NGHIÊN CỨU XÂY DỰNG MÔ HÌNH DỰ BÁO LƯỢNG NHIÊN LIỆU TIÊU THỤ CỦA TÀU BIỂN ỨNG DỤNG TRÍ TUỆ NHÂN TẠO

RESEARCH ON DEVELOPING A PREDICTIVE MODEL FOR FUEL CONSUMPTION OF SHIPS USING ARTIFICIAL INTELLIGENCE

TRẦN HỒNG HÀ*, YE SI THU AUNG

Khoa Máy tàu biển, Trường Đại học Hàng hải Việt Nam

**Email liên hệ: tranhongha@vimaru.edu.vn*

DOI: https://doi.org/10.65154/jmst.2025.i84.883

Tóm tắt

Trong bài báo nghiên cứu này, chúng tôi trình bày một mô hình dự báo mức tiêu thụ nhiên liệu cho tàu biển bằng cách áp dụng thuật toán Random Forest được đào tạo trên các thông số vận hành của tàu trong báo cáo buổi trưa của tàu bờ, chẳng hạn như khoảng cách tàu chạy, tốc độ, tải trọng hàng hóa, chiều cao sóng, tốc độ gió và độ chúi của tàu. Mô hình được tích hợp với giao diện người dùng tương tác trên nền tảng PyQt5 để người vận hành sử dụng theo thời gian thực một cách dễ dàng. Mô hình đạt hiệu suất cao với kết quả thử nghiệm là sai số tuyệt đối trung bình (MAE) là 0,78 tấn và hệ số xác định (R^2) là 0,715, phản ánh khả năng dự báo khá chính xác so với dữ liệu thực tế. Người vận hành có thể dễ dàng nhập dữ liệu đầu vào và nhận kết quả dự báo trong vài giây trên giao diện người dùng và nó cũng cung cấp biểu đồ so sánh giữa giá trị thực tế và giá trị dự báo. Nghiên cứu này cho thấy tiềm năng ứng dụng to lớn của các phương pháp học máy để đóng góp vào quản lý năng lượng hàng hải bằng cách tối ưu hóa hoạt động của tàu và chi phí vận hành.

Từ khóa: *Dự báo tiêu thụ nhiên liệu, Học máy, Random Forest.*

Abstract

In this paper, we present a fuel consumption prediction model for ships by applying Random Forest algorithm that was trained on operating parameters of noon reports of a banker ship, such as distance, speed, cargo load, wave height, wind speed and ship trim. The model is integrated with an interactive user interface on the PyQt5 platform for the operators to use in real time at ease. The model achieved high performance with the test results of a Mean Absolute Error (MAE) of 0.78 tons and a coefficient of determination (R^2) of 0.715, reflecting a fairly accurate predicting ability compared to actual data. The operators can easily enter input data and receive predicted results in a few seconds in the user interface and it also provides comparison charts between actual and predicted values. This study shows that huge potential applications of maching learning methods to contribute for marine energy management by optimizing ship operations and operating costs.

Keywords: *Fuel consumption prediction, Machine learning, Random Forest.*

1. Mở đầu

Trong bối cảnh ngành vận tải biển toàn cầu ngày càng chịu áp lực từ yêu cầu giảm phát thải khí nhà kính và tối ưu hóa chi phí vận hành, việc dự báo chính xác mức tiêu thụ nhiên liệu của tàu biển trở thành một vấn đề mang tính cấp thiết. Nhiên liệu chiếm tỷ trọng lớn nhất trong cơ cấu chi phí khai thác tàu, đồng thời là nguồn phát thải chủ yếu của CO_2 và các khí gây ô nhiễm khác [1]. Do đó, các giải pháp dự báo và tối ưu hóa nhiên liệu không chỉ mang ý nghĩa kinh tế mà còn góp phần đáp ứng các tiêu chuẩn môi trường quốc tế ngày càng khắt khe. Các nghiên cứu gần đây đã cho thấy tiềm năng của các mô hình dự báo dựa trên trí tuệ nhân tạo (Artificial Intelligence - AI) trong việc xử lý dữ liệu vận hành phức tạp của tàu biển. So với các phương pháp truyền thống, mô hình AI có khả năng học từ dữ liệu thực nghiệm, nhận diện các quan hệ phi tuyến và phụ thuộc lẫn nhau giữa các thông số khai thác và điều kiện môi trường [2, 3]. Đặc biệt, các thuật toán như phân phối chuẩn đa biến (Multivariate Normal Distribution - MVN) và thuật toán di truyền (Genetic Algorithm - GA) đã được ứng dụng thành công trong mô phỏng dữ liệu và tối ưu hóa tham số trong nhiều lĩnh vực kỹ thuật [4, 5].

