

Áp dụng mô hình học máy trong lập dự toán các công trình thủy lợi ở Việt Nam

Lê Minh Thoa^{1*}

¹ Trường Đại học Thủy lợi

TỪ KHOÁ

Học máy
Dự toán
Công trình thủy lợi
Việt Nam

TÓM TẮT

Trong bối cảnh phát triển hạ tầng nông nghiệp và thích ứng với biến đổi khí hậu, các công trình thủy lợi ở Việt Nam ngày càng có quy mô lớn, kỹ thuật phức tạp và chịu nhiều yếu tố bất định như điều kiện địa chất, thủy văn, vật liệu và biến động giá. Do vậy cần phải tiếp cận mô hình học máy nhằm nâng cao độ chính xác trong lập dự toán đầu tư xây dựng. Mô hình học máy, đặc biệt là Random Forest, có độ chính xác cao hơn đáng kể so với phương pháp truyền thống. Nghiên cứu phân tích tầm quan trọng của các biến đầu vào, chỉ ra rằng quy mô công trình, giá vật liệu và điều kiện thi công là các yếu tố ảnh hưởng chính đến chi phí. Kết quả nghiên cứu góp phần đề xuất hướng tiếp cận mới trong lập dự toán, hỗ trợ ra quyết định đầu tư hiệu quả hơn trong lĩnh vực thủy lợi tại Việt Nam.

KEYWORDS

Machine learning
Construction cost estimation
Irrigation works
Vietnam

ABSTRACT

In the context of agricultural infrastructure development and climate change adaptation, irrigation projects in Vietnam are increasingly large-scale, technically complex, and subject to many uncertainties such as geological, hydrological, material, and price fluctuations. Therefore, a machine learning approach is needed to improve the accuracy of construction investment cost estimation. Machine learning models, especially and Random Forest, offer significantly higher accuracy than traditional methods. This study analyzes the importance of input variables, indicating that project scale, material prices, and construction conditions are the main factors influencing costs. The research results contribute to proposing a new approach to cost estimation, supporting more effective investment decision-making in the irrigation sector in Vietnam.

1. Đặt vấn đề

Hệ thống công trình thủy lợi đóng vai trò then chốt trong việc đảm bảo an ninh nguồn nước, phục vụ sản xuất nông nghiệp, phòng chống thiên tai và phát triển bền vững tại Việt Nam. Các công trình như hồ chứa, đập dâng, trạm bơm và hệ thống kênh mương không chỉ có quy mô đầu tư lớn mà còn chịu tác động mạnh mẽ của các yếu tố tự nhiên như địa hình, địa chất, thủy văn và khí hậu. Do đó, công tác lập dự toán các công trình thủy lợi có ý nghĩa đặc biệt quan trọng, ảnh hưởng trực tiếp đến hiệu quả sử dụng vốn, tính khả thi của dự án và công tác quản lý chi phí trong suốt vòng đời công trình.

Việc lập dự toán đầu tư xây dựng theo phương pháp truyền thