thu thập và xây dựng. Bộ dữ liệu này được sử dụng nhằm nghiên cứu mối quan hệ giữa các đặc điểm hành trình, vị trí địa lý và điều kiện thời tiết với thời gian di chuyển thực tế, từ đó làm cơ sở cho việc phát triển các mô hình học máy trong bài toán dự báo thời gian di chuyển. Dữ liệu được thu thập tại khu vực nội thành Hà Nội, Việt Nam trong tháng 12, thời điểm mùa đông với đặc trưng nhiệt độ thấp, độ ẩm cao và đôi khi xuất hiện mưa phùn.

Sau quá trình trích xuất và tiền xử lý, tập dữ

liệu gồm 20,493 bản ghi, mỗi bản ghi đại diện cho một hành trình di chuyển giữa điểm xuất phát và điểm đến. Tập dữ liệu bao gồm 14 biến số đầu vào và đầu ra mô tả đặc điểm của hành trình và điều kiện môi trường tại thời điểm di chuyển. Các biến đầu vào bao gồm tọa độ địa lý của điểm xuất phát và điểm đến (kinh độ và vĩ độ), thông tin về thời điểm bắt đầu hành trình và khoảng cách di chuyển, cho phép mô hình xem xét sự khác biệt của điều kiện giao thông theo từng thời điểm trong ngày. Bên cạnh đó, dữ liệu còn tích hợp các yếu tố khí tượng tại thời điểm hành trình diễn ra, bao gồm lượng mưa, độ ẩm, tốc độ gió, hướng gió, cùng với nhiệt độ thấp nhất và nhiệt độ cao nhất trong ngày. Những yếu tố khí tượng này có thể ảnh hưởng đáng kể đến tình trạng giao thông và làm thay đổi thời gian di chuyển của các phương tiện. Trong nghiên cứu này, thời gian di chuyển thực tế của hành trình được sử dụng làm biến mục tiêu, là giá trị cần được dự báo dựa trên các đặc trưng đầu vào.

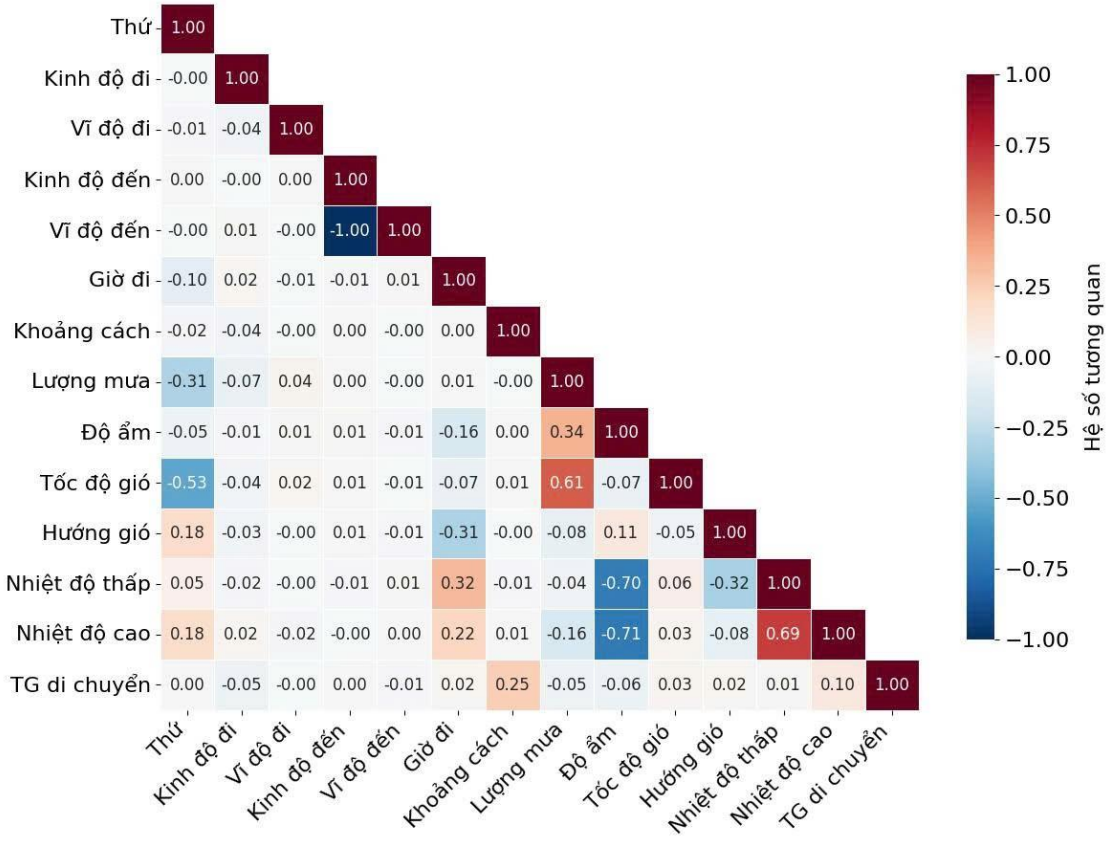
Để xây dựng và đánh giá mô hình, tập dữ liệu được chia ngẫu nhiên thành hai phần, trong đó 70% dữ liệu được sử dụng để huấn luyện mô hình và 30% dữ liệu còn lại được sử dụng để kiểm chứng và đánh giá hiệu suất dự báo. Thống kê mô tả chi tiết của các đặc trưng được tổng hợp tại Bảng 1.

Bảng 1. Thống kê mô tả bộ dữ liệu trong nghiên cứu này

Đặc trưng	Đơn vị	Trung bình	Độ lệch chuẩn	Nhỏ nhất	Q1	Q2	Q3	Lớn nhất
Thứ (ngày trong tuần từ 2 – 8)		5.69	1.88	2	4	6	7	8
Kinh độ đi	Độ (°)	105.81	0.04	105.53	105.79	105.81	105.83	105.94
Vĩ dộ đi	Độ (°)	21.18	3.75	20.93	21	21.02	21.03	105.78
Kinh độ đến	Độ (°)	105.8	0.04	105.49	105.79	105.81	105.83	106
Vĩ độ đến	Độ (°)	21.02	0.03	20.92	21	21.02	21.03	21.18
Giờ đi	phút	753.01	443.85	0	352	799	1190	1439
Quãng đường	km	6.98	4.59	0.17	3.9	5.9	8.9	82
Lượng mưa	mm	0.05	0.22	0	0	0	0	1
Độ ẩm	%	65.78	15.18	33	57	70	76	92
Tốc độ gió	km/h	8.15	4.65	4	5	7	9	92
Hướng gió (chỉ số hướng từ 0 – 8)		1.73	2.97	0	0	0	2	8
Nhiệt độ thấp	°C	16.97	3.21	12	14	17	20	24
Nhiệt độ cao	°C	21.23	3.35	15	18	21	24	27
Thời gian di chuyển	Phút	18.53	10.69	1	11	17	24	336

Bên cạnh đó, mối tương quan giữa các thông số đầu vào và giữa thông số đầu vào với các thông số đầu ra được thực hiện và thể hiện thông qua ma trận tương quan Pearson giữa các biến (Hình 1). Kết quả cho thấy phần lớn các biến độc lập có hệ số tương quan rất thấp (dao động từ -0.1 đến 0.1), đảm bảo tính độc lập và hạn chế nhiễu thông tin khi huấn luyện mô hình học máy. Đối với biến mục tiêu “Thời gian di chuyển”, “Quãng đường” là yếu tố có tương quan tuyến tính dương cao nhất (0.25). Các biến khí tượng thể hiện tương quan tuyến tính yếu với thời gian di chuyển, nhưng phản ánh đúng quy luật tự nhiên. Tuy nhiên, các yếu tố thời tiết vẫn có thể tác động gián tiếp đến tình trạng giao thông và có thể được khai thác tốt hơn bởi các mô hình học máy phi tuyến. Ngoài ra, có thể nhận thấy một số cặp biến khí tượng có mức tương quan tương đối cao, chẳng hạn như nhiệt độ thấp nhất và nhiệt độ cao nhất trong ngày với hệ số tương quan khoảng 0.69**,** hoặc độ ẩm và nhiệt độ với hệ số

tương quan âm khá lớn (khoảng -0.70). Điều này phản ánh mối quan hệ tự nhiên giữa các yếu tố khí tượng trong cùng một ngày. Nhìn chung, ma trận tương quan cho thấy các biến trong tập dữ liệu không có hiện tượng tương quan tuyến tính quá mạnh, điều này là thuận lợi cho quá trình huấn luyện các mô hình dự báo vì giúp giảm nguy cơ dư thừa thông tin giữa các đặc trưng.



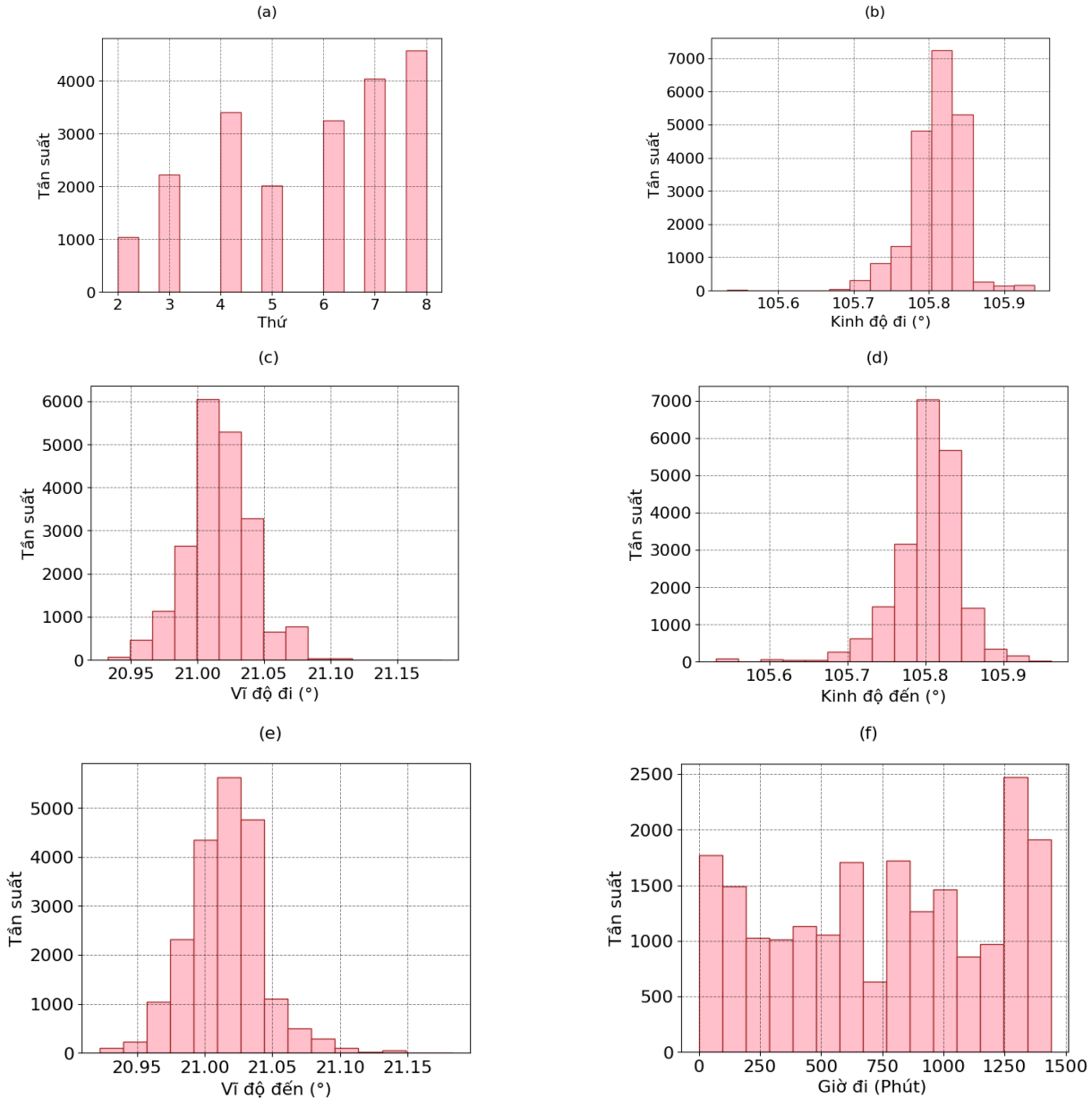
Hình 1. Ma trận tương quan giữa các biến đầu vào, đầu ra trong nghiên cứu này

Cấu trúc phân phối của các biến được trực quan hóa thông qua hệ thống biểu đồ tần suất (Hình 2). Phân tích đồ thị cho thấy bộ dữ liệu phản ánh rất sát đặc thù vận tải hàng hóa đô thị: phần lớn các hành trình có cự ly ngắn dưới 15 km, thời gian hoàn thành dưới 40 phút và tọa độ xuất phát/đến tập trung trong một độ phân giải không gian hẹp. Tương tự, dữ liệu khí tượng cũng hội tụ tại các giá trị mang tính đại diện cho thời tiết mùa đông miền Bắc: lượng mưa thấp, tốc độ gió nhẹ, độ ẩm cao. Nhìn chung, tính không đồng nhất trong phân phối dữ liệu cho thấy sự cần thiết phải ứng dụng các thuật toán học máy có khả năng xử lý các quan hệ phi tuyến tính phức tạp ở bước tiếp theo.

