XÂY DỰNG
TẠP CHÍ ĐIỆN TỬ CỦA BỘ XÂY DỰNG
JOURNAL OF CONSTRUCTION

TẠP CHÍ XÂY DỰNG - eISSN 3030-4482

# Dự báo thời gian di chuyển đô thị bằng thuật toán CatBoost: Phát triển nền tảng NovaLog AI hỗ trợ tối ưu hóa logistics tại Việt Nam

Urban travel time prediction using the CatBoost algorithm: Developing the NovaLog AI platform for logistics optimization in Vietnam

**TS Nguyễn Trọng Dũng[1,*], Nguyễn Thị Thu Hương[1], Phạm Quỳnh Anh[2]**
[1]Nhóm nghiên cứu Ứng dụng công nghệ 4.0 trong giao thông vân tải (I4T), Trường Đại học Công nghệ Giao thông vận tải
[2]Trường Đại học Giáo dục - Đại học Quốc gia Hà Nội
[*]Email: dungnt88@utt.edu.vn

THÔNG TIN BÀI BÁO
Chuyên mục: Khoa học công nghệ
Ngày nhận bài: 21/5/2026
Ngày sửa bài: 05/6/2026
Ngày chấp nhận đăng: 25/6/2026
Ngày xuất bản Online: 23/7/2026
Tác giả liên hệ:
Email: dungnt88@utt.edu.vn

## TÓM TẮT

Sự bất định của thời gian di chuyển đô thị làm giảm hiệu quả điều phối và tăng chi phí logistics tại Việt Nam lên mức 16,8% GDP. Nghiên cứu này đề xuất hệ thống NovaLog AI nhằm dự báo thời gian di chuyển và hỗ trợ tối ưu hóa lộ trình. Dựa trên bộ dữ liệu vi mô gồm hơn 20.000 hành trình với 11 đặc trưng đầu vào, nghiên cứu triển khai phương pháp xác thực chéo kết hợp 30 lần mô phỏng Monte Carlo nhằm giảm độ lệch chọn mẫu. Kết quả cho thấy thuật toán CatBoost đạt $R^2$ = 0,930, RMSE = 2,738 và A20 = 88,30% trên tập kiểm tra. Phân tích SHAP xác định quãng đường, hướng gió, kinh độ đi và vĩ độ đi là 4 yếu tố ảnh hưởng nhiều nhất đến độ trễ. Nghiên cứu cung cấp công cụ định tuyến động, hỗ trợ giảm ùn tắc và tối ưu chi phí vận hành chuỗi cung ứng.

**Từ khóa:** Logistics; dự báo thời gian di chuyển; học máy; CatBoost; Monte Carlo.

## ABSTRACT

The uncertainty of urban travel time reduces coordination efficiency and increases logistics costs in Vietnam to 16.8% of GDP. This study proposes the NovaLog AI system to predict travel time and support route optimization. Based on a microscopic dataset comprising over 20,000 trajectories with 11 input features, the study deploys cross-

validation combined with 30 Monte Carlo simulations to minimize sampling bias. The results show that the CatBoost algorithm achieves $R^2$ = 0.930, RMSE = 2.738, and A20 = 88.30% on the test set. SHAP analysis identifies travel distance, wind direction, departure longitude, and departure latitude as the four factors with the greatest influence on delay. The study provides a dynamic routing tool, supporting congestion reduction and supply chain operational cost optimization.

**Keywords:** Logistics; travel time prediction; machine learning; CatBoost; Monte Carlo.

## 1. ĐẶT VẤN ĐỀ

Trong bối cảnh bùng nổ của thương mại điện tử và những biến động phức tạp của chuỗi cung ứng toàn cầu thời kỳ hậu đại dịch, nhu cầu vận tải hàng hóa đang gia tăng mạnh mẽ, đặt ra áp lực khổng lồ lên hệ thống hạ tầng đô thị [1, 2]. Trên bình diện quốc tế, việc ứng dụng trí tuệ nhân tạo (AI) đã chứng minh hiệu quả vượt trội trong tối ưu hóa vận hành logictic, từ tự động hóa quy trình, nâng cao khả năng phục hồi của chuỗi cung ứng đến độ dự báo chính xác thời gian di chuyển và thời gian hoàn thành chuyến đi [3]. Đặc biệt, việc tích hợp dữ liệu thời gian thực kết hợp với các yếu tố ngoại cảnh như thời tiết và tình trạng giao thông giúp các nhà quản trị đưa ra quyết định điều phối nhanh chóng, chính xác và hiệu quả hơn [4].

Đối với Việt Nam, yêu cầu tối ưu hóa chuỗi cung ứng đang trở nên đặc biệt cấp thiết. Mặc dù ngành Logistics duy trì tốc độ tăng trưởng ổn định khoảng 14 - 16% mỗi năm, chi phí logistics vẫn ở mức cao, chiếm khoảng 16,8 - 18% GDP, vượt xa mức trung bình của nhiều quốc gia phát triển [5 - 7]. Đồng thời, chỉ số hiệu suất logistics (LPI) của Việt Nam cũng giảm xuống vị trí thứ 43 toàn cầu trong năm 2023, cho thấy những hạn chế về năng lực logistics và khả năng theo dõi đơn hàng [8]. Thực trạng này phần lớn xuất phát từ cấu trúc ngành còn phân mảnh, với đa số doanh nghiệp có quy mô nhỏ, nguồn lực công nghệ hạn chế, khiến việc ứng dụng AI trong điều phối và quản lý vận hành vẫn còn thấp. Điều đó đặt ra nhu cầu cấp thiết về các công cụ dự báo và hỗ trợ ra quyết định dựa trên AI.