Trong nghiên cứu này, nhóm tác giả tập trung vào việc xây dựng mô hình toán học và ứng dụng trí tuệ nhân tạo nhằm dự báo lượng nhiên liệu tiêu thụ của

TẠP CHÍ ISSN: 1859-316X, e-ISSN: 3093-3161
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

tàu biển, dựa trên dữ liệu khai thác thực tế từ tàu Victoria của Công ty Ashico. Dữ liệu bao gồm các thông số vận hành (tốc độ tàu, tải trọng, quãng đường di chuyển, độ chúi) và các điều kiện môi trường (cấp sóng, cấp gió). Do số lượng dữ liệu gốc còn hạn chế, nhóm nghiên cứu đã tiến hành bổ sung dữ liệu tổng hợp bằng phương pháp MVN, đồng thời áp dụng GA để tối ưu mô hình dự báo. Kết quả kỳ vọng sẽ đóng góp vào việc phát triển công cụ hỗ trợ ra quyết định cho khai thác tàu biển hiệu quả và bền vững.

2. Cơ sở lý thuyết

2.1. Phương pháp thu thập dữ liệu và tạo bộ dữ liệu chuẩn

Nhóm nghiên cứu đã tiến hành thu thập dữ liệu thực nghiệm từ tàu Victoria của công ty Ashico. Bộ dữ liệu bao gồm các thông số khai thác đặc trưng như tốc độ hành trình, trọng tải, quãng đường di chuyển và độ chúi của tàu, kết hợp với các yếu tố môi trường bên ngoài như cấp sóng và cấp gió. Sau quá trình làm sạch và xử lý, tổng cộng có 82 bộ dữ liệu hoàn chỉnh được sử dụng làm cơ sở cho phân tích như trong Bảng 1.

Với mục tiêu khám phá quy luật và mối quan hệ giữa các thông số vận hành và điều kiện thời tiết, nhóm nghiên cứu đã xây dựng một mô hình toán nhằm phân tích cấu trúc tiềm ẩn trong tập dữ liệu. Trên cơ sở đó, hai thuật toán đã được áp dụng:

- Phân phối Chuẩn Đa Biến (Multivariate Normal Distribution - MVN): Dùng để tổng hợp và mở rộng dữ liệu, bảo toàn được đặc trưng thống kê cũng như mối tương quan giữa các biến.

- Thuật toán Di truyền (Genetic Algorithm - GA): Triển khai như một phương pháp tối ưu hóa heuristic, hỗ trợ tìm kiếm và hiệu chỉnh các quy luật phi tuyến phức tạp trong dữ liệu.

Sự kết hợp giữa hai phương pháp này không chỉ giúp khắc phục hạn chế về quy mô dữ liệu, mà còn tạo nền tảng cho việc xây dựng các mô hình dự báo chính xác và có khả năng ứng dụng trong thực tiễn khai thác tàu biển. Tổng hợp dữ liệu phân phối chuẩn đa biến MVN là một phương pháp thống kê nhằm tạo ra các tập dữ liệu tổng hợp dựa trên giả định rằng các biến quan sát tuân theo phân phối chuẩn đa biến (MVN). Trong khuôn khổ bài báo này, vector trung bình và ma trận hiệp phương sai được ước lượng từ dữ liệu gốc. Các mẫu tổng hợp sau đó được sinh ra từ phân phối MVN tương ứng nhằm duy trì các đặc tính thống kê của dữ liệu ban đầu, chẳng hạn như phương sai và mối tương quan giữa các biến. Để xử lý các vấn đề bất ổn số học, bao gồm tính suy biến hoặc giá trị riêng âm trong ma trận hiệp phương sai, một thủ tục điều chuẩn được áp dụng. Điều này thường bao gồm việc cộng thêm một hằng số dương nhỏ vào các phần tử đường chéo hoặc cắt ngưỡng giá trị riêng, nhằm đảm bảo rằng ma trận hiệp phương sai duy trì được tính xác định dương và khả nghịch. Nhờ đó, phương pháp này cho phép tạo ra dữ liệu một cách ổn định và đáng tin cậy, đồng thời vẫn bảo toàn được cấu trúc tương quan của tập dữ liệu gốc