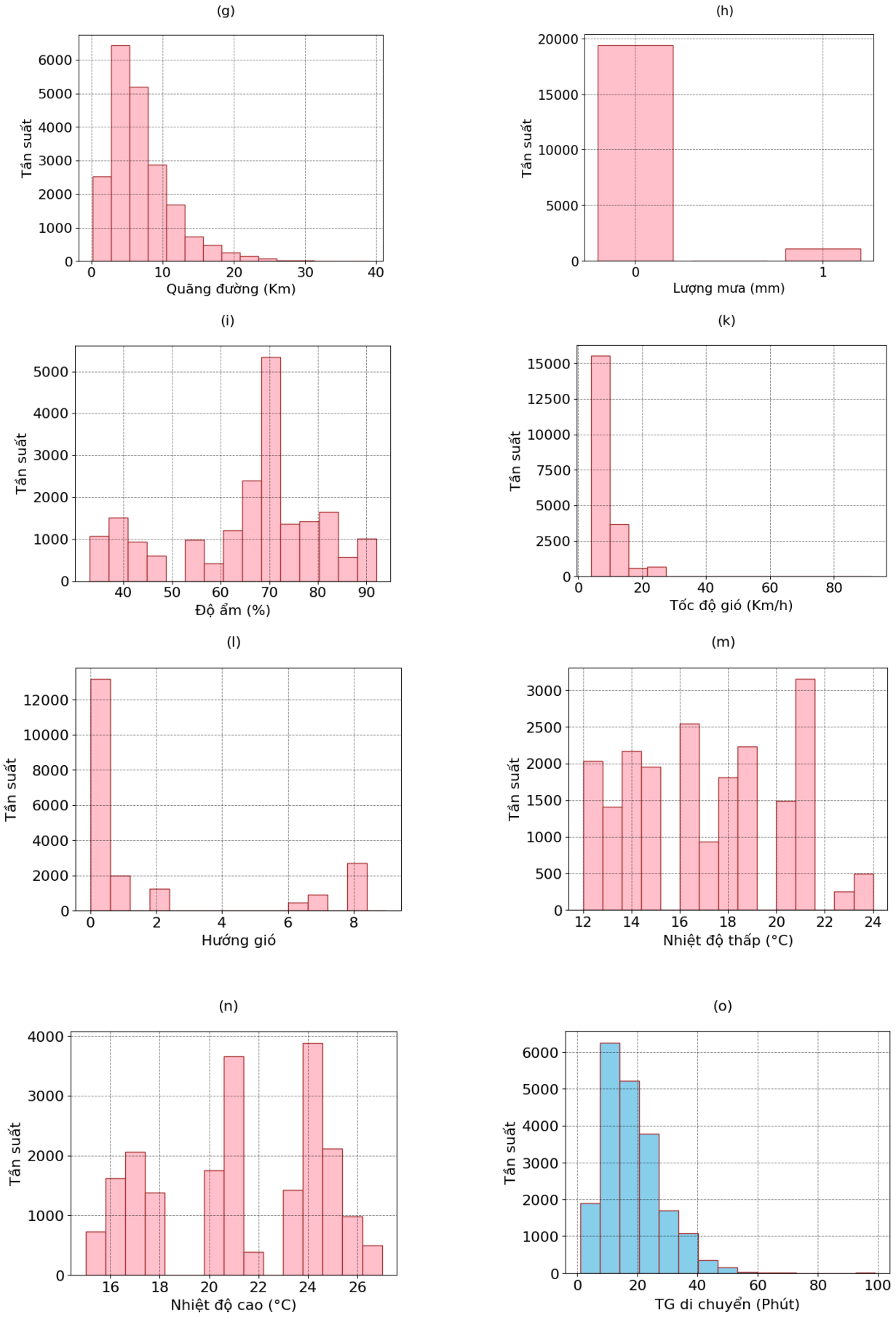
Hình 2 trình bày phân bố tần suất của các biến trong tập dữ liệu thông qua các biểu đồ histogram. Kết quả cho thấy các biến có những đặc điểm phân bố khác nhau, phản ánh đặc trưng của dữ liệu hành trình và điều kiện khí tượng tại khu vực nghiên cứu.Các biến tọa độ địa lý của điểm xuất phát và điểm đến (kinh độ và vĩ độ) có phân bố tập trung trong một khoảng giá trị hẹp, điều này phù hợp với thực tế vì dữ liệu được thu thập trong phạm vi thành phố Hà Nội. Phần lớn các giá trị tập trung quanh các tọa độ đặc trưng của khu vực nghiên cứu. Biến thời gian xuất phát có phân bố khá rộng, cho thấy các hành trình diễn ra ở nhiều thời điểm khác nhau trong ngày. Điều này giúp mô hình có khả năng học được sự thay đổi của điều kiện giao thông theo từng khoảng thời gian. Đối với quãng đường di chuyển, phần lớn các hành trình có khoảng cách tương đối ngắn, chủ yếu tập trung ở mức dưới khoảng 15 km. Điều này phản ánh đặc trưng của các chuyến di chuyển trong khu vực đô thị, nơi các hành trình thường có khoảng cách không quá lớn.

Các biến khí tượng như lượng mưa, độ ẩm, tốc độ gió, hướng gió và nhiệt độ cũng thể hiện các đặc điểm phân bố khác nhau. Biến lượng mưa chủ yếu tập trung ở giá trị thấp, cho thấy phần lớn thời gian trong tập dữ liệu không có mưa hoặc lượng mưa nhỏ. Độ ẩm có xu hướng phân bố ở mức tương đối cao, phù hợp với đặc điểm khí hậu mùa đông tại Hà Nội. Trong khi đó, tốc độ gió chủ yếu tập trung ở các giá trị thấp, cho thấy điều kiện gió mạnh ít xuất hiện trong khoảng thời gian thu thập dữ liệu. Hai biến nhiệt độ thấp nhất và nhiệt độ cao nhất trong ngày có phân bố tập trung trong khoảng từ khoảng 13°C đến 27°C, phù hợp với điều kiện thời tiết mùa đông tại khu vực nghiên cứu.

Đối với biến mục tiêu là thời gian di chuyển, phần lớn các giá trị tập trung trong khoảng dưới 40 phút, cho thấy đa số các hành trình trong tập dữ liệu có thời gian di chuyển tương đối ngắn. Tuy nhiên vẫn tồn tại một số giá trị lớn hơn, phản ánh các trường hợp di chuyển dài hơn hoặc có thể bị ảnh hưởng bởi tình trạng giao thông. Nhìn chung, các biểu đồ phân bố cho thấy dữ liệu có đặc điểm phù hợp với bối cảnh giao thông đô thị tại Hà Nội, đồng thời cung cấp thông tin quan trọng giúp hiểu rõ cấu trúc dữ liệu trước khi tiến hành xây dựng các mô hình dự báo.



Hình 2. Phân phối dữ liệu các biến đầu vào, đầu ra được sử dụng trong nghiên cứu này


(g)
Tần suất
Quãng đường (Km)
(h)
Tần suất
Lượng mưa (mm)
(i)
Tần suất
Độ ẩm (%)
(k)
Tần suất
Tốc độ gió (Km/h)
(l)
Tần suất
Hướng gió
(m)
Tần suất
Nhiệt độ thấp (°C)
(n)
Tần suất
Nhiệt độ cao (°C)
(o)
Tần suất
TG di chuyển (Phút)

Hình 2. (tiếp)

3. Phương pháp nghiên cứu

3.1. Hồi quy tuyến tính

Hồi quy tuyến tính (Linear Regression) là kỹ thuật thống kê nền tảng nhằm thiết lập mối quan hệ tuyến tính giữa một biến phụ thuộc (y - thời gian di chuyển) và nhiều biến độc lập (x_i). Phương trình tổng quát của mô hình có dạng [10]:

$$y=\beta_0 +\beta_1 x_1 +\beta_2 x_2 2+...+\beta_n x_n+ \varepsilon \quad (1)$$

Trong đó, β_0 là các hệ số hồi quy và ε là sai số ngẫu nhiên. Các tham số được ước lượng bằng phương pháp bình phương tối thiểu, nhằm tối thiểu hóa tổng bình phương sai số giữa giá trị thực tế và giá trị dự đoán.

Để đảm bảo độ tin cậy, cấu trúc dữ liệu đầu vào bắt buộc phải thỏa mãn các giả định khắt khe: tồn tại mối quan hệ tuyến tính, phần dư phân phối chuẩn, phương sai đồng nhất và không xảy ra hiện tượng đa cộng tuyến. Sự phù hợp của mô hình được chẩn đoán qua hệ số xác định R^2. Theo Roustaei (2024), nếu R^2 tiến gần về 0, điều này minh chứng dữ liệu đã vi phạm các giả định nền tảng và chứa đựng các cấu trúc phi tuyến phức tạp [11]. Sự ràng buộc khắt khe này chính là nguyên nhân khiến hồi quy tuyến tính truyền thống thường gặp hạn chế khi xử lý các bộ dữ liệu logistics đa chiều.

3.2. Hồi quy đỉnh

Hồi quy đỉnh (Ridge Regression) là một kỹ thuật mở rộng của hồi quy tuyến tính, được thiết kế để khắc phục những hạn chế của phương pháp ước lượng bình phương tối thiểu thông thường (OLS) khi xử lý các bộ dữ liệu đa chiều. Theo Saleh và cộng sự (2019), hạt nhân của phương pháp này là cơ chế quy chuẩn L_2, giúp giảm thiểu tác động của hiện tượng đa cộng tuyến và ngăn chặn tình trạng quá khớp. Hồi quy đỉnh kiểm soát sự bất ổn định của mô hình bằng cách tích hợp một siêu tham số phạt λ có giá trị dương vào hàm mục tiêu [12]. Phương trình hàm mất mát cần tối thiểu hóa có dạng:

$$L_{Ridge}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \quad (2)$$

Trong đó, sự thay đổi của siêu tham số λ mang tính quyết định đến cấu trúc mô hình: nếu $\lambda\to 0$, các hệ số hồi quy sẽ gia tăng và tương đương với phương pháp OLS; ngược lại, khi $\lambda\to\infty$, giá trị của các hệ số β sẽ bị ép tiến dần về 0, mang lại một mô hình có phương sai thấp và độ ổn định cao hơn. Mặc dù khắc phục được nhiễu do đa cộng tuyến, nhưng do bản chất vẫn là một mô hình tuyến tính, hồi quy đỉnh thường bộc lộ giới hạn cốt lõi khi phải lập bản đồ các mối quan hệ phi tuyến phức tạp giữa không gian, chu kỳ thời gian và đặc trưng khí tượng trong mạng lưới giao thông.

3.3. Cây quyết định

Cây quyết định (Decision Tree) là một phương pháp học máy phi tham số được sử dụng để xây dựng các mô hình dự báo thông qua kỹ thuật phân vùng đệ quy không gian dữ liệu. Đối với biến mục tiêu mang giá trị liên tục như thời gian di chuyển trong logistics, mô hình cây hồi quy được áp dụng bằng cách phân chia tập dữ liệu thành các phân vùng nhỏ hơn và thiết lập một mô hình dự báo đơn giản tại mỗi phân vùng, với sai số thường được đo lường bằng bình phương độ lệch giữa giá trị quan sát và giá trị dự báo [13].