Tuy nhiên, các nghiên cứu tại Việt Nam hiện nay chủ yếu dừng ở mức tổng quan lý thuyết hoặc phân tích định tính, trong khi còn thiếu các mô hình học máy phục vụ trực tiếp cho vận hành thực tế. Đặc biệt, khoảng trống lớn nằm ở các mô hình dự báo thời gian di chuyển ở cấp độ vi mô, tức từng chuyến đi cụ thể, vốn là nền tảng cho bài toán định tuyến và tối ưu vận hành tại các trạm trung chuyển đô thị [4, 8]. Do đó, việc xây dựng mô hình dự báo chi tiết ở cấp độ chuyến đi là hướng nghiên cứu cần thiết và là cơ sở cho phần phân tích tiếp theo.

Trên cơ sở khoảng trống nghiên cứu và nhu cầu thực tiễn, nghiên cứu này đề xuất hệ thống NovaLog AI nhằm dự báo thời gian di chuyển và hỗ trợ tối ưu hóa vận hành logistics đô thị. Điểm nổi bật là việc xây dựng bộ dữ liệu vi mô thực chứng tại Hà Nội với 11 đặc trưng đầu vào bao quát yếu tố không gian, thời gian và khí tượng. Trên nền dữ liệu đó, nghiên cứu áp dụng thuật toán học máy phù hợp với nhiều biến phân loại, cho phép xử lý hiệu quả các đặc trưng như khung giờ, ngày trong tuần và hướng gió, đồng thời giảm nguy cơ rò rỉ mục tiêu. Kết quả không chỉ đề xuất một khung phương pháp đáng tin cậy cho bài toán dự báo logistics ngắn hạn mà còn cung cấp công cụ hỗ trợ doanh nghiệp chủ động lập kế hoạch, giảm chậm trễ và tối ưu chi phí vận hành.

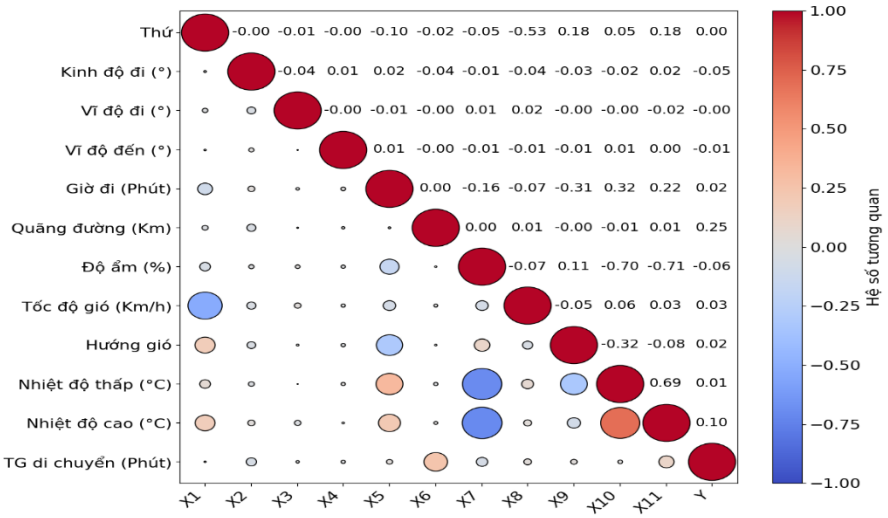
## 2. PHƯƠNG PHÁP NGHIÊN CỨU

### 2.1. Cơ sở dữ liệu

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

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Thứ (2-8) | Kinh độ đi (°) | Vĩ độ đi (°) | Vĩ độ đến (°) | Giờ đi (00:00) | Quãng đường (km) | Độ ẩm (%) | Tốc độ gió (km/h) | Hướng gió (0-8) | Nhiệt độ thấp (°C) | Nhiệt độ cao (°C) | TG di chuyển (phút) |
| Min (0%) | 2,00 | 105,53 | 20,93 | 20,92 | 0,00 | 0,17 | 33,00 | 4,00 | 0,00 | 12,00 | 15,00 |
| 10% | 3,00 | 105,77 | 20,99 | 20,98 | 114,00 | 2,50 | 40,00 | 4,00 | 0,00 | 13,00 | 16,00 |
| 20% | 4,00 | 105,79 | 21,00 | 20,99 | 264,00 | 3,40 | 48,00 | 5,00 | 0,00 | 14,00 | 17,00 |
| 30% | 4,00 | 105,80 | 21,01 | 21,00 | 457,00 | 4,30 | 63,00 | 6,00 | 0,00 | 15,00 | 20,00 |
| 40% | 5,00 | 105,81 | 21,01 | 21,01 | 617,00 | 5,10 | 66,00 | 6,00 | 0,00 | 16,00 | 21,00 |
| Median (50%) | 6,00 | 105,81 | 21,02 | 21,02 | 799,00 | 5,90 | 70,00 | 7,00 | 0,00 | 17,00 | 21,00 |
| 60% | 7,00 | 105,82 | 21,02 | 21,02 | 911,00 | 6,90 | 72,00 | 7,00 | 0,00 | 18,00 | 23,00 |
| 70% | 7,00 | 105,83 | 21,03 | 21,03 | 1.060,00 | 8,10 | 72,00 | 9,00 | 1,00 | 19,00 | 24,00 |
| 80% | 8,00 | 105,84 | 21,04 | 21,03 | 1.263,00 | 9,80 | 78,00 | 10,00 | 2,00 | 20,00 | 24,00 |
| 90% | 8,00 | 105,85 | 21,05 | 21,04 | 1.339,00 | 12,70 | 84,00 | 13,00 | 8,00 | 21,00 | 25,00 |
| Max (100%) | 8,00 | 105,94 | 105,78 | 105,83 | 1.439,00 | 800,00 | 92,00 | 92,00 | 9,00 | 24,00 | 27,00 |
| Mean | 5,69 | 105,81 | 21,18 | 21,03 | 753,07 | 7,13 | 65,78 | 8,15 | 1,73 | 16,98 | 21,22 |

Nghiên cứu sử dụng bộ dữ liệu vi mô thực chứng gồm hơn 20.000 hành trình giao nhận hàng hóa tại khu vực nội thành Hà Nội. Dữ liệu được thu thập trong tháng 12, phản ánh sự biến động giao thông dưới điều kiện vi khí hậu mùa đông. Mỗi quan sát gồm 1 biến mục tiêu là thời gian di chuyển (phút) và 11 đặc trưng đầu vào chia làm 3 nhóm: Không gian (kinh độ đi, vĩ độ đi, vĩ độ đến, quãng đường), thời gian (thứ, giờ đi) và khí tượng (nhiệt độ, độ ẩm, tốc độ gió, hướng gió). Thống kê mô tả chi tiết của các đặc trưng được tổng hợp tại Bảng 1.




Hình 1. Ma trận tương quan giữa các biến đầu vào, đầu ra trong bộ dữ liệu NovaLogAI

Để hiểu rõ các mối quan hệ giữa các biến số này, nghiên cứu tiến hành thiết lập ma trận tương quan (Hình 1). Kết quả từ biểu đồ nhiệt cho thấy "Quãng đường" là biến có hệ số tương quan dương lớn nhất với thời gian di chuyển (0,25). Điều này phù hợp với thực tiễn vật lý của ngành vận tải, khi cự ly càng dài thì thời gian tiêu tốn càng lớn. Đáng chú ý, các biến khí hậu bộc lộ mối tương quan chéo với nhau. Cụ thể, độ ẩm có tương quan âm lớn với nhiệt độ thấp (-0,70) và nhiệt độ cao (-0,71), trong khi hai dải nhiệt độ này lại có tương quan dương chặt chẽ với nhau (0,69). Sự liên kết này cảnh báo nguy cơ xảy ra hiện tượng đa cộng tuyến, đòi hỏi mô hình dự báo phải có năng lực chọn lọc đặc trưng tốt để loại bỏ các thông tin trùng lặp gây nhiễu. Ngược lại, hầu hết các biến số không gian, thời gian và hướng gió, tốc độ gió lại có hệ số tương quan nhỏ với thời gian di chuyển (-0,06 đến 0,05). Đây là minh chứng cho việc các biến số tương tác với nhau theo cơ chế phi tuyến vô cùng phức tạp, khẳng định sự cần thiết của việc áp dụng các thuật toán học máy tiên tiến thay vì mô hình hồi quy tuyến tính truyền thống.

### 2.2. Phương pháp nghiên cứu

Sau bước tiền xử lý, dữ liệu được phân tách ngẫu nhiên thành tập huấn luyện (80%) và tập kiểm tra (20%). Nghiên cứu lựa chọn CatBoost làm thuật toán cốt lõi nhờ ưu điểm xử lý trực tiếp các biến phân loại, ngăn chặn hiện tượng rò rỉ mục tiêu và kiểm soát tốt tình trạng quá khớp đối với dữ liệu có độ nhiễu cao [9]. Để thiết lập cấu hình tối ưu, phương pháp tìm kiếm lưới (Grid Search) được triển khai nhằm tinh chỉnh 4 siêu tham số chính: Độ sâu cây, tốc độ học, số lượng vòng lặp và hệ số chuẩn hóa L2. Độ tin cậy và năng lực tổng quát hóa của hệ thống được kiểm định nghiêm ngặt thông qua kỹ thuật xác thực chéo phân tầng kết hợp với 30 lần chạy mô phỏng Monte Carlo độc lập [10]. Hiệu suất của CatBoost được đối chiếu với các thuật toán tham chiếu (MLP, Extra Trees, AdaBoost, ElasticNet) và được định lượng qua 5 chỉ số thống kê: $R^2$, MAE, RMSE, MAPE và tỷ lệ A20. Cuối cùng, phương pháp SHAP (Shapley Additive Explanations) được tích hợp nhằm phân tích mô hình, định lượng chính xác cả tác động toàn cục lẫn cục bộ của từng đặc trưng đầu vào đối với thời gian di chuyển [11 - 14].

## 3. KẾT QUẢ VÀ THẢO LUẬN

### 3.1. Xây dựng mô hình

Nhằm khai thác tối đa năng lực dự báo và hạn chế rủi ro quá khớp đối với dữ liệu giao thông có độ nhiễu cao, phương pháp Grid Search được triển khai để tinh chỉnh không gian siêu tham số của thuật toán CatBoost. Không gian khảo sát bao gồm: Độ sâu cây {6 - 14}, tốc độ học {0,03, 0,05, 0,08}, số vòng lặp {600 - 1.700} và hệ số chuẩn hóa L2 {3, 7, 9}. Thông qua kiểm định chéo phân tầng, cấu hình tối ưu nhất (CAT_1) thiết lập mức vặn khớp tốt nhất với $R^2CV$ = 0,855, tương ứng với 1.700 vòng lặp, độ sâu cây ở mức 12, tốc độ học 0,08 và hệ số chuẩn hóa L2 = 9. Phân tích động học mô hình cho thấy, thuật toán ưu tiên phát triển cấu trúc cây phân nhánh sâu kết hợp số lượng vòng lặp lớn nhằm có đủ không gian giả thuyết để bóc tách các tương tác phi tuyến phức tạp giữa hệ tọa độ không gian và khí hậu. Tuy nhiên, để triệt tiêu rủi ro học vẹt sinh ra từ cấu trúc sâu và tốc độ học cao, một lực phạt L2 = 9 là cơ chế đối trọng nhằm kiểm soát phương sai phân nhánh. Sự vắng mặt của cơ chế điều chuẩn này ở các cấu hình cây nông hơn (độ sâu 10, L2 = 5) đều dẫn đến giảm hiệu suất. Đánh giá trên tập kiểm tra độc lập, cấu hình CAT_1 chứng minh năng lực tổng quát hóa tốt hơn khi duy trì hệ số $R^2$ đạt 0,9228 và giới hạn sai số RMSE ở mức 7,1970 (so với $R^2$ = 0,9889 và RMSE = 1,3204 trên tập huấn luyện). Sự chênh lệch hiệu năng này nằm trong dung sai thống kê cho phép, khẳng định thuật toán đã xấp xỉ hóa thành công động học của dòng lưu thông mà không bị quá khớp, qua đó hoàn toàn đáp ứng độ tin cậy để vận hành nền tảng NovaLog AI.

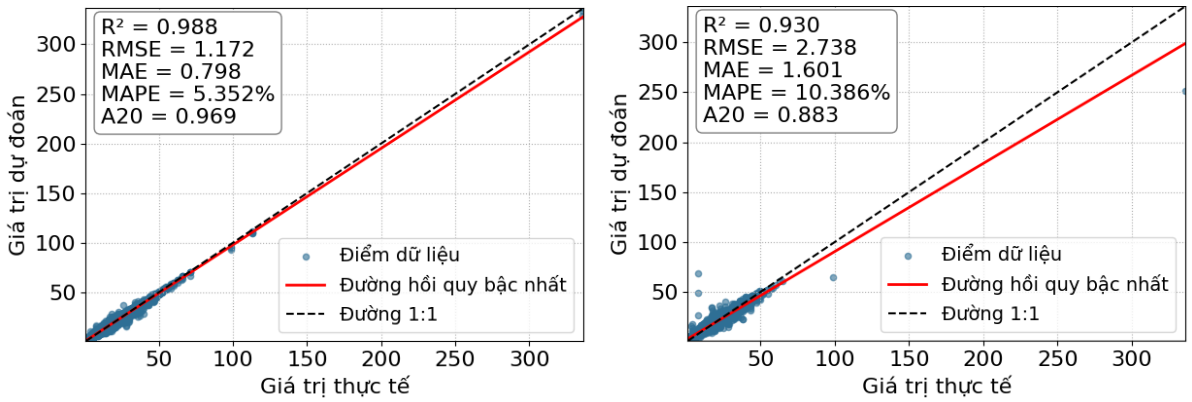
### 3.2. Kết quả dự đoán đại diện

Để định lượng chi tiết độ chính xác và năng lực tổng quát hóa của mô hình CatBoost tối ưu (CAT_1), Bảng 2 trình bày chỉ tiết 5 chỉ tiêu thống kê trên cả hai tập dữ liệu. Trên tập huấn luyện, mô hình dự báo rất cao với hệ số xác định $R^2$ = 0,988 và sai số RMSE xuống mức rất thấp (1,172). Quan trọng hơn, khi xử lý dữ liệu kiểm chứng chưa từng được quan sát trên tập kiểm tra, mô hình vẫn duy trì hiệu suất dự báo với $R^2$ = 0,930. MAE = 1,601 và RMSE = 2,738. Đặc biệt, chỉ số A20 trên tập kiểm tra đạt 0,883, minh chứng rằng có tới 88,3% số chuyến đi được dự báo chính xác với biên độ sai lệch không vượt quá ±20% so với thời gian thực tế. Dưới góc độ quản trị logistics, tỷ lệ an toàn A20 cao vượt trội này cung cấp một cơ sở định lượng vững chắc, giúp các nhà điều phối tự tin thiết lập các biên độ dự phòng thời gian trong bài toán định tuyến phương tiện, từ đó loại bỏ các rủi ro trễ hạn dây chuyền.

Bảng 2. Các chỉ số đánh giá thống kê của mô hình

| Tiêu chí đánh giá | $R^2$ | MAE | MAPE | RMSE | A20 |
|---|---|---|---|---|---|
| Tập huấn luyện | 0,988 | 0,798 | 5,352 | 1,172 | 0,969 |
| Tập kiểm tra | 0,930 | 1,601 | 10,386 | 2,738 | 0,883 |

Khả năng học tập của thuật toán tiếp tục được trực quan hóa thông qua biểu đồ hồi quy tại Hình 2. Đồ thị cho thấy các điểm dữ liệu tập trung dọc theo đường lý tưởng 1:1 (nét đứt) và đường hồi quy bậc nhất (nét liền) trên cả hai tập dữ liệu. Ở tập huấn luyện (Hình 2a), sự trùng khớp giữa đường hồi quy và đường 1:1 khẳng định mô hình đã nắm bắt trọn vẹn cấu trúc phi tuyến của dữ liệu. Đối với tập kiểm tra (Hình 2b), mặc dù xuất hiện sự phân tán nhẹ ở các chuyến đi có thời gian kéo dài đột biến (trên 150 phút), đường hồi quy bậc nhất vẫn duy trì độ dốc bám sát đường tham chiếu. Điều này chứng tỏ CatBoost thật sự đã khái quát hóa được động học phức tạp của dòng lưu thông, kể cả khi gặp phải với các cự ly vận chuyển bất thường.