Trong Bảng 2 cho thấy sự sai lệch tương quan giữa tập dữ liệu gốc và tập dữ liệu tổng hợp được tạo ra thông qua thuật toán di truyền GA cho thấy những hiểu biết quan trọng về mức độ trung thực của dữ liệu tổng hợp. Sai số tuyệt đối trung bình (MAE) giữa các ma trận tương quan được ghi nhận ở mức 0,2306, cho thấy rằng, xét trên tổng thể, tập dữ liệu tổng hợp dựa trên GA duy trì cấu trúc tương quan của tập dữ liệu gốc ở mức tương đối tốt. Tuy nhiên, vẫn tồn tại các sai lệch cục bộ đáng kể đối với một số cặp biến. Sai lệch lớn nhất được quan sát thấy ở cặp biến (3,25, 18,40)

Bảng 1. Bộ dữ liệu được thu thập trên tàu

distance_nm	speed_knots	cargo_tons	wave_height_m	wind_speed_knots	trim_deg	fuel_consumed_tons
18	9	9417	3,25	24,3	-0,02	0,88
226	9,4	9417	3,25	18,4	-0,02	10,5
257	10,7	9417	3,25	24,3	-0,02	10,48
229	9,54	9417	3,25	24,3	-0,02	10,49
219	9,1	9417	3,25	24,3	-0,02	10,5
60	9,2	2340	1,88	18,4	-2,5	9,2
236	9,7	2340	1,88	18,4	-2,5	9,79
248	10	2340	1,88	18,4	-2,5	9,8
240	10		1,88	18,4	-2,35	9,8
200	8,9	7000	0,88	18,4	-0,94	9,375
224	9,33	7000	0,88	12,7	-0,94	9,96
203	8,7	7000	0,88	12,7	-0,94	9,95
202	8,7	7000	0,88	12,7	-0,94	9,98
246	8,9	7000	0,88	7,8	-0,94	10,208
227	9	7000	0,3	7,8	-0,94	9,99
237	9,1	7000	0,3	7,8	-0,94	9,98
237	9,9	7000	0,3	7,8	-0,94	9,98
244	9,3	7000	0,3	7,8	-0,94	10,417
235	9,8	7000	0,88	12,7	-0,94	10
218	9,1	7000	0,3	7,8	-0,94	10

với độ chênh lệch tương quan tuyệt đối là 0,9630, cho thấy gần như sự đảo ngược hoặc mất mối quan hệ gốc. Tương tự, các cặp như (226,00, 10,50) và (9417,0, -0,02) cũng thể hiện sai lệch đáng kể (0,9084 và 0,8349), nhấn mạnh thách thức trong việc duy trì sự phụ thuộc giữa các biến trong quá trình tạo dữ liệu tổng hợp. Ngược lại, một số cặp biến như (9,40, 9417,00) và (12,00, -0,02) lại cho thấy sai lệch thấp hơn (0,4332 và 0,2053), phản ánh khả năng của GA trong việc bảo toàn tốt hơn các mối quan hệ này.

Bảng 2. Top 5 cặp biến có độ lệch tương quan lớn nhất giữa dữ liệu gốc và dữ liệu synthetic (GA)

STT	Var1	Var2	AbsDiff
1	3.25	18.40	0.963013
2	226.00	10.50	0.908392
3	9417.0	-0.02	0.834882
4	9.40	9417.00	0.433167
5	12.00	-0.02	0.205282