Ưu điểm vượt trội của phương pháp này là tính linh hoạt cao, ít đòi hỏi các giả định khắt khe về cấu trúc mô hình thống kê, và đặc biệt bền vững ngay cả trong điều kiện dữ liệu bị khuyết thiếu. Hơn nữa, kết quả đầu ra được biểu diễn dưới cấu trúc đồ thị dạng cây trực quan, giúp các nhà quản trị dễ dàng diễn giải quy luật dữ liệu. Mặc dù khắc phục được điểm yếu của các mô hình tuyến tính, giới hạn lớn nhất của cây quyết định là tính thiếu ổn định; mô hình rất nhạy cảm với các nhiễu động nhỏ trong tập mẫu và cực kỳ dễ rơi vào trạng thái quá khớp [13]. Đặc tính rủi ro này chính là tiền đề buộc nghiên cứu phải tiến tới các thuật toán học tổ hợp nhằm kiểm soát phương sai trong các bước tiếp theo.

3.4. Rừng ngẫu nhiên

Để khắc phục rủi ro mất ổn định và hiện tượng quá khớp của một mô hình cây quyết định đơn lẻ, rừng ngẫu nhiên (Random Forest) áp dụng như một bước tiến đột phá thuộc nhóm phương pháp học tổ hợp. Thuật toán này hoạt động dựa

trên việc thiết lập một quần thể gồm số lượng lớn các cây quyết định hồi quy độc lập, được xây dựng thông qua hai cơ chế ngẫu nhiên hóa cốt lõi. Thứ nhất, mô hình áp dụng kỹ thuật tổng hợp tái lặp, trong đó mỗi cây được huấn luyện trên một tập dữ liệu con được trích xuất ngẫu nhiên có hoàn lại từ tập dữ liệu gốc. Thứ hai, tại mỗi nút quyết định, thuật toán không đánh giá toàn bộ các đặc trưng mà chỉ lựa chọn ngẫu nhiên một tập hợp con các biến dự báo để tìm ra ngưỡng phân chia tối ưu [14].

Đối với bài toán hồi quy nhằm dự báo thời gian di chuyển, kết quả đầu ra cuối cùng của mô hình được xác định bằng cách lấy trung bình cộng các giá trị dự báo từ toàn bộ các cây trong rừng. Sức mạnh vượt trội của rừng ngẫu nhiên nằm ở khả năng tự động nhận diện các tương tác phức tạp và hiệu ứng phi tuyến tính giữa các biến số đầu vào đa chiều mà không đòi hỏi các giả định tiên nghiệm khắt khe. Tuy nhiên, sự đánh đổi cho độ chính xác cao này là bản chất của thuật toán; cấu trúc phức tạp từ hàng trăm cây quyết định khiến mô hình thiếu đi sự diễn giải trực quan, đòi hỏi phải sử dụng thêm các kỹ thuật đo lường độ quan trọng của biến [14]. Đồng thời, chi phí tính toán cao của mô hình này cũng chính là tiền đề để nghiên cứu tiếp tục đánh giá các kỹ thuật tăng cường trọng số hiệu quả hơn ở các bước tiếp theo.

3.5. Tăng cường độ dốc

Khác với mô hình rừng ngẫu nhiên, tăng cường độ dốc (Gradient Boosting) là một kỹ thuật học tổ hợp tiếp cận theo chiến lược tuần tự. Theo Natekin và Knoll (2013), nguyên lý cốt lõi của phương pháp này là liên tục bổ sung các mô hình học cơ sở yếu để tối ưu hóa trực tiếp hàm mất mát trong không gian hàm thông qua thuật toán giảm dần đạo hàm [15]. Đối với bài toán hồi quy, quá trình này tương đương với việc liên tục khớp lại các phần dư giả từ các bước lặp trước đó. Mặc dù sở hữu năng lực mô hình hóa phi tuyến xuất sắc, tăng cường Gradient lại bộc lộ hạn chế lớn về chi phí thời gian huấn luyện do bản chất học tuần tự không cho phép tính toán song song. Rào cản cơ học này chính là tiền đề để nghiên cứu áp dụng một phiên bản cải tiến đột phá hơn là XGBoost ở phần tiếp theo.

3.6. Tăng cường độ dốc cực hạn

Kế thừa và khắc phục triệt để những hạn chế của phương pháp học tuần tự trước đó, tăng cường độ dốc cực hạn (XGBoost) là một kỹ thuật học tổ hợp có khả năng mở rộng cao, được thiết kế chuyên biệt nhằm tối ưu hóa đồng thời cả tốc độ huấn luyện lẫn hiệu suất tổng quát hóa. Để giải quyết rào cản về thời gian tính toán của Gradient Boosting truyền thống, XGBoost tích hợp cơ chế xử lý song song ở cấp độ khối dữ liệu trong quá trình tìm kiếm ngưỡng phân chia tối ưu của các cây quyết định [16]. Về mặt toán học, điểm đột phá cốt lõi của thuật toán này nằm ở việc đưa trực tiếp các thành phần điều chuẩn vào phương trình hàm mục tiêu. Cơ chế này giúp hệ thống kiểm soát chặt chẽ độ phức tạp của cấu trúc mô hình, ngăn chặn hiệu quả hiện tượng quá khớp khi đối mặt với các bộ dữ liệu giao thông mang tính ngẫu nhiên và chứa nhiều nhiễu động. Nhờ sự kết hợp toàn diện giữa sức mạnh mô hình hóa phi tuyến, tốc độ xử lý vượt trội và tham số cấu hình linh hoạt, XGBoost hội tụ đầy đủ các đặc tính cơ học ưu việt nhất để được lựa chọn làm thuật toán hạt nhân phục vụ bài toán dự báo thời gian di chuyển của hệ thống NovaLog AI.

Trong nghiên cứu này, XGBoost hoạt động dưới dạng một mô hình hồi quy dự báo. Kết quả dự báo thời gian di chuyển có độ chính xác cao từ XGBoost sẽ đóng vai trò cung cấp các tham số đầu vào vững chắc, hỗ trợ đắc lực cho khâu ra quyết định và điều phối logistics của nhà quản lý ở các bước tiếp theo.

3.7. Hồi quy véc-tơ hỗ trợ

Hồi quy véc-tơ hỗ trợ (SVR) là một thuật toán học máy tiên tiến được phát triển từ phương pháp Máy véc-tơ hỗ trợ (SVM) để giải quyết các bài toán dự báo giá trị liên tục [17]. Điểm độc đáo của SVR so với các mô hình thông thường là cách nó xử lý sai số. Thay vì cố gắng triệt tiêu mọi sai số nhỏ nhất, SVR tạo ra một ống dung sai bao quanh đường dự báo. Thuật toán sẽ bỏ qua và hoàn toàn không phạt các điểm dữ liệu rơi vào bên trong

không gian ống này. Chỉ những điểm dữ liệu dị thường, nằm ngoài ranh giới ống mới bị tính là sai số và chịu hình phạt. Mục tiêu cốt lõi của SVR là tìm ra một mô hình cân bằng giữa việc giảm thiểu các sai số lớn và giữ cho đường dự báo phẳng, ổn định nhất có thể.

Bên cạnh đó, để xử lý các mối quan hệ phi tuyến tính phức tạp trong dữ liệu giao thông, SVR sử dụng kỹ thuật hàm nhân nhằm ánh xạ dữ liệu lên một không gian nhiều chiều hơn để dễ dàng nhận diện quy luật. Nhờ cơ chế dung sai này, SVR hoạt động cực kỳ bền vững và ít bị đánh lừa bởi các dữ liệu nhiễu ngẫu nhiên trên đường truyền.

3.8. K – láng giềng gần nhất

K-láng giềng gần nhất (KNN) là một trong những thuật toán học máy có giám sát cơ bản và đơn giản nhất. Khác với các mô hình tham số, KNN không thiết lập một phương trình toán học tổng quát ngay từ ban đầu mà dựa hoàn toàn vào tập dữ liệu đã được huấn luyện để tiến hành nhận diện và dự báo cho các điểm dữ liệu mới [18]. Về mặt cơ học, thuật toán KNN vận hành tuần tự qua 4 bước cốt lõi. Bước 1: Khởi tạo tham số K, đại diện cho số lượng các láng giềng (điểm dữ liệu) gần nhất sẽ được đưa vào xem xét. Bước 2: Đo lường khoảng cách từ điểm dữ liệu mới (chuyến hàng cần dự báo) đến toàn bộ các điểm dữ liệu đang tồn tại trong tập huấn luyện. Phương pháp tính toán tiêu chuẩn và phổ biến nhất được sử dụng là công thức khoảng cách Euclid. Bước 3: Sàng lọc và trích xuất ra đúng K láng giềng có khoảng cách Euclid nhỏ nhất so với điểm dữ liệu mới, đồng thời loại bỏ các láng giềng ở khoảng cách xa. Bước 4: Đưa ra quyết định dự báo dựa trên K láng giềng đã chọn, sử dụng cơ chế bầu chọn đa số đối với bài toán phân lớp, hoặc tính giá trị trung bình đối với bài toán hồi quy thời gian di chuyển.

Mặc dù có ưu điểm về sự trực quan và dễ triển khai, nhưng việc dựa hoàn toàn vào đo lường khoảng cách khiến KNN dễ bộc lộ giới hạn hiệu suất khi phải xử lý các bộ dữ liệu logistics có quy mô quá lớn hoặc chứa nhiều điểm nhiễu động dị thường.

3.9. Xác thực chéo

Để đánh giá một cách khách quan năng lực tổng quát hóa của các thuật toán dự báo và kiểm soát triệt để hiện tượng quá khớp, nghiên cứu áp dụng phương pháp xác thực chéo (CV). Đây là một trong những kỹ thuật lấy mẫu lại dữ liệu nền tảng và chuẩn mực nhất trong quy trình lựa chọn và đánh giá mô hình học máy [19]. Cụ thể, nghiên cứu sử dụng chiến lược xác thực chéo k bước. Tập dữ liệu thực chứng ban đầu được phân chia ngẫu nhiên thành k tập con rời rạc có dung lượng xấp xỉ bằng nhau. Quá trình huấn luyện và đánh giá mô hình được tiến hành lặp lại đệ quy k lần. Tại mỗi vòng lặp, một tập con duy nhất được trích xuất để làm tập xác thực nhằm kiểm định mô hình, trong khi k−1 tập con còn lại được tích hợp để tạo thành tập dữ liệu huấn luyện. Cấu trúc quay vòng này đảm bảo rằng tuyệt đối mọi điểm dữ liệu trong mẫu đều được sử dụng làm dữ liệu kiểm tra đúng một lần. Kết quả hiệu suất cuối cùng của mô hình dự báo được xác định bằng cách lấy trung bình cộng các cấu phần sai số thu được từ toàn bộ k tập xác thực.

Tỷ lệ phân chia tập dữ liệu 70% huấn luyện và 30% kiểm tra được thiết lập ban đầu là bước khởi tạo cơ sở. Đối với một tập dữ liệu thực chứng mang đậm tính đặc thù về sự đan xen phức tạp giữa các hệ tọa độ không gian và chu kỳ thời gian như mạng lưới logistics đô thị, việc chỉ dựa vào một lần chia tách ngẫu nhiên đơn lẻ sẽ tiềm ẩn rủi ro rất lớn về độ chệch và hoàn toàn không đủ cơ sở thống kê để khẳng định tính bền vững của mô hình học máy. Để khắc phục triệt để giới hạn này và đáp ứng tiêu chuẩn kiểm định khắt khe nhất, nghiên cứu đã thiết lập một khung đánh giá mở rộng bằng cách tích hợp phương pháp xác thực chéo kết hợp cùng kỹ thuật mô phỏng Monte Carlo thông qua 30 lần chạy. Quá trình xáo trộn và lấy mẫu ngẫu nhiên lại liên tục 30 lần giúp hệ thống triệt tiêu hoàn toàn các sai lệch có thể phát sinh do yếu tố may rủi trong lần cắt mẫu ban đầu. Cách tiếp cận này đảm bảo mọi kết quả đối sánh hiệu suất giữa các thuật toán đều mang tính bền vững về mặt thống kê, qua đó cung cấp một thước đo trung thực và khách quan nhất về năng lực tổng quát hóa của hệ thống

NovaLog AI khi đối mặt với những nhiễu động ngẫu nhiên trong thực tiễn.