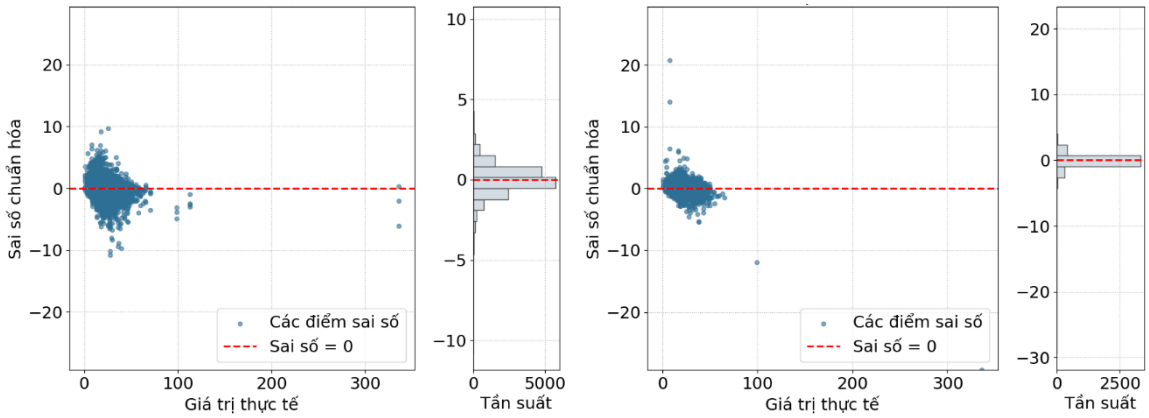



*a) - Tập huấn luyện* *b) - Tập kiểm tra*

Hình 2. Kết quả hồi quy của mô hình trên giá trị dự báo và giá trị thực tế

Hình 3 đi sâu vào phân tích cấu trúc phân phối của các phần dư để kiểm định sâu hơn tính bền vững của các kết quả trên. Kết quả Hình 3a và Hình 3b cho thấy các điểm sai số phân bổ đối xứng và bám sát trục hoành, không xảy ra hiện tượng phương sai thay đổi khi thời gian chuyến đi tăng lên. Kết hợp với biểu đồ tần suất ở cạnh bên, có thể thấy hình thái phân phối của phần dư tiệm cận hoàn hảo với phân phối chuẩn, với đỉnh tháp nhọn tập trung tại giá trị 0. Việc triệt tiêu được các độ chệch hệ thống khẳng định thuật toán CatBoost đã phân tích thành công sự tương tác chéo giữa tọa độ không gian,

thời điểm xuất phát và các lực cản vi khí hậu, biến mô hình này thành một công cụ dự báo có độ tin cậy cao để tích hợp vào hệ thống NovaLog AI.




*a) - Tập huấn luyện* *b) - Tập kiểm tra*

Hình 3. Kết quả sai số dự báo

### 3.3. So sánh các mô hình

Với mục tiêu định lượng khách quan sự vượt trội và phổ quát của mô hình CatBoost, phần này đã triển khai một khung kiểm định thông qua 30 lần chạy mô phỏng Monte Carlo độc lập cho 6 thuật toán học máy. Tại mỗi vòng lặp, dữ liệu được xáo trộn và phân tách lại ngẫu nhiên. Kết quả trung bình từ các vòng lặp được tổng hợp chi tiết tại Bảng 3.

Bảng 3. So sánh Catboost với các thuật toán khác sau khi tối ưu siêu tham số

| Thuật toán | | CatBoost | ExtraTrees | Hist Gradient Boosting | AdaBoost | MLP | ElasticNet |
|---|---|---|---|---|---|---|---|
| Tập huấn luyện | $R^2$ | 0,988 | 0,998 | 0,840 | 0,480 | 0,569 | 0,112 |
| | RMSE | 1,171 | 0,420 | 4,265 | 7,686 | 6,993 | 10,044 |
| Tập kiểm tra | $R^2$ | 0,929 | 0,765 | 0,821 | 0,502 | 0,536 | -0,129 |
| | RMSE | 2,737 | 5,199 | 4,537 | 7,575 | 7,317 | 11,418 |