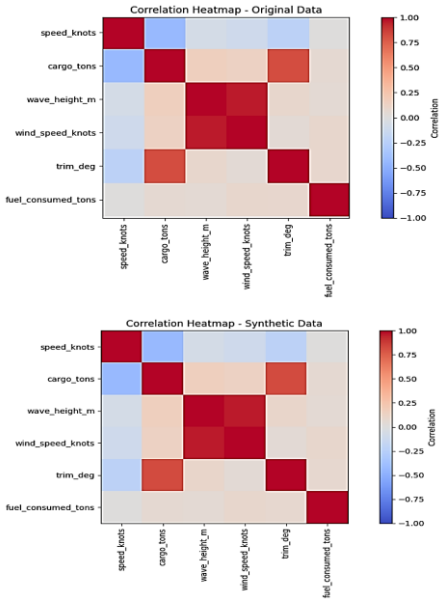
Số liệu trong Bảng 3 là tương quan giữa bộ dữ liệu gốc và bộ dữ liệu tổng hợp cho thấy mức độ tương đồng rất cao. Sai số tuyệt đối trung bình (MAE) giữa hai ma trận tương quan đạt 0,0072, cho thấy bộ dữ liệu tổng hợp đã tái hiện thành công các quan hệ phụ thuộc nền tảng của dữ liệu gốc với sai lệch chỉ ở mức tối thiểu. Cụ thể, năm cặp biến có độ sai khác lớn nhất dao động trong khoảng 0,0104 đến 0,0219, một giá trị có thể coi là không đáng kể trong thực tiễn. Đáng chú ý, cặp biến (226,00, 10,50) thể hiện sai lệch lớn nhất (0,0219), nhưng mức chênh lệch này vẫn rất nhỏ khi xét trong toàn bộ cấu trúc tương quan. So với các lần thử nghiệm trước đây, khi MAE đạt mức cao hơn đáng kể (≈0,2306), kết quả hiện tại chứng minh sự cải thiện rõ rệt về mức độ trung thực của quá trình sinh dữ liệu tổng hợp. Những phát hiện này củng cố tính hợp lệ của quy trình tiền xử lý và tối ưu hóa đã áp dụng, khẳng định rằng bộ dữ liệu tổng hợp có thể được xem như một đại diện đáng tin cậy cho bộ dữ liệu gốc. Điều này không chỉ nâng cao tiềm năng ứng dụng trong các phân tích tiếp theo mà còn góp phần giải quyết các vấn đề liên quan đến bảo mật và quyền riêng tư dữ liệu.

Bảng 3. Top 5 cặp biến có độ lệch tương quan lớn nhất giữa dữ liệu gốc và dữ liệu synthetic (MVN)

STT	Var1	Var2	AbsDiff
1	226.00	10.50	0.021888
2	9417.00	-0.02	0.019085
3	18.40	10.50	0.012726
4	34.00	10.50	0.012098
5	3.25	10.50	0.010405

2.2. Mô hình toán tạo ra bộ dữ liệu huấn luyện

Hình 1 trình bày bản đồ nhiệt ma trận tương quan giữa các biến vận hành và môi trường trong bộ dữ liệu gốc. Thang màu trải dài từ xanh đậm, biểu thị các tương quan âm mạnh (giá trị gần -1), đến đỏ sẫm, đại diện cho các tương quan dương mạnh (giá trị gần +1). Một số mô hình đáng chú ý có thể được quan sát. Thứ nhất, tốc độ tàu (speed_knots) thể hiện mối tương quan âm với tải trọng hàng hóa (cargo_tons) và góc trim (trim_deg), cho thấy khi tải trọng và độ nghiêng trim tăng thì tốc độ tàu có xu hướng giảm. Thứ hai, tồn tại mối tương quan dương mạnh giữa cargo_tons và trim_deg, phản ánh thực tế rằng tải trọng lớn hơn thường đi kèm với độ nghiêng trim cao hơn. Thứ ba, các yếu tố môi trường như chiều cao sóng (wave_height_m) và tốc độ gió (wind_speed_knots) có mối tương quan dương, phù hợp với điều kiện hải dương học thực tế khi gió mạnh thường đi kèm với sóng cao.



Hình 1. Bản đồ nhiệt ma trận tương quan giữa các biến vận hành và môi trường trong bộ dữ liệu gốc, bộ dữ liệu được tạo ra

Đáng chú ý, biến mục tiêu là lượng tiêu thụ nhiên liệu (fuel_consumed_tons) cho thấy mối tương quan dương ở mức trung bình với hầu hết các yếu tố vận hành và môi trường. Điều này chỉ ra rằng mức tiêu thụ nhiên liệu chịu ảnh hưởng đồng thời bởi tốc độ tàu, tải trọng, điều kiện sóng, cường độ gió và độ nghiêng trim. Tổng thể, các phát hiện này khẳng định rằng nhiều yếu tố tương tác cùng lúc tác động đến tiêu thụ nhiên liệu, qua đó nhấn mạnh sự cần thiết của các phương pháp mô hình hóa đa biến nhằm dự đoán và tối ưu hóa hiệu quả sử dụng năng lượng của tàu.