3.10. Chỉ tiêu đánh giá hiệu suất mô hình

Để định lượng và so sánh khách quan năng lực tổng quát hóa của các thuật toán học máy, nghiên cứu áp dụng ba chỉ tiêu thống kê chuẩn mực, bao gồm: sai số tuyệt đối trung bình (MAE), sai số căn bình phương trung bình (RMSE) và hệ số xác định (R^2). Cụ thể:

Sai số tuyệt đối trung bình đo lường mức độ sai lệch tuyệt đối trung bình giữa giá trị thực tế và giá trị dự báo của mô hình. Phép đo này cung cấp một cái nhìn trực quan về biên độ sai số trung bình mà không bị chi phối quá mạnh bởi các điểm dữ liệu dị thường.

$$MAE = \frac{1}{n}\Sigma|y_i - \hat{y}_i| \qquad (3)$$

Sai số căn bình phương trung bình phản ánh mức độ sai lệch tổng thể của chuỗi dự báo. Do cơ chế bình phương phần dư trước khi lấy trung bình, RMSE đặc biệt nhạy cảm và áp dụng hình phạt nặng đối với những sai số dự báo có độ lớn bất thường. Điều này khiến RMSE trở thành một chỉ tiêu khắt khe và phù hợp để đánh giá độ tin cậy của hệ thống logistics.

$$RMSE = \sqrt{\frac{1}{n}\Sigma(y_i - y_i)^2} \qquad (4)$$

Hệ số xác định (R^2) là thước đo đánh giá mức độ giải thích phương sai của biến phụ thuộc bởi các đặc trưng đầu vào trong mô hình. Giá trị R^2 càng tiệm cận 1 minh chứng mô hình có độ vặn khớp với dữ liệu thực chứng càng cao.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad (5)$$

Sai số phần trăm tuyệt đối trung bình **(**MAPE) là chỉ số đo sai số dự báo thường dùng trong thống kê, phân tích dữ liệu và dự báo.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100\% \qquad (6)$$

Trong đó: *n* là số lượng mẫu quan sát, y_i là giá trị thời gian di chuyển thực tế, $\hat{y}_i$ là giá trị dự báo từ thuật toán, $\bar{y}$ là giá trị trung bình của tập dữ liệu thực tế, A_t là giá trị thực tế, *Ft* là giá trị dự báo.

Tỷ lệ dự báo chính xác trong phạm vi ±20% (A20) là chỉ số đánh giá độ chính xác dự báo trong phạm vi ±20% so với giá trị thực tế.

$$A20 = \frac{\text{Số dự báo có sai số} \le 20\%}{\text{Tổng số dự báo}} \times 100\% \qquad (7)$$

4. Kết quả và thảo luận

4.1. Khảo sát và tối ưu siêu tham số XGBoost

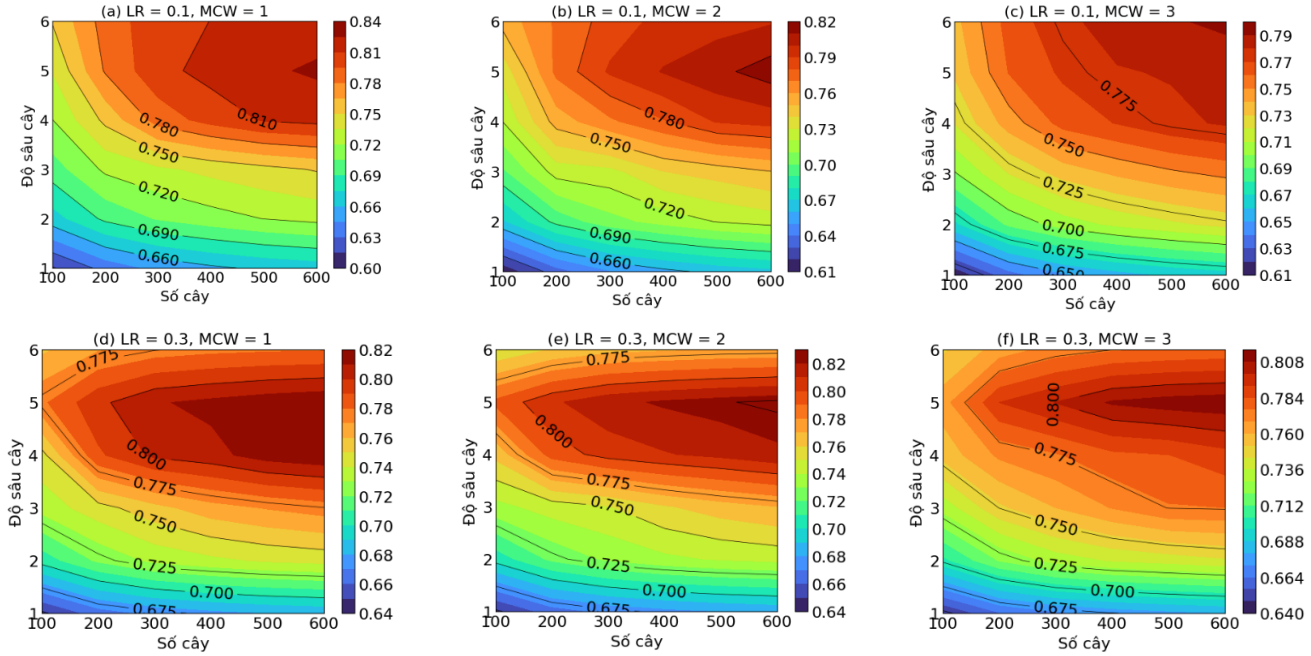
Trong nghiên cứu này, phương pháp tìm kiếm lưới (Grid Search) kết hợp với xác thực chéo được lựa chọn để tối ưu hóa siêu tham số cho mô hình XGBoost. Mặc dù các thuật toán tối ưu hóa bầy đàn có khả năng tìm kiếm toàn cục mạnh mẽ trong không gian tham số liên tục, phương pháp Grid Search vẫn được ưu tiên sử dụng do tính tường minh, khả năng bao phủ toàn diện các kết hợp tham số trong miền khảo sát xác định. Việc sử dụng Grid Search giúp đảm bảo rằng mọi kịch bản cấu hình quan trọng đều được thực chứng, đồng thời tránh rủi ro hội tụ sớm vào các cực trị cục bộ của các thuật toán meta-heuristic khi kích thước không gian tham số ở mức vừa phải. Tuy nhiên, việc tích hợp các thuật toán tối ưu hóa thông minh sẽ là một định hướng mở rộng quan trọng trong các nghiên cứu tiếp theo để nâng cao hơn nữa hiệu suất của hệ thống NovaLog AI.

Để phát huy tối đa năng lực mô hình hóa phi tuyến của thuật toán XGBoost và kiểm soát chặt chẽ hiện tượng quá khớp đối với bộ dữ liệu thời gian di chuyển, quá trình khảo sát và tinh chỉnh các siêu tham số đóng vai trò mang tính quyết định. Không gian tìm kiếm được thiết lập tập trung vào bốn siêu tham số cốt lõi chi phối trực tiếp đến độ phức tạp của cấu trúc cây và tốc độ hội tụ của thuật toán, bao gồm: độ sâu tối đa của cây (khảo sát từ 1 đến 6), số lượng cây (từ 100 đến 600), tốc độ học (ở các mức 0.1, 0.3, 0.5) và trọng số con tối thiểu (ở các mức 1, 2, 3) như được trình bày chi tiết tại Bảng 2.

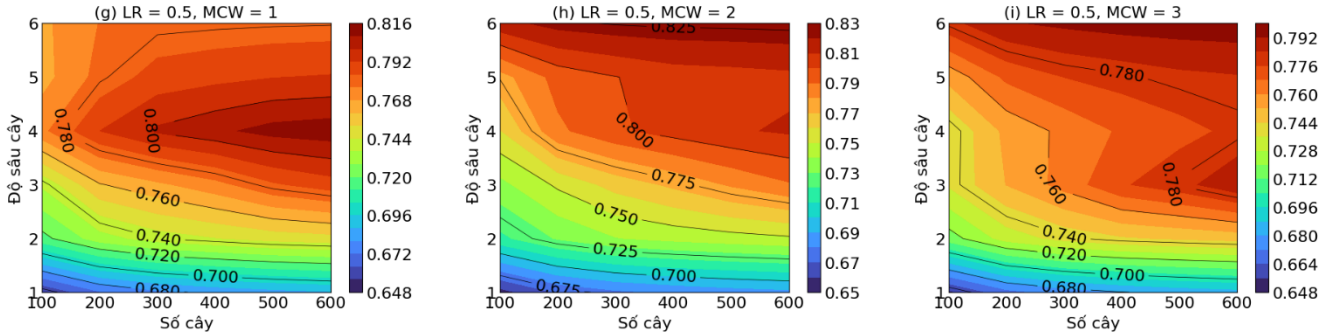
Bảng 2. Miền tìm kiếm của các siêu tham số của mô hình XGBoost

Tên siêu tham số	API	Miền giá trị khảo sát
Độ sâu cây	max_depth	1, 2, 3, 4, 5, 6
Số cây	n_estimators	100, 200, 300, 400, 500, 600
Tốc độ học (LR)	learning_rate	0.1, 0.3, 0.5
Trọng số con tối thiểu (MCW)	min_child_weight	1, 2, 3

Sự biến thiên và mức độ ảnh hưởng của các siêu tham số này đến hiệu năng tổng thể của mô hình đã được trực quan hóa chi tiết trong Hình 3. Dưới góc độ học máy, biểu đồ này cung cấp một cái nhìn sâu sắc về không gian tối ưu: khi tăng số lượng cây dự báo và độ sâu của từng cây, năng lực học các quy luật phi tuyến của mô hình tăng lên rõ rệt, đẩy giá trị lên cao. Tuy nhiên, sự gia tăng độ phức tạp này đòi hỏi hệ thống phải được kiểm soát chặt chẽ bằng tốc độ học và ngưỡng trọng số phân chia phù hợp, nếu không mô hình sẽ rơi vào trạng thái học thuộc lòng nhiễu động của tập dữ liệu huấn luyện. Thông qua kỹ thuật xác thực chéo trên toàn bộ lưới không gian tham số, nghiên cứu đã sàng lọc và trích xuất thành công 5 mô hình tốt nhất trên tập xác thực, được trình bày chi tiết tại Bảng 3. Kết quả phân tích Bảng 3 cho thấy một quy luật hội tụ học máy rất rõ ràng: toàn bộ 5 mô hình tốt nhất (được ký hiệu từ M1 đến M5) đều yêu cầu xây dựng một quần thể rừng với số lượng cây đạt mức tối đa trong không gian khảo sát bằng 600. Điều này minh chứng một thực tế khoa học rằng bộ dữ liệu logistics đang xét chứa đựng các mối quan hệ tương tác đan xen rất phức tạp giữa không gian, chu kỳ thời gian và các điều kiện vi khí hậu, đòi hỏi một hệ thống học tổ hợp quy mô lớn và đủ sâu mới có thể bóc tách toàn diện các quy luật tiềm ẩn. Bên cạnh sự đồng nhất về số lượng cây, sự phân hóa hiệu suất giữa các cấu hình top đầu chủ yếu nằm ở cách kết hợp giữa độ sâu cây và tốc độ học. Mô hình M5 đạt được điểm xác thực chéo cao nhất (R^2 = 0.8275) nhờ sử dụng cấu trúc cây sâu nhất bằng 6 kết hợp với LR = 0.5 và MCW = 2. Tiếp theo là mô hình M1 (R^2 = 0.8271) và M4 (R^2 = 0.8267). Cả hai mô hình này thiết lập cấu trúc cây nông hơn một chút là 5 để giảm rủi ro, nhưng M1 chọn phương án học rất chậm là 0.1, trong khi M4 chọn tốc độ học trung bình là 0.3. Còn lại, các mô hình M3 (R^2 = 0.8182) và M2 (R^2 = 0.8150) cũng duy trì độ sâu cây bằng 5 và số lượng cây bằng 600, tạo ra các phương án dự phòng với sự thay đổi về tổ hợp trọng số con.