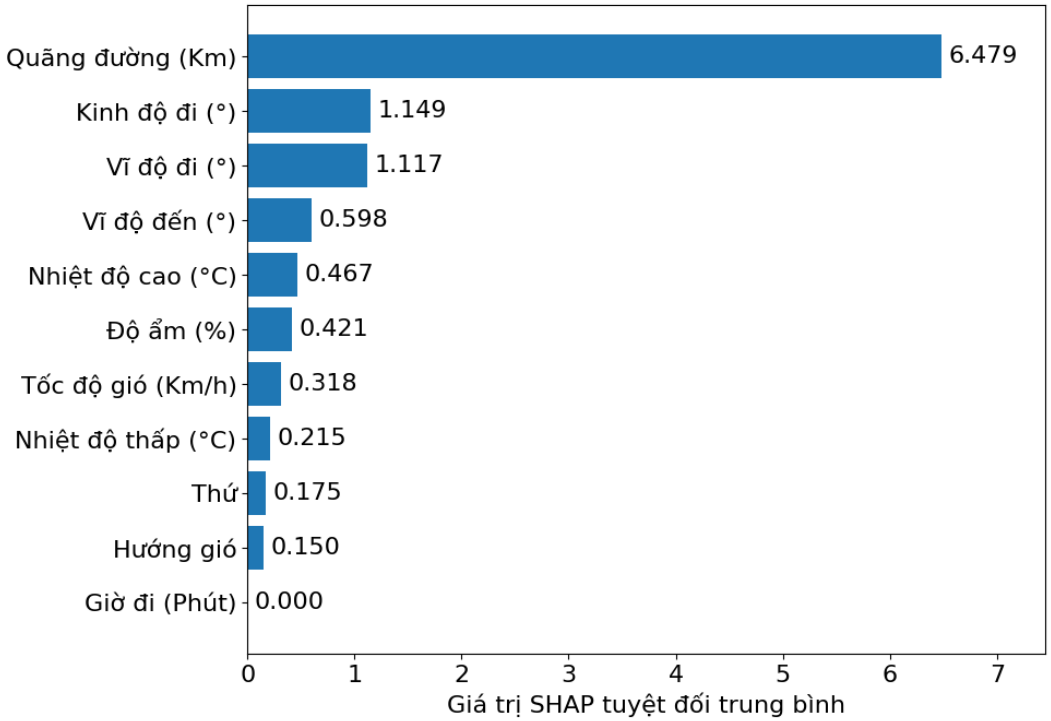
Phân tích các chỉ số thống kê cho thấy sự thất bại rõ rệt của các mô hình tuyến tính truyền thống khi xử lý dữ liệu không - thời gian. Cụ thể, mô hình ElasticNet ghi nhận hiệu suất kém nhất với hệ số xác định $R^2$ trên tập kiểm tra xuống mức âm (-0,129) và sai số RMSE lên tới 11,418. Kết quả này minh chứng bằng cơ sở toán học rằng động học giao thông mang tính phi tuyến cực kỳ phức tạp, hoàn toàn vượt quá khả năng biểu diễn của các hàm tuyến tính. Các thuật toán học máy như mạng nơ-ron truyền thẳng và tăng cường thích nghi cũng bộc lộ nhiều hạn chế với $R^2$ trên tập kiểm tra chỉ đạt mức trung bình (dao động từ 0,5028 đến 0,5361), cho thấy sự thiếu ổn định của các mô hình này khi gặp các giá trị nhiễu ngoại lai đặc thù của dòng giao thông đô thị.

Ngoài ra, phân tích đối chiếu giữa $R^2$ huấn luyện và $R^2$ kiểm tra cho thấy hiện tượng quá khớp của thuật toán cây cực kỳ ngẫu nhiên (ExtraTrees). Mặc dù ExtraTrees thiết lập mức vặn khớp gần như tuyệt đối trên tập huấn luyện ($R^2$ = 0,998, RMSE = 0,420), hiệu suất của mô hình lại suy giảm với biên độ lớn khi đánh giá trên tập kiểm tra ($R^2$ = 0,765, RMSE = 5,199). Sự suy giảm này khẳng định mô hình đã học vẹt các nhiễu động cục bộ của tập dữ liệu huấn luyện thay vì khái quát hóa được quy luật giao thông cốt lõi. Trong khi đó, thuật toán tăng cường độ dốc dựa trên biểu đồ tần suất (HGB) thể hiện tính ổn định tốt hơn giữa hai tập dữ liệu, nhưng năng lực dự báo tổng thể trên tập kiểm tra

vẫn chưa đạt đến ngưỡng tối ưu để ứng dụng thực tiễn ($R^2$ = 0,821). Nhìn chung, mô hình CatBoost chứng minh hiệu suất dự báo và tính phổ quát. Mô hình này không chỉ duy trì độ chính xác cao trên tập huấn luyện mà còn thiết lập đỉnh cao hiệu suất trên tập kiểm tra với $R^2$ = 0,929 và sai số RMSE mức thấp nhất là 2,737. Điều này xác nhận rằng, mô hình CatBoost đã kiểm soát được rủi ro quá khớp, khẳng định cơ chế học tổ hợp tăng cường có thứ tự kết hợp với khả năng xử lý tốt các biến phân loại. Do đó, mô hình CatBoost được khẳng định là mô hình tối ưu để vận hành hệ thống dự báo NovaLog AI.

### 3.4. SHAP

Để hiểu rõ cơ chế học tập của thuật toán CatBoost và định lượng mức độ đóng góp của từng biến số đầu vào đối với thời gian di chuyển, nghiên cứu ứng dụng phương pháp giải thích mô hình SHAP. Khác với các kỹ thuật đánh giá tầm quan trọng truyền thống, SHAP cho phép tính toán chính xác cả tác động tổng thể lẫn mức độ ảnh hưởng cục bộ của từng đặc trưng. Hình 6 biểu diễn tầm quan trọng toàn cục, trong đó các biến được xếp hạng theo giá trị SHAP tuyệt đối trung bình. Kết quả định lượng cho thấy "Quãng đường" là yếu tố mang tính quyết định nhất với giá trị SHAP đạt mức 6,479. Sự chi phối này phù hợp với các nguyên tắc vật lý cơ học trong vận tải: Cự ly chuyến đi càng dài, thời gian tiêu tốn càng lớn. Nằm ở nhóm quan trọng thứ hai là các biến đại diện cho cấu trúc không gian, bao gồm Kinh độ đi (1,149), Vĩ độ đi (1,117) và Vĩ độ đến (0,598). Thứ hạng này khẳng định hệ thống đã nhận diện được sự phức tạp của tọa độ địa lý, nơi các khu vực xuất phát từ vùng trung tâm thành phố sẽ tạo ra lực cản giao thông lớn hơn đáng kể so với khu vực ngoại ô.