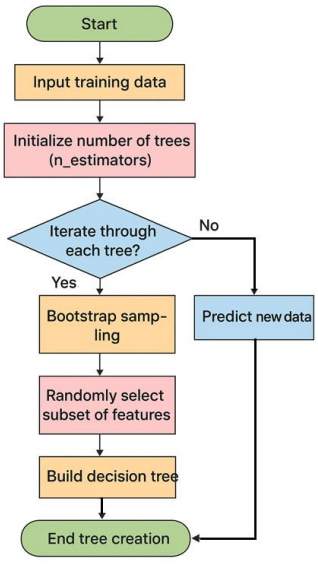
TẠP CHÍ ISSN: 1859-316X, e-ISSN: 3093-3161
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

Nhìn chung, dữ liệu tổng hợp đã tái tạo thành công các cấu trúc phụ thuộc chính quan sát được trong bộ dữ liệu gốc. Đáng chú ý, các mối tương quan âm giữa speed_knots với cargo_tons và trim_deg được duy trì, phản ánh các đánh đổi vận hành thực tế khi tải trọng và độ nghiêng mạn tàu (trim) tăng thường dẫn đến giảm tốc độ tàu. Tương tự, mối liên hệ dương mạnh giữa cargo_tons và trim_deg vẫn hiện diện, cho thấy tải trọng lớn hơn thường gắn liền với việc điều chỉnh độ nghiêng nhiều hơn. Các biến môi trường như wave_height_m và wind_speed_knots tiếp tục thể hiện mối tương quan dương mạnh, phù hợp với các điều kiện hải dương học khi gió mạnh thường tạo ra sóng cao hơn.

Bản đồ nhiệt chênh lệch tuyệt đối (absolute difference heatmap) giữa ma trận tương quan của dữ liệu gốc và dữ liệu tổng hợp. Kết quả cho thấy toàn bộ các ô ma trận có giá trị gần bằng 0, được thể hiện bằng tông màu xám đồng nhất. Điều này chứng tỏ sự khác biệt về hệ số tương quan giữa hai bộ dữ liệu là rất nhỏ và không có sai lệch đáng kể nào trong việc bảo toàn cấu trúc quan hệ giữa các biến. Đáng chú ý, việc tái tạo chính xác các mối tương quan cả về độ lớn và chiều hướng cho thấy phương pháp sinh dữ liệu tổng hợp dựa trên phân phối Gaussian đa biến đã hoạt động hiệu quả trong việc duy trì các đặc tính thống kê của tập dữ liệu ban đầu. Do đó, dữ liệu tổng hợp có thể được xem là đại diện đáng tin cậy, thích hợp cho các ứng dụng phân tích và mô hình dự báo mà không gây ra sai lệch thống kê đáng kể so với dữ liệu gốc. Khi so sánh ba bản đồ nhiệt tương quan, có thể nhận thấy sự tương đồng rõ rệt giữa dữ liệu gốc và dữ liệu tổng hợp. Heatmap của dữ liệu gốc thể hiện các mối quan hệ nội tại giữa các biến, chẳng hạn như mối liên hệ mạnh giữa lượng hàng hóa, tốc độ gió và mức tiêu thụ nhiên liệu. Heatmap của dữ liệu tổng hợp tái tạo gần như toàn bộ các mô hình này, cho thấy phương pháp sinh dữ liệu duy trì được cấu trúc thống kê cốt lõi. Bản đồ nhiệt chênh lệch tuyệt đối, với các giá trị gần bằng 0 trên toàn bộ ma trận, củng cố thêm bằng chứng rằng các mối quan hệ giữa các biến đã được bảo toàn một cách nhất quán. Sự nhất quán giữa ba heatmap chứng minh rằng dữ liệu tổng hợp không chỉ phản ánh các phân phối biên mà còn tái hiện chính xác các mối quan hệ tương quan đa biến, do đó có thể được xem là nguồn dữ liệu thay thế đáng tin cậy trong các nghiên cứu phân tích và mô hình hóa.