Hình 3. Ảnh hưởng của các siêu tham số đến hiệu suất mô hình



Hình 3. (tiếp)

Mặc dù cả 5 mô hình đều cho thấy năng lực tổng quát hóa tốt khi vượt qua ngưỡng R^2 trên 0.81 trên tập xác thực, nhưng dưới góc độ thiết kế hệ thống dự báo giao thông, những cấu hình có độ sâu lớn và tốc độ học quá cao như M5 hoặc học quá chậm như M1 luôn tiềm ẩn rủi ro mất cân bằng phương sai - độ chệch khi đưa vào vận hành trên dữ liệu chưa từng tiếp xúc. Do đó, việc chỉ dựa vào trong Bảng 3 là chưa đủ cơ sở khoa học để chốt cấu hình vận hành cuối cùng cho hệ thống NovaLog AI. Trong nghiên cứu học máy ứng dụng vào giao thông vận tải, kết quả xác thực chéo mang tính chất tham khảo ban đầu về không gian siêu tham số. Để đánh giá chính xác năng lực tổng quát hóa của thuật toán và loại bỏ hoàn toàn các rủi ro thống kê, 5 mô hình xuất sắc nhất (từ M1 đến M5) bắt buộc phải được đưa vào kiểm định độc lập trên cả hai tập dữ liệu: tập huấn luyện và tập kiểm tra.

Phân tích số liệu thực chứng từ Bảng 4 cho thấy sự chênh lệch về hiệu suất khá lớn của các mô hình khi làm việc với dữ liệu chưa từng tiếp xúc. Mô hình M1 đạt hiệu suất dự báo gần như tuyệt đối trên tập huấn luyện với R^2 = 0.9989, RMSE = 0.3475. Tuy nhiên, khi dự báo trên tập kiểm tra, hiệu suất của mô hình này giảm đáng kể với R^2 = 0.9143 và RMSE = 3.2302. Sự chênh lệch đáng kể giữa hiệu suất huấn luyện và kiểm tra là minh chứng điển hình của hiện tượng quá khớp. Tương tự, mô hình M5 có điểm xác thực chéo cao nhất nhưng lại cho thấy sự thiếu ổn định về hiệu suất dự báo với sai số RMSE = 3.3022 (cao nhất trong 5 mô hình).

Bảng 3. Top 5 mô hình XGBoost tốt nhất dựa trên R^2 của tập xác thực

STT	Độ sâu cây	Số cây	Tốc độ học (LR)	Trọng số con tối thiểu (MCW)	R^2 (CV)
M1	5	600	0.1	1	0.8271
M2	5	600	0.1	2	0.8150
M3	5	600	0.3	1	0.8182
M4	5	600	0.3	2	0.8267
M5	6	600	0.5	2	0.8275

Bảng 4. Kết quả dự báo của 5 mô hình XGBoost trên tập huấn luyện và tập kiểm tra

Mô hình	R^2 huấn luyện	RMSE huấn luyện	R^2 kiểm tra	RMSE kiểm tra
M1	0.9989	0.3475	0.9143	3.2302
M2	0.9620	2.0498	0.8922	3.6234
M3	0.9879	1.1192	0.9375	2.8965
M4	0.9887	1.1154	0.9119	3.2745
M5	0.9598	2.1086	0.9104	3.3022

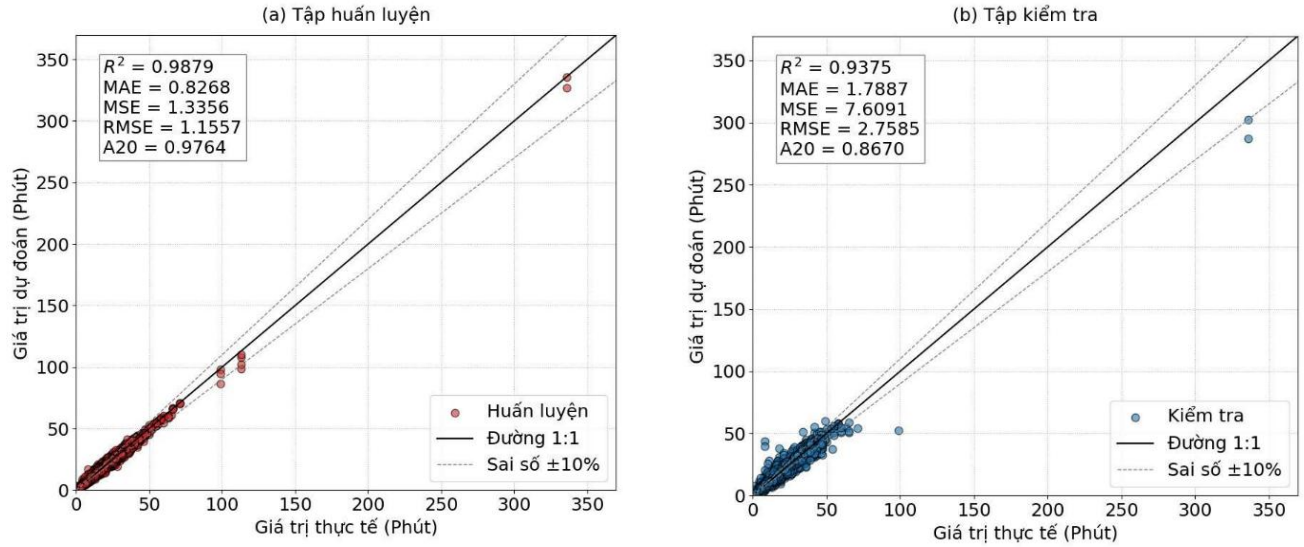
Trái ngược với mô hình M1 và M5, mô hình M3 có hiệu suất dự báo cao và phổ quát hơn với R^2 = 0.9879, RMSE = 1.1192 trên tập huấn luyện và R^2 = 0.9375, RMSE = 2.8965 trên tập kiểm tra. Hai mô hình còn lại là M2 và M4 có hiệu suất dự báo thấp hơn mô hình M3 trên tập dữ liệu kiểm tra

và bị hiện tượng quá khớp tương tự như mô hình M1 và M5. Từ những phân tích bên trên, mô hình M3 được lựa chọn là mô hình đại diện và tiếp tục được đánh giá ở phần tiếp theo.

4.2. Kết quả dự đoán đại diện

Phần này trình bày hiệu suất dự báo của mô hình XGBoost (M3) tối ưu được đánh giá thông qua biểu đồ hồi quy và biểu đồ sai số.. Hình 4 thể hiện mối quan hệ giữa giá trị dự báo và giá trị thực tế của mô hình XGBoost. Các điểm dữ liệu tập trung chặt quanh đường chéo 45°, đặc biệt trong vùng giá trị phổ biến của thời gian di chuyển, cho thấy mức độ hội tụ giữa dự báo và thực tế. Độ phân tán của điểm dữ liệu tăng nhẹ ở các giá trị lớn, phản ánh đặc trưng khó dự báo hơn của các hành trình dài, tuy nhiên không xuất hiện xu hướng lệch có hệ thống.

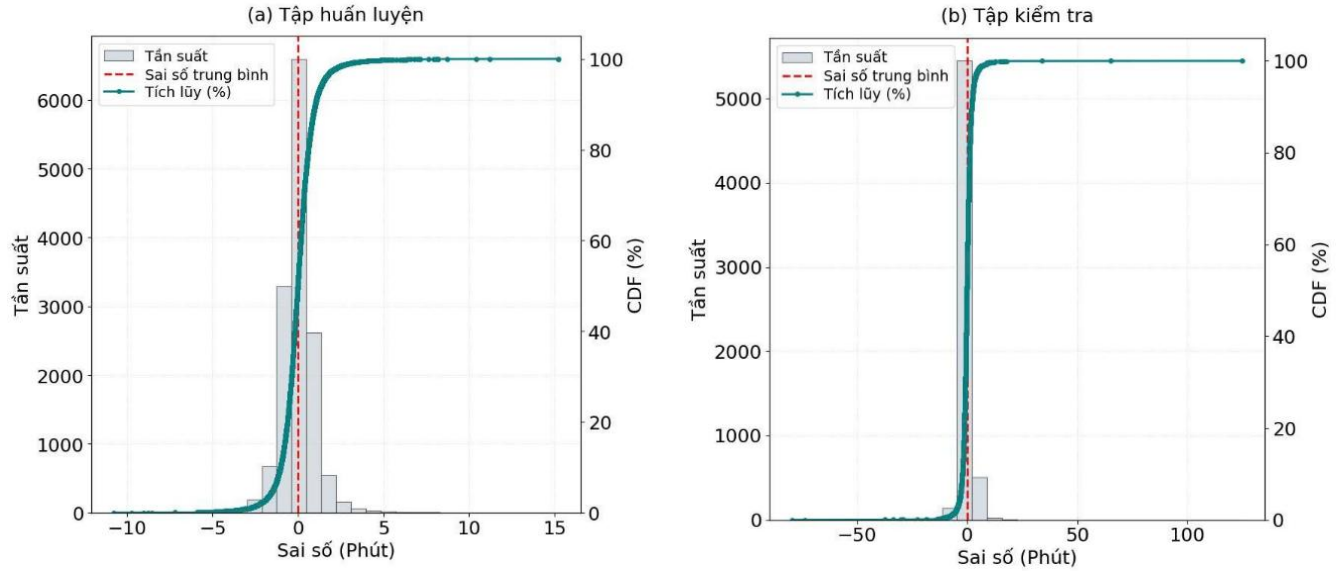
Điều này cho thấy mô hình đã học hiệu quả các quan hệ phi tuyến nhưng vẫn đảm bảo tính phổ quát.



Hình 4. Kết quả hồi quy của mô hình trên giá trị dự báo và giá trị thực tế

Bảng 5. Các chỉ số đánh giá thống kê của mô hình đại diện

Tiêu chí đánh giá	R^2	MAE	MSE	RMSE	A20
Tập huấn luyện	0.9879	0.8268	1.3356	1.1557	0.9764
Tập kiểm tra	0.9375	1.7887	7.6091	2.7585	0.8670



Hình 5. Kết quả sai số dự báo

Các kết quả định lượng của mô hình đại diện được trình bày chi tiết trong Bảng 5. Trên tập huấn luyện, mô hình đạt R^2 = 0.9879, MAE = 0.8268 và RMSE = 1.1557, cho thấy khả năng mô tả dữ liệu rất cao với sai số nhỏ. Trên tập kiểm tra, mô hình vẫn duy trì hiệu suất tốt với R^2 = 0.9375 MAE = 1.7887 và RMSE = 2.7585. Hơn nữa, chỉ số A20 đạt 0.8670 trên tập dữ liệu kiểm tra cho thấy phần lớn dự báo có sai lệch dưới 20%, đáp ứng yêu cầu ứng dụng trong thực tiễn logistics. Cuối cùng, Hình 5 cung cấp phân tích sâu hơn về phân phối sai số dự báo. Kết quả cho thấy sai số chủ yếu tập trung trong khoảng [-3; 3] (phút) đối với tập huấn luyện và mở rộng trong khoảng [-7; 7] (phút) đối với tập kiểm tra. Đáng chú ý, trên 95% dữ liệu của cả hai tập đều có sai số nằm trong khoảng [-3; 3] (phút), cho thấy phần lớn dự báo có độ chính xác cao. Phân phối sai số có dạng xấp xỉ chuẩn và đối xứng quanh 0, khẳng định mô hình không bị chệch và có độ ổn định tốt khi triển khai trên dữ liệu thực tế.

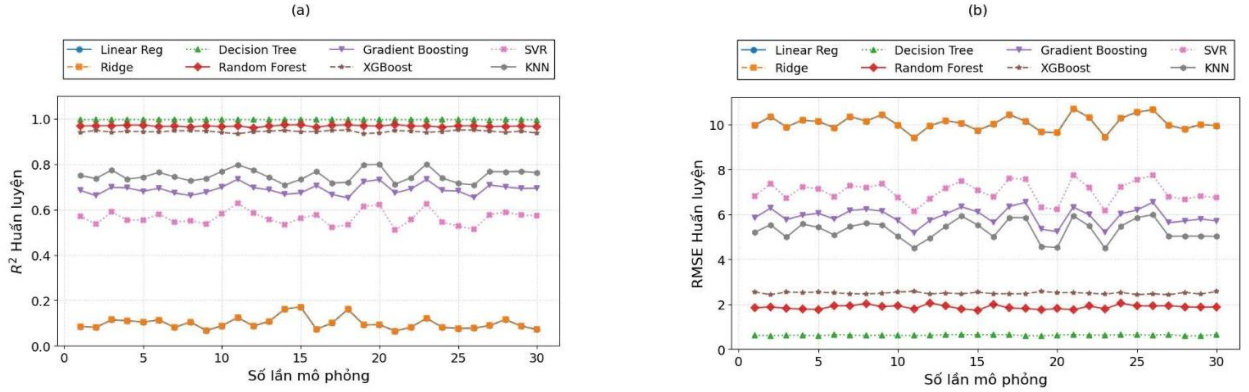
4.3. So sánh các mô hình

Để đánh giá tính ổn định và độ tin cậy của mô hình XGBoost, nghiên cứu tiến hành mô phỏng Monte Carlo với 30 lần lặp lại để so sánh mô hình này với bảy mô hình khác.