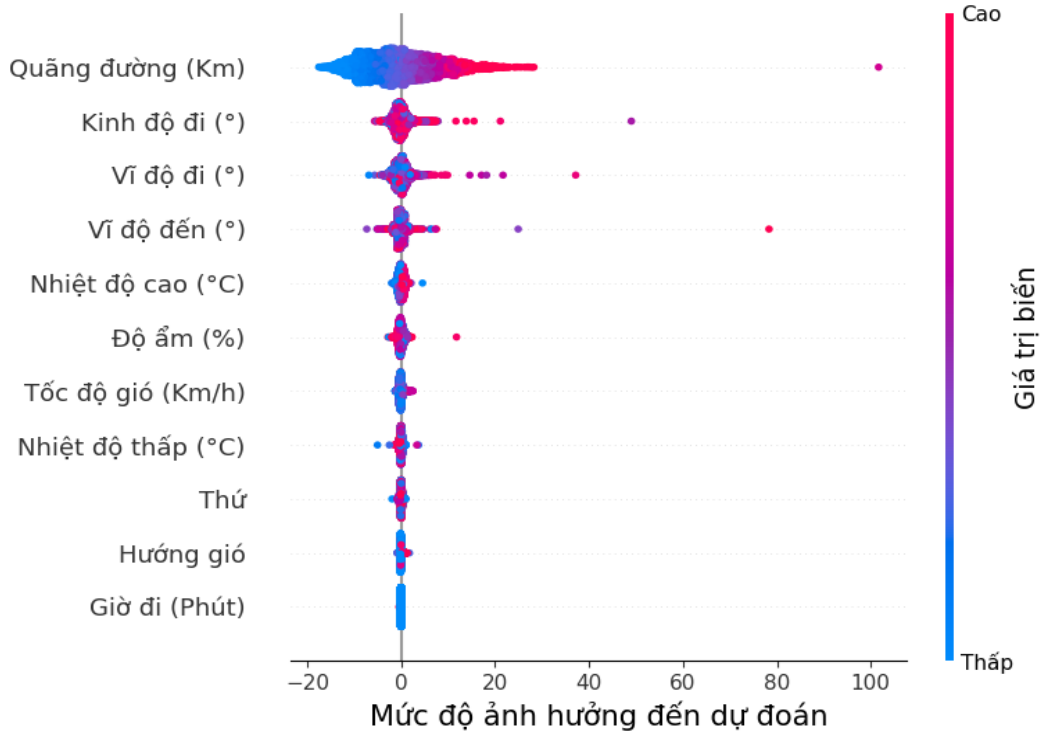



Hình 4. Biểu đồ tầm quan trọng của đặc trưng (SHAP)

Bên cạnh tầm quan trọng tổng thể, Hình 4 đi sâu phân tích chiều hướng và cường độ tác động của từng điểm dữ liệu cụ thể. Trong biểu đồ này, trục hoành biểu diễn mức độ ảnh hưởng đến kết quả dự báo (thời gian tăng hoặc giảm), trong khi dải màu từ xanh đến đỏ cho thấy độ lớn thực tế của biến số. Đối với "Quãng đường", đồ thị thể hiện cấu trúc đồng biến tuyến tính: Các chuyến đi có cự ly dài (tập hợp điểm màu đỏ) đều có giá trị SHAP dương lớn, làm kéo dài thời gian dự báo. Ngược lại, các cự ly ngắn (tập hợp điểm màu xanh) tập trung ở phía âm, làm giảm thời gian di chuyển. Tuy nhiên, đối với các biến tọa độ không gian, đồ thị bộc lộ một cơ chế học phi tuyến phức tạp. Cả các giá

trị cao và thấp đều phân tán rộng ở cả hai phía âm và dương của trục hoành. Điều này chứng minh bằng toán học rằng thời gian chậm trễ không tịnh tiến theo một hướng cố định của vĩ độ hay kinh độ, mà phụ thuộc vào việc tổ hợp tọa độ đó có rơi vào các điểm nóng ùn tắc cục bộ của mạng lưới giao thông đô thị hay không.

Nhóm các biến số vi khí hậu thể hiện mức độ đóng góp trung bình (giá trị SHAP từ 0,150 đến 0,467), đại diện cho các lực cản môi trường đặc thù. Sự phân tán các điểm màu đỏ của biến "Độ ẩm" và "Tốc độ gió" về phía biên dương của trục hoành phản ánh các rủi ro của điều kiện thời tiết: Độ ẩm cao gây trơn trượt hoặc sức gió lớn sẽ cản trở động năng của phương tiện, buộc tài xế phải giảm tốc độ lưu thông, từ đó trực tiếp gia tăng thời gian hoàn thành chuyến đi.




Hình 5. Biểu đồ phân bố mức độ ảnh hưởng của đặc trưng (SHAP)

Một điểm đặc biệt cần phân tích từ kết quả SHAP là biến "Giờ đi" nằm ở vị trí cuối cùng trong bảng xếp hạng Hình 5 với giá trị SHAP biên tiệm cận 0. Nếu chỉ đánh giá dựa trên ảnh hưởng đơn chiều, kết quả này dường như mâu thuẫn với động học giao thông thực tiễn, nơi biên độ thời gian giữa khung giờ cao điểm và thấp điểm là rất lớn. Mặc dù vậy, dưới góc độ học máy, hiện tượng này không đồng nghĩa với việc thông tin thời gian bị bỏ qua, mà phản ánh khả năng lập bản đồ đa chiều của CatBoost. Mô hình đã học được rằng biến "Giờ đi" không hoạt động độc lập mà có sự tương tác chéo mật thiết với các biến không gian. Trên bình diện tổng thể của toàn mạng lưới, tác động của một khung giờ đơn lẻ bị trung bình hóa, bởi vì thời gian di chuyển chỉ thật sự gia tăng đột biến khi phương tiện đi qua những dải tọa độ cụ thể thuộc vùng trung tâm thành phố vào đúng các khung giờ cao điểm. Sự phân mảnh trọng số này khẳng định mô hình CatBoost không đánh giá các biến số một cách cô lập mà đã giải mã thành công sự cộng hưởng phi tuyến giữa không gian và thời gian, cung cấp một công cụ dự báo sát với thực tế vận hành.