2.3. Thuật toán Random Forest

Lưu đồ trên mô tả quy trình nghiên cứu và thực hiện mô hình Random Forest, một phương pháp học máy tiên tiến được sử dụng phổ biến trong các bài toán phân loại và hồi quy. Quá trình bắt đầu với việc nhập dữ liệu huấn luyện, bao gồm các đặc trưng đầu vào và nhãn mục tiêu. Tiếp theo, mô hình được khởi tạo số lượng cây (n_estimators) - đây là tham số quyết định quy mô và độ chính xác của mô hình. Trong giai đoạn lặp qua từng cây, hệ thống thực hiện lấy mẫu bootstrap, tức là chọn ngẫu nhiên có hoàn lại từ tập dữ liệu gốc để tạo ra các tập con huấn luyện riêng biệt cho mỗi cây, giúp tăng tính đa dạng của mô hình. Sau đó, mô hình chọn ngẫu nhiên một tập con đặc trưng tại mỗi nút chia của cây để giảm tương quan giữa các cây và hạn chế hiện tượng overfitting. Tiếp đến, từng cây quyết định được xây dựng dựa trên tập dữ liệu và đặc trưng đã chọn. Khi một cây hoàn thành, quy trình quay lại bước lặp để tiếp tục huấn luyện các cây còn lại cho đến khi đạt đủ số lượng đã định. Sau khi toàn bộ rừng cây được tạo, mô hình tiến hành dự đoán dữ liệu mới, trong đó kết quả đầu ra được tổng hợp từ tất cả các cây bằng cách bình chọn đa số (đối với phân loại) hoặc lấy trung bình giá trị dự đoán (đối với hồi quy). Quy trình kết thúc khi mô hình sẵn sàng áp dụng cho các bài toán thực tiễn. Toàn bộ quá trình thể hiện rõ bản chất của Random Forest: Kết hợp nhiều mô hình yếu (các cây quyết định) thành một mô hình mạnh, nhờ đó cải thiện độ chính xác, tăng khả năng khái quát hóa và giảm thiểu sai số trong dự đoán.



Hình 2. Lưu đồ thuật toán của mô hình Random

Random Forest là một mô hình ensemble learning dựa trên tập hợp nhiều cây quyết định (Decision Trees). Đối với bài toán hồi quy, công thức tổng quát được viết như sau [6]:

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (1)$$

Trong đó:

$\hat{y}(x)$: Giá trị dự báo của mô hình tại điểm dữ liệu;

B: số lượng cây trong rừng (n_estimators);

$T_b(x)$: Kết quả dự báo của cây quyết định thứ b.

Mỗi cây T_b được huấn luyện trên một mẫu bootstrap từ tập dữ liệu được tạo ra, đồng thời tại mỗi nút chia, một tập con ngẫu nhiên các đặc trưng được lựa chọn để giảm tương quan giữa các cây.

3. Huấn luyện mô hình

Với 10000 bộ dữ liệu được tạo ra, các mẫu này được đưa vào mô hình Random Forest để huấn luyện.

Quá trình huấn luyện

Tạo tập bootstrap: Từ tập dữ liệu gốc, B tập con ngẫu nhiên bằng phương pháp lấy mẫu có hoàn lại.

Huấn luyện cây quyết định: Với mỗi tập bootstrap, huấn luyện một cây hồi quy: Tại mỗi nút, lựa chọn ngẫu nhiên m đặc trưng từ tổng số p đặc trưng. Chọn đặc trưng và ngưỡng chia giúp giảm thiểu độ lệch bình phương trung bình (MSE):

$$MSE(S) = \frac{1}{|S|}\sum_{i\in S}(y_i - \bar{y}_S)^2 \quad (2)$$

Trong đó:

$|S|$: Là số lượng mẫu có trong nút đó;

S: Là tập con các mẫu (subset) tại nút hiện tại trong cây quyết định;

y_i: Là giá trị mục tiêu (target) thực tế của mẫu thứ i;

$\bar{y}_S$: Là giá trị trung bình của tất cả các nhãn trong tập con S.

Để đánh giá mô hình sử dụng sai số trung bình tuyệt đối:

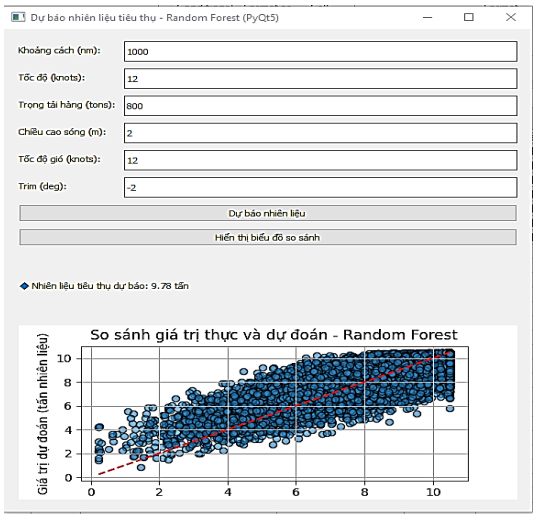
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