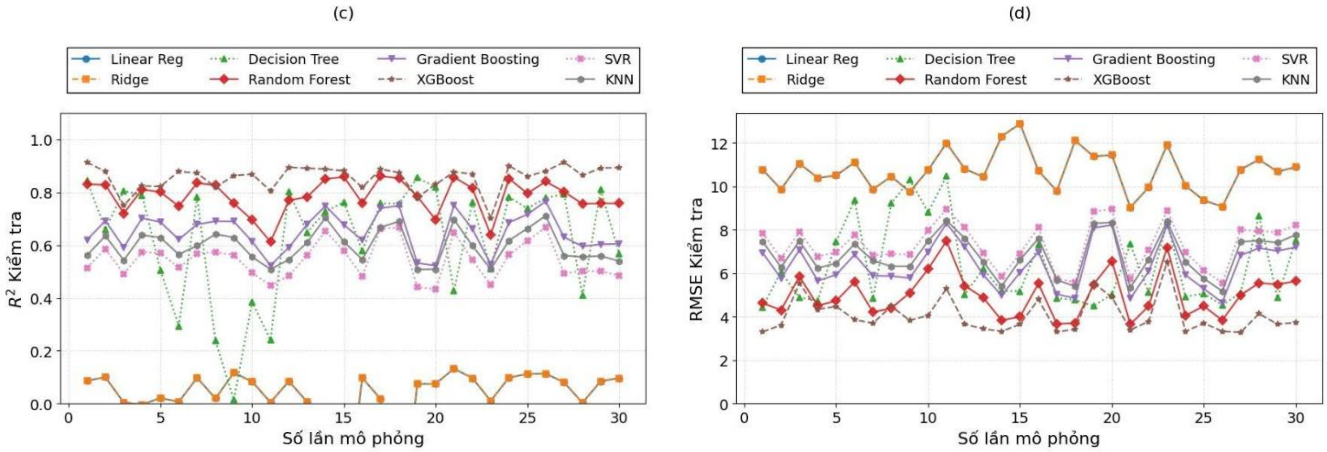
Kết quả tổng hợp Bảng 6 cho thấy sự khác biệt rõ rệt giữa các thuật toán, đặc biệt khi xét đồng thời trên tập huấn luyện và tập kiểm tra.

Bảng 6. So sánh XGB với các thuật toán khác sau khi tối ưu siêu tham số

	Thuật toán	R^2	RMSE
Tập huấn luyện	Linear Reg	0.0991	10.0787
	XGBoost	0.9879	1.1557
	Decision Tree	0.9965	0.6254
	Ridge	0.0991	10.0787
	Random Forest	0.9687	1.8756
	Gradient Boosting	0.6892	5.9219
	KNN	0.7508	5.3019
	SVR	0.5641	7.0137
Tập kiểm tra	Linear Reg	0.0051	10.7055
	Ridge	0.0051	10.7056
	Decision Tree	0.6286	6.3353
	Random Forest	0.7857	4.9782
	Gradient Boosting	0.6507	6.3821
	XGBoost	0.9375	2.7585
	SVR	0.5446	7.2918
	KNN	0.5972	6.8576



Hình 6. Hiệu suất dự báo các mô hình học máy sau 30 mô phỏng Monte Carlo trên tập huấn luyện, tập kiểm tra



Hình 6. (tiếp)

Trên tập huấn luyện, Decision Tree đạt hiệu suất cao nhất với R^2 = 0.9965 và RMSE = 0.6254, tiếp theo là XGBoost với R^2 = 0.9879 và RMSE = 1.1557. Tuy nhiên, mức độ phù hợp gần như tuyệt đối của Decision Tree là dấu hiệu của hiện tượng quá khớp khi so sánh với kết quả trên tập kiểm tra. Các mô hình tuyến tính (Linear Regression, Ridge) cho hiệu suất rất thấp R^2 = 0.099, RMSE > 10), cho thấy không phù hợp với dữ liệu có tính phi tuyến. Các mô hình như Random Forest, Gradient Boosting và KNN đạt kết quả trung bình, nhưng vẫn kém hơn XGBoost về độ chính xác.

Trên tập kiểm tra, XGBoost thể hiện ưu thế vượt trội với R^2 = 0.9375 và RMSE = 2.7585, cao hơn đáng kể so với tất cả các mô hình còn lại. Trong khi đó, Decision Tree giảm mạnh xuống R^2 = 0.6286, xác nhận hiện tượng quá khớp. Random Forest là mô hình có hiệu suất đứng thứ hai (R^2 = 0.7857, RMSE = 4.9782) nhưng vẫn kém XGBoost. Các mô hình Gradient Boosting, KNN và SVR có hiệu suất trung bình, với R^2 chỉ dao động từ khoảng 0.54 đến 0.65. Đặc biệt, các mô hình tuyến tính gần như không có khả năng dự báo (R^2 =0).

Tổng kết lại, XGBoost là mô hình duy nhất đạt được sự cân bằng tốt giữa độ chính xác cao trên tập huấn luyện và khả năng tổng quát hóa trên tập kiểm tra. Điều này khẳng định hiệu quả của việc tối ưu siêu tham số đối với XGBoost, đồng thời cho thấy mô hình phù hợp nhất cho bài toán dự báo trong bối cảnh dữ liệu có tính phi tuyến và biến động cao như logistics đô thị.

4.4. Phân tích ảnh hưởng của các yếu tố đầu vào đến hiệu suất dự báo của mô hình

Phần này trình bày biểu đồ ảnh hưởng từng phần nhằm cung cấp sâu hơn về sự ảnh hưởng của từng biến đầu vào đơn lẻ đối với giá trị dự báo đầu ra.

Kết quả từ Bảng 7 cho thấy trong số 13 yếu tố đầu vào, có 4 yếu tố ảnh hưởng mạnh mẽ nhất đến thời gian di chuyển của một chuyến hàng: quãng đường, hướng gió, kinh độ đi, vĩ độ đi. Quãng đường là yếu tố có vai trò quyết định lớn nhất với mức độ tác động cao nhất với hiệu số là 25.4472. Khoảng cách di chuyển càng dài thì thời gian giao hàng càng tăng, điều này hoàn toàn phù hợp với thực tế vật lý của ngành vận tải. Yếu tố thời tiết "Hướng gió" xếp ở vị trí thứ hai về mức độ ảnh hưởng với hiệu số là 9.3813. Trong vận chuyển, sức cản của gió tác động đáng kể đến vận tốc di chuyển của các phương tiện chở hàng nhẹ hoặc xe máy giao hàng tận nơi. Kinh độ đi và Vĩ độ đi lần lượt xếp ở vị trí thứ 3 với hiệu số 4.2691 và thứ 4 với hiệu số 3.9557. Cặp tọa độ không gian này cho thấy vị trí xuất phát ảnh hưởng rất lớn đến sự chậm trễ. Việc xuất phát từ những khu vực trung tâm đông đúc, nhiều nút giao sẽ mất nhiều thời gian hơn so với việc đi từ khu vực ngoại thành.

Một điểm đáng lưu ý là biến Giờ đi có trọng số đóng góp thấp (vị trí thứ 11) trong mô hình hiện tại. Nếu chỉ đánh giá dựa trên thang đo tầm quan trọng đặc trưng một chiều, kết quả này dường như

mâu thuẫn với thực tiễn giao thông đô thị, nơi sự chênh lệch thời gian di chuyển giữa khung giờ cao điểm và giờ thấp điểm là khá lớn. Việc biến Giờ đi xếp hạng thấp về trọng số tương đối không phủ nhận vai trò của thời điểm xuất phát. Nghiên cứu khẳng định thời gian di chuyển là hệ quả của sự tương tác phi tuyến giữa tọa độ không gian và chu kỳ thời gian. Biến "Giờ đi" không hoạt động độc lập mà có sự tương tác phi tuyến cực kỳ mạnh mẽ với các biến không gian (Kinh độ đi, Vĩ độ đi). Cụ thể, trên bình diện tổng thể của toàn mạng lưới, tác động của một khung giờ có thể bị trung bình hóa; nhưng khi kết hợp đồng thời không gian và thời gian trên bề mặt PDP-2D, thời gian di chuyển lập tức gia tăng đột biến khi phương tiện đi qua những dải tọa độ cụ thể thuộc vùng trung tâm chật hẹp của thành phố vào đúng các khung giờ cao điểm. Sự cộng hưởng chặt chẽ giữa không - thời gian này khiến thuật toán phân tán trọng số của biến Giờ đi khi đánh giá đơn lẻ, khẳng định mô hình đã học được động học phức tạp của dòng giao thông thực tế.

Bảng 7. Bảng xếp hạng độ quan trọng của các biến

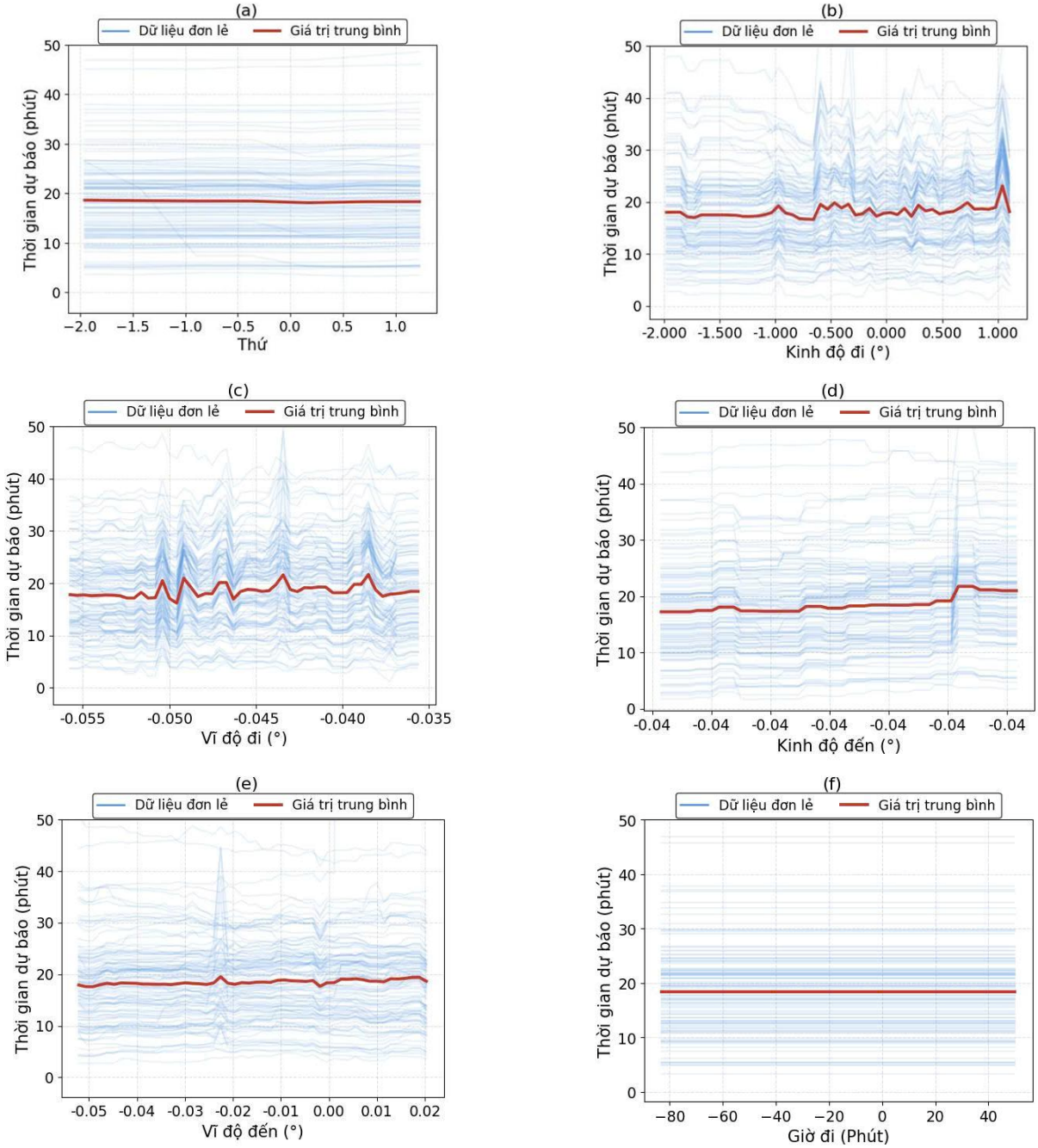
Xếp hạng	Đặc trưng	Nhỏ nhất	Lớn nhất	Hiệu số
1	Quãng đường (km)	6.5253	31.9725	25.4472
2	Hướng gió	9.4142	18.7956	9.3813
3	Kinh độ đi (°)	16.3083	20.5774	4.2691
4	Vĩ độ đi (°)	16.8171	20.7728	3.9557
5	Độ ẩm (%)	17.3611	19.5211	2.1599
6	Tốc độ gió (km/h)	17.3299	18.9164	1.5866
7	Vĩ độ đến (°)	17.7735	19.3186	1.5451
8	Nhiệt độ thấp (°C)	17.8221	18.9474	1.1254
9	Nhiệt độ cao (°C)	18.0405	19.0970	1.0565
10	Thứ	18.1006	18.5143	0.4137
11	Giờ đi (Phút)	18.4442	18.4442	0.0000
12	Kinh độ đến (°)	18.4589	21.6100	0.0000
13	Lượng mưa (mm)	18.4627	18.4537	0.0000