## 4. KẾT LUẬN

Trong bối cảnh hệ thống hạ tầng giao thông đô thị ngày càng chịu nhiều áp lực, sự bất định về thời gian di chuyển đã trở thành rào cản lớn đối với năng lực tối ưu hóa chi

phí của ngành Logistics Việt Nam. Để giải quyết thách thức này, nghiên cứu đã phát triển thành công nền tảng NovaLog AI dựa trên thuật toán CatBoost để dự báo thời gian di chuyển thực chứng từ bộ dữ liệu vi mô đa chiều. Thông qua khung đánh giá nghiêm ngặt kết hợp giữa xác thực chéo và 30 lần mô phỏng Monte Carlo, nghiên cứu đã giảm thiểu ảnh hưởng của chệch chọn mẫu và nâng cao độ tin cậy của kết quả dự báo. Kết quả cho thấy sự vượt trội của mô hình CatBoost so với các mô hình tuyến tính và mạng nơ-ron truyền thống, đạt $R^2$ = 0,930, RMSE = 2,738 và A20 lên tới 88,30% trên tập kiểm tra. Đồng thời, phân tích SHAP đã làm rõ bản chất phi tuyến của động học thời gian di chuyển, chỉ ra rằng độ trễ chủ yếu chịu tác động tổng hợp của quãng đường, điều kiện vi khí hậu và đặc trưng không gian tọa độ, đặc biệt tại các điểm nóng ùn tắc trong giờ cao điểm.

Mặc dù đạt được hiệu suất dự báo cao, nghiên cứu vẫn tồn tại một số hạn chế liên quan đến phạm vi dữ liệu đầu vào. Cụ thể, bộ dữ liệu chỉ được thu thập trong tháng 12, do đó các kết luận định lượng về ảnh hưởng của các biến số khí tượng chủ yếu phản ánh đặc trưng khí hậu mùa đông miền Bắc và chưa đủ cơ sở để khái quát cho các điều kiện thời tiết khác trong năm. Trên cơ sở đó, hướng nghiên cứu tiếp theo sẽ tập trung mở rộng dữ liệu theo mùa và tích hợp trực tiếp đầu ra dự báo của NovaLog AI với các thuật toán tối ưu hóa siêu heuristic, như tối ưu bầy đàn, nhằm hoàn thiện hệ thống hỗ trợ định tuyến phương tiện và xây dựng một hệ sinh thái logistics thông minh, khép kín và thích ứng theo thời gian thực.

**Lời cảm ơn:** Nghiên cứ này là một phần nội dung nằm trong 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 ĐTTĐ2425-19, 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] C.-N. Wang and T.-T. Truong. Assessing Financial Performance Disparities in the Vietnam Logistics Sector Amidst COVID-19: A Hybrid MCDM Evaluation Approach Using MEREC-COBRA, Sep. 09, 2025, In Review, doi: 10.21203/rs.3.rs-6484793/v1.

[2] F. A. Saraiva and H. T. Y. Yoshizaki. Logistics demand forecasting: a literature review. Transp. Res. Procedia, vol. 79, pp. 100-107, 2024, doi: 10.1016/j.trpro.2024.03.015.

[3] A. R. Teixeira, J. V. Ferreira and A. L. Ramos. Intelligent Supply Chain Management: A Systematic Literature Review on Artificial Intelligence Contributions. Information, vol. 16, no. 5, p. 399, May 2025, doi: 10.3390/info16050399.

[4] Lê Thanh Phương. 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, Số 9, Tháng 4 năm 2024.

[5] L. T. K. Hoa and N. N. Long. Application of artificial intelligence in logistics enterprises in vietnam: A new direction in the industrial revolution 4.0, 2025.

[6] Truong Trung Hieu and Pham Trinh Hong Phi. AI Applications in Logistics and Supply Chain Management in Vietnam. Int. J. Res. Appl. Technol., 2025.

[7] Quân T. N. P. and Vinh T. D., ‘Chuyển đổi xanh ngành logistics: Tiềm năng và thách thức từ xe điện’, Kỷ Yếu Hội Thảo Quốc Gia Logistics Và Vận Tải Xanh, Tháng 6 năm 2025.

[8] D. M. Tran et al. The nexus of supply chain managerial competence expectation and possession in the new era: the case of Vietnam. Int. J. Logist. Manag., vol. 36, no. 7,

pp. 63-98, Dec. 2025, doi: 10.1108/IJLM-11-2023-0488.

[9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin. CatBoost: unbiased boosting with categorical features, Jan. 20, 2019, arXiv: arXiv:1706.09516, doi: 10.48550/arXiv.1706.09516.

[10] D. Berrar. Cross-Validation. Encycl. Bioinforma. Comput. Biol., Mar. 2024, doi: 10.1016/B978-0-323-95502-7.00032-4.

[11] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu and N. Mastorakis. Multilayer Perceptron and Neural Networks. WSEAS Trans. CIRCUITS Syst., vol. 8, no. 7, Jul. 2009.

[12] P. Geurts, D. Ernst and L. Wehenkel. Extremely randomized trees. Mach. Learn., vol. 63, no. 1, pp. 3-42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[13] Y. Freund, R. E. Schapire, T. Labs, S. Laboratory, P. Avenue and F. Park. A Short Introduction to Boosting. J. Jpn. Soc. Artif. Intell., vol. 14, no. 5, pp. 771-780, Sep. 1999.

[14] C. De Mol, E. De Vito and L. Rosasco. Elastic-net regularization in learning theory. J. Complex., vol. 25, no. 2, pp. 201-230, Apr. 2009, doi: 10.1016/j.jco.2009.01.002.