Hệ số xác định R^2 để đo mức độ biến thiên của dữ liệu:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Các chỉ số đánh giá cho thấy sai số tuyệt đối trung bình (MAE) đạt mức 0,78 tấn, trong khi hệ số xác định (R^2) đạt 0,715. Điều này chứng tỏ mô hình có khả năng dự báo tương đối chính xác, khi sai số so với dữ liệu thực tế ở mức thấp và mức độ giải thích phương sai dữ liệu đạt trên 70%. Kết quả này khẳng định tiềm năng của việc ứng dụng học máy, đặc biệt là mô hình Random Forest, trong việc xây dựng công cụ dự báo nhiên liệu phục vụ tối ưu hóa khai thác tàu biển.

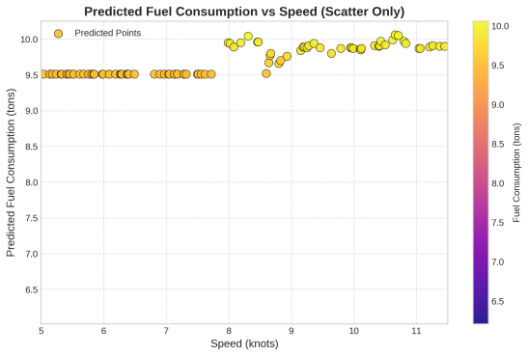
Hình 3 trên minh họa giao diện phần mềm dự báo nhiên liệu tiêu thụ của tàu biển được phát triển trên nền tảng PyQt5 và tích hợp mô hình Random Forest. Người dùng có thể nhập các thông số vận hành quan trọng như khoảng cách hành trình, tốc độ, trọng tải hàng hóa, chiều cao sóng, tốc độ gió và góc trim để mô hình đưa ra kết quả dự báo.



Hình 3. Các ảnh được xử lý và dãn nhãn

4. Kết quả và thảo luận

Hình 4 thể hiện biểu đồ phân tán mối quan hệ giữa tốc độ tàu (Speed, tính bằng knots) và mức tiêu thụ nhiên liệu dự đoán (Predicted Fuel Consumption, tính bằng tấn). Các điểm dữ liệu được mã hóa màu theo giá trị tiêu thụ nhiên liệu, với thang màu từ tím (thấp) đến vàng (cao). Kết quả cho thấy mức tiêu thụ nhiên liệu nằm chủ yếu trong khoảng 9,5-10 tấn, với sự phân bố khá đồng đều ở các dải tốc độ từ 6 đến 11 knots.



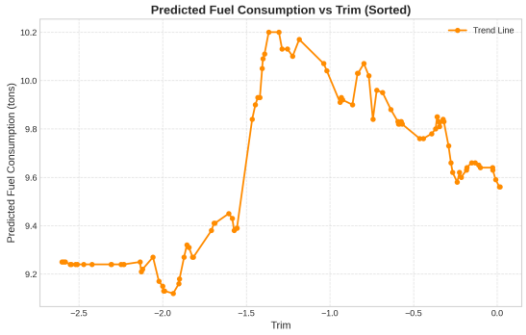
Hình 4. Ảnh hưởng của tốc độ tàu tới lượng tiêu thụ nhiên liệu

Tuy nhiên, tại vùng tốc độ thấp (khoảng 6-7 knots), dữ liệu có xu hướng tập trung mạnh ở mức 9,5 tấn, thể hiện sự ổn định cao. Ngược lại, khi tốc độ tăng trên 8knots, mức tiêu thụ có độ dao động nhẹ quanh giá trị

TẠP CHÍ ISSN: 1859-316X, e-ISSN: 3093-3161
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

trung bình 9,6-9,8 tấn. Điều này gợi ý rằng trong phạm vi tốc độ khảo sát, tiêu thụ nhiên liệu không biến thiên nhiều theo tốc độ, và có thể bị chi phối bởi các yếu tố khác ngoài tốc độ tàu.