Bên cạnh động học phức tạp tại điểm xuất phát, cấu trúc không gian tại điểm đến cũng mang tính đặc thù cao và chi phối trực tiếp đến cơ chế học của mô hình. Cụ thể, Hình 1 ghi nhận hệ số tương quan tuyến tính nghịch hoàn hảo (-1.00) giữa Kinh độ đến và Vĩ độ đến. Dưới góc độ phân tích không gian, điều này phản ánh trực diện sự phân bố hình học đặc thù của các trạm trung chuyển trong tập mẫu khảo sát. Các điểm đến này được sắp xếp dọc theo một trục giao thông chính duy nhất (theo hướng Đông Bắc - Tây Nam), khiến vĩ độ và kinh độ biến thiên tỷ lệ nghịch với nhau. Sự phân bổ tuyến tính này tạo ra một ràng buộc hình học chặt chẽ, hỗ trợ thuật toán XGBoost dễ dàng nhận diện cấu trúc mạng lưới vận tải hiện tại. Tuy nhiên, đặc tính hình học này cũng đặt ra yêu cầu phải thận trọng khi ngoại suy hoặc tổng quát hóa mô hình sang các khu vực đô thị có quy hoạch mạng lưới dạng nan quạt hoặc bàn cờ.

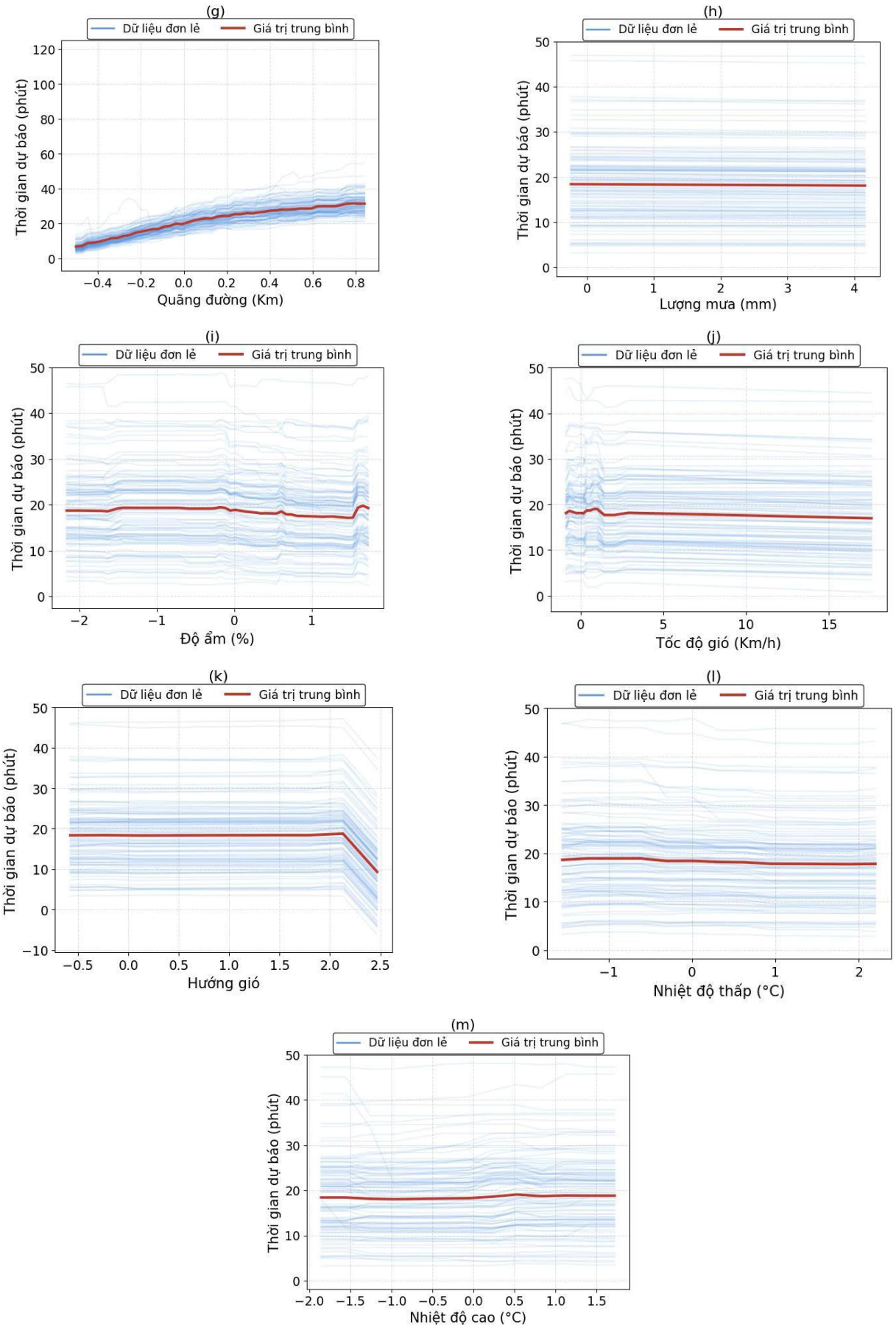
Mức độ tác động cụ thể của từng yếu tố này được thể hiện rõ qua Hình 7. Biểu đồ này giúp người quản lý nhìn thấy được những mức giới hạn mà tại đó thời gian giao hàng sẽ tăng vọt, chẳng hạn như khi quãng đường vượt qua một cự ly nhất định. Đối với biến Quãng đường (Hình 7g), đồ thị thể hiện một đường dốc lên, khẳng định quãng đường càng dài thì thời gian giao hàng càng tăng.

Tuy nhiên, đường dốc này không phải là một đường thẳng đều đặn. Ở những cự ly ngắn (ví dụ dưới 5 km), thời gian di chuyển có xu hướng tăng vọt rất nhanh. Điều này phản ánh thực tế giao thông đô thị: ở cự ly ngắn, xe chủ yếu đi qua các ngõ ngách, khu dân cư đông đúc hoặc phải chờ nhiều đèn đỏ. Khi cự ly đủ dài và phương tiện tiếp cận được các trục đường lớn hoặc đại lộ, tốc độ tăng của thời gian di chuyển bắt đầu ổn định và

thoải hơn. Đối với biến Hướng gió (Hình 7k), đồ thị của hướng gió không đi lên hay đi xuống theo một chiều, mà tạo thành các đỉnh và đáy dao động. Điều này cho thấy thời gian di chuyển bị kéo dài đáng kể ở một số hướng gió nhất định. Trong thực tế môi trường đô thị vào tháng 12, sự tăng cường của các đợt gió mùa Đông Bắc, kết hợp với hiệu ứng hút gió tại các khu vực có nhiều tòa nhà cao tầng, sẽ tạo ra lực cản vật lý lớn. Điều này làm giảm tốc độ di chuyển an toàn của các phương tiện chở hàng nhẹ hoặc xe máy, từ đó làm tăng thời gian hoàn thành chuyến đi. Đối với "Kinh độ đi" và "Vĩ độ đi": Đồ thị của hai yếu tố tọa độ này cho thấy thời gian dự báo sẽ bị đẩy lên mức cao nhất tại một số dải tọa độ cụ thể. Những dải giá trị nhô cao này chính là tọa độ của các khu vực lõi trung tâm thành phố, các khu vực tập trung đông văn phòng, hoặc các nút giao thông thường xuyên xảy ra kẹt xe. Nếu một chuyến hàng xuất phát từ các tọa độ này, mô hình sẽ tự động cộng thêm một khoảng thời gian trễ đáng kể do phương tiện phải nhích từng chút một để thoát ra khỏi vùng ùn tắc.



Hình 7. Biểu đồ phụ thuộc một phần mô tả ảnh hưởng của các yếu tố

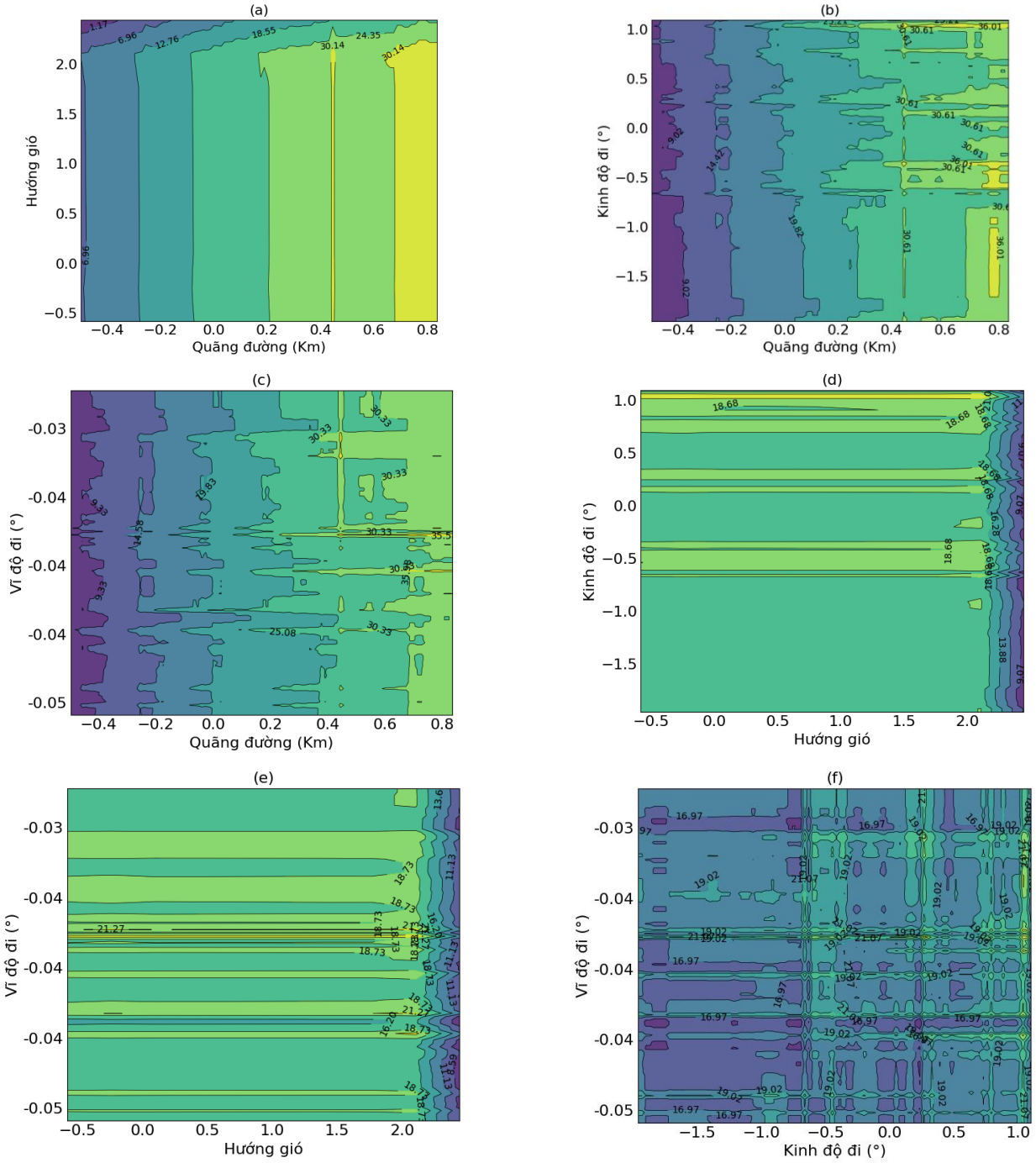

(g)
Dữ liệu đơn lẻ
Giá trị trung bình
Thời gian dự báo (phút)
Quãng đường (Km)
(h)
Lượng mưa (mm)
(i)
Độ ẩm (%)
(j)
Tốc độ gió (Km/h)
(k)
Hướng gió
(l)
Nhiệt độ thấp (°C)
(m)
Nhiệt độ cao (°C)

Hình 7. (tiếp)

Hình 8 trình bày các biểu đồ phụ thuộc từng phần hai chiều (2D PDP) cho bốn biến quan trọng nhất, qua đó làm rõ không chỉ ảnh hưởng riêng lẻ mà còn cả tương tác giữa các yếu tố đầu vào đến thời gian vận chuyển.

Nhìn chung, các bề mặt PDP đều có dạng phi tuyến, xuất hiện các vùng cong, đỉnh và thung lũng rõ rệt, cho thấy biến mục tiêu phụ thuộc mạnh vào sự kết hợp của các biến thay vì từng yếu tố độc lập. Đối với cặp "Kinh độ đi" và "Vĩ độ đi" (Hình 8f), biểu đồ thể hiện một bản đồ nhiệt không gian của khu vực giao nhận, trong đó các vùng có giá trị cao tương ứng với các "điểm nóng" về nguy cơ chậm trễ.

Kết quả này cho thấy rủi ro ùn tắc mang tính cục bộ và phụ thuộc vào sự giao thoa vị trí địa lý, cung cấp cơ sở để tối ưu hóa quy hoạch mạng lưới và vị trí các điểm trung chuyển.



Hình 8. Biểu đồ phụ thuộc từng phần (PDP-2D) mô tả ảnh hưởng của các cặp biến đầu vào đến dự báo giá trị đầu ra

Đối với cặp "Quãng đường" và "Hướng gió" (Hình 8a), bề mặt PDP cho thấy hiệu ứng cộng hưởng rõ rệt: khi khoảng cách ngắn, ảnh hưởng của hướng gió là không đáng kể; tuy nhiên, khi quãng đường tăng, đặc biệt trong điều kiện gió bất lợi, thời gian vận chuyển gia tăng mạnh do tác động tích lũy của lực cản môi trường. Điều này phản ánh bản chất phi tuyến của bài toán và tầm quan trọng của việc xem xét đồng thời các yếu tố vật lý và thời tiết.

Các cặp biến còn lại cũng cho thấy những vùng biến thiên cục bộ, trong đó thời gian vận chuyển chỉ tăng hoặc giảm đáng kể tại các tổ hợp giá trị cụ thể. Sự xuất hiện của các vùng đỉnh và thung lũng khẳng định vai trò của tương tác biến trong việc hình thành các kịch bản vận hành đặc thù, đồng thời chứng minh rằng mô hình đã học được các quy luật phức tạp trong dữ liệu.

Tóm lại, phân tích PDP hai chiều không chỉ xác nhận tính phi tuyến và đa biến của bài toán mà còn cung cấp các hiểu biết trực quan, hỗ trợ hiệu quả cho việc ra quyết định trong vận hành logistics, đặc biệt trong việc nhận diện các điều kiện rủi ro và tối ưu hóa chiến lược điều phối.

5. Kết luận

Nghiên cứu này được thực hiện nhằm giải quyết bài toán cấp thiết về tối ưu hóa chi phí và nâng cao hiệu quả vận hành chuỗi cung ứng tại Việt Nam. Thông qua việc phát triển hệ thống NovaLog AI, nghiên cứu tập trung vào mục tiêu xây dựng một mô hình dự báo chính xác thời gian di chuyển làm cơ sở để cung cấp tham số đầu vào hỗ trợ việc ra quyết định lộ trình tại các trạm trung chuyển thương mại điện tử.

Về mặt phương pháp luận, đóng góp nổi bật của bài báo là việc thiết lập thành công bộ dữ liệu thực chứng đa chiều với 20,493 bản ghi tại khu vực nội thành Hà Nội. Bộ dữ liệu tích hợp đồng thời các đặc trưng về không gian, thời gian và đặc biệt là các điều kiện khí tượng học phức tạp như lượng mưa, độ ẩm, tốc độ và hướng gió. Qua quá trình kiểm định khắt khe 8 thuật toán học máy bằng phương pháp xác thực chéo và 30 lần mô phỏng Monte Carlo, kết quả đã khẳng định sự vượt trội của thuật toán XGBoost. Mô hình này không chỉ kiểm soát tốt hiện tượng quá khớp mà còn đạt độ chính xác và tính ổn định cao nhất với hệ số xác định R^2 đạt 0.9375 và chỉ số A20 đạt 0.8670 trên tập kiểm tra.

Bên cạnh khả năng dự báo, nghiên cứu đã xác định được 4 đặc trưng có sức ảnh hưởng quyết định nhất đến động học thời gian di chuyển bao gồm: quãng đường, hướng gió, kinh độ đi và vĩ độ đi. Trong đó, quãng đường là yếu tố chi phối mạnh mẽ nhất, trong khi hướng gió và tọa độ không gian phản ánh các tác động lũy tích của lực cản môi trường và tình trạng ùn tắc giao thông cục bộ tại khu vực lõi đô thị.

Về mặt ứng dụng, NovaLog AI cung cấp một giải pháp chuyển đổi số thiết thực, giúp các doanh nghiệp logistics tối ưu hóa nguồn lực phương tiện và cam kết thời gian giao hàng với độ tin cậy cao. Việc tối thiểu hóa thời gian xe chạy trên đường còn đóng góp trực tiếp vào mục tiêu phát triển logistics xanh, giúp cắt giảm nhiên liệu tiêu thụ và phát thải carbon. Mặc dù đạt kết quả khả quan, nghiên cứu vẫn tồn tại hạn chế khi dữ liệu mới chỉ tập trung vào đặc trưng khí hậu mùa đông tại Hà Nội. Do đó, các kết quả định lượng về mức độ tác động của các biến số khí tượng học mới chỉ phản ánh đúng quy luật trong điều kiện đặc trưng vi khí hậu mùa đông miền Bắc. Tập dữ liệu này chưa bao hàm tính đa dạng của các hiện tượng thời tiết cực đoan trong cả năm. Vì vậy, nghiên cứu chưa thể ngoại suy hay tổng quát hóa trực tiếp các kết luận về ảnh hưởng của thời tiết hiện tại cho các kiểu hình khí hậu khác trong năm. Trong các giai đoạn tiếp theo, nhóm tác giả hướng tới việc mở rộng độ bao phủ dữ liệu thời gian thực qua công nghệ IoT và kết hợp XGBoost với các mô hình tối ưu hóa bầy đàn để giải quyết trọn vẹn bài toán định tuyến xe.

Lời cảm ơn

Nghiên cứu này là một phần nội dung của đề tài trọng điểm cấp trường năm học 2024 – 2025: "NOVALOG AI: Hệ thống AI dự báo nhu cầu và tối ưu hóa logistic"; Mã số ĐTTĐ-19, Chủ nhiệm đề tài: Triệu Vân Anh.

Nhóm tác giả chân thành cảm ơn sự hỗ trợ

về kinh phí của Trường Đại học Công nghệ Giao thông vận tải để nhóm tác giả thực hiện các nội dung trong đề tài.

Tài liệu tham khảo

[1] F.A. Saraiva, H.T.Y. Yoshizaki. (2024). Logistics demand forecasting: a literature review. *Transportation Research Procedia,* 79, 100–107. https://doi.org/10.1016/j.trpro.2024.03.015

[2] Case Study: Walmart's AI-Enhanced Supply Chain Operations. (2023). *AI Expert Network*. https://aiexpert.network/case-study-walmarts-ai-enhanced-supply-chain-operations/

[3] T.T. Hieu, P.T.H. Phi. (2025). AI Applications in Logistics and Supply Chain Management in Vietnam. *International Journal of Research and Applied Technology*, 5(1), 229-240.

[4] N.P, Quân, T.D. Vinh (2025). Chuyển đổi xanh ngành logistics: Tiềm năng và thách thức từ xe điện. *Kỷ yếu hội thảo Quốc gia: Logistics và vận tải xanh*. NXB Khoa học và Kỹ thuật.

[5] L.T.K. Hoa, N.N. Long. (2025). Application of artificial intelligence in logistics enterprises in Vietnam: A new direction in the industrial revolution 4.0. *Business Administration International Conference.*

[6] D.M. Tran et al. (2025). The nexus of supply chain managerial competence expectation and possession in the new era: the case of Vietnam. *The International Journal of Logistics Management,* 36(7), 63–98. https://doi.org/10.1108/IJLM-11-2023-0488

[7] T.-A. Pham, T. L.-Anh, M.-C. Duong. (2025). A study of green logistics practice in Vietnam – the roles of intellectual capital and digital transformation. *Asia Pacific Journal of Marketing and Logistics*. https://doi.org/10.1108/APJML-02-2025-0265

[8] C.-N. Wang, T.-T. Truong. (2025). Assessing Financial Performance Disparities in the Vietnam Logistics Sector Amidst COVID-19: A Hybrid MCDM Evaluation Approach Using MEREC-COBRA. https://doi.org/10.21203/rs.3.rs-6484793/v1

[9] L.T. Phương. (2024). Ứng dụng trí tuệ nhân tạo trong ngành Logistics và quản lý chuỗi cung ứng trên thế giới và Việt Nam. *Tạp chí công thương*, 9

[10] D.C. Montgomery, E.A. Peck, G.G. Vining. (2021). Introduction to Linear Regression Analysis. John Wiley & Sons.

[11] N. Roustaei. (2024). Application and interpretation of linear-regression analysis. *Medical Hypothesis Discovery and Innovation in Ophthalmology*, 13(3), 151–159. https://doi.org/10.51329/mehdiophthal1506

[12] A.K.M.E. Saleh, M. Arashi, B.M.G. Kibria. (2019). Theory of Ridge Regression Estimation with Applications. John Wiley & Sons. https://doi.org/10.1002/9781118644478

[13] B. de Ville. (2013). Decision trees. *WIREs Computational Statistics,* 5(6), 448–455. https://doi.org/10.1002/wics.1278

[14] S.J. Rigatti. (2017). Random Forest. *Journal of Insurance Medicine,* 47(1), 31-39. https://doi.org/10.17849/insm-47-01-31-39.1

[15] A. Natekin, A. Knoll. (2013). Gradient Boosting Machines, a tutorial. *Frontiers in Neurorobotics*, 7, 21. https://doi.org/10.3389/fnbot.2013.00021

[16] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz. (2019). A Comparative Analysis of XGBoost. https://doi.org/10.48550/arXiv.1911.01914

[17] M. Awad, R. Khanna. (2015). Support Vector Regression. Efficient Learning Machines, 67-80. https://doi.org/10.1007/978-1-4302-5990-9_4

[18] M. Suyal, P. Goyal. (2022). A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning. *International Journal of Engineering Trends and Technology*, 70(7),

43-48. https://doi.org/10.14445/22315381/IJETT-V70I7P205

[19] D. Berrar. (2024). Cross validation. Encyclopedia of Bioinformatics and Computational Biology. https://doi.org/10.1016/B978-0-323-95502-7.00032-4