Hình 5 thể hiện mối quan hệ giữa độ trim và mức tiêu thụ nhiên liệu dự đoán. Dữ liệu đã được sắp xếp theo giá trị trim và được biểu diễn bằng đường xu hướng liên tục. Kết quả cho thấy khi trim nằm trong khoảng từ -2,5 đến -2,0m, mức tiêu thụ nhiên liệu duy trì ổn định quanh giá trị thấp (≈ 9,2 tấn). Tuy nhiên, khi trim tăng từ -2,0 lên khoảng -1,5m, tiêu thụ nhiên liệu tăng nhanh, đạt cực đại trên 10,2 tấn. Sau đó, khi trim tiếp tục tiến gần về 0m, mức tiêu thụ nhiên liệu giảm dần và dao động quanh 9,6-9,8 tấn. Xu hướng này cho thấy mối liên hệ phi tuyến tính rõ rệt giữa trim và hiệu suất nhiên liệu, trong đó tồn tại một vùng trim bất lợi (khoảng -1,5 đến -1,0m) gây ra tiêu hao nhiên liệu cao nhất. Điều này gợi ý rằng việc tối ưu hóa trim có thể đóng vai trò quan trọng trong việc giảm thiểu chi phí nhiên liệu và nâng cao hiệu quả vận hành tàu.



Hình 5. Ảnh hưởng của trim tới lượng tiêu thụ nhiên liệu

6. Kết luận

Nghiên cứu này đã đề xuất và triển khai mô hình dự báo lượng nhiên liệu tiêu thụ của tàu biển dựa trên bộ dữ liệu được tạo ra sử dụng thuật toán phân phối chuẩn đa biến MVN và thuật toán di truyền GA. Phương pháp MVN được sử dụng để sinh dữ liệu tổng hợp nhằm mở rộng bộ dữ liệu gốc, đồng thời bảo toàn đặc trưng thống kê và mối tương quan giữa các biến và có sai số nhỏ hơn so với thuật toán GA. Trong khi đó, GA được áp dụng để tối ưu hóa các tham số dự báo, nâng cao độ chính xác và giảm sai số so với các phương pháp truyền thống.

Kết quả mô phỏng dự báo tiêu thụ nhiên liệu bằng mô hình Random Forest cho thấy mô hình có khả năng dự đoán tương đối chính xác mối quan hệ giữa Trim (độ chúi của tàu) và mức tiêu thụ nhiên liệu. Đường xu hướng trong đồ thị thể hiện sự biến thiên rõ ràng của lượng nhiên liệu tiêu thụ theo giá trị Trim, phản ánh độ nhạy của hệ thống đối với các thay đổi trong điều kiện vận hành. Các chỉ số đánh giá cho thấy sai số tuyệt đối trung bình (MAE) đạt 0,78 tấn, chứng tỏ sai lệch giữa giá trị dự đoán và thực tế ở mức thấp, đảm bảo độ tin cậy cho mô hình trong bối cảnh dự báo kỹ thuật. Bên cạnh đó, hệ số xác định (R^2) đạt 0,715, cho thấy mô hình giải thích được khoảng 71,5% biến thiên của dữ liệu thực nghiệm, thể hiện mức độ phù hợp khá cao giữa giá trị dự đoán và giá trị quan sát. Như vậy, mô hình Random Forest được xem là một công cụ dự báo hiệu quả cho bài toán ước lượng tiêu thụ nhiên liệu theo Trim, giúp hỗ trợ ra quyết định tối ưu hóa vận hành tàu nhằm giảm chi phí và nâng cao hiệu quả năng lượng.

Tổng thể, kết quả nghiên cứu khẳng định rằng trí tuệ nhân tạo, khi được kết hợp với các phương pháp thống kê và tối ưu hóa, có thể trở thành một công cụ hữu hiệu trong dự báo nhiên liệu tiêu thụ, hướng tới khai thác tàu biển hiệu quả và bền vững hơn.

TÀI LIỆU THAM KHẢO

[1] International Maritime Organization (IMO). (2020). *Fourth IMO GHG Study 2020*. IMO.

[2] Nguyen, T. H., & Lee, Y. H. (2021). *Machine learning approaches for ship fuel consumption prediction*. Ocean Engineering, Vol.233, 109089.

[3] Zhang, H., Liu, Z., & Wang, J. (2019). *Artificial intelligence applications in ship energy efficiency: A review.* Applied Energy, Vol.255, 113801.

[4] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[5] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

[6] Breiman, L. (2001). Random Forests. *Machine Learning*, Vol.45(1), pp.5-32.

Ngày nhận bài: 30/09/2025
Ngày nhận bản sửa lần 1: 12/10/2025
Ngày nhận bản sửa lần 2: 17/10/2025
Ngày duyệt đăng: 17/10/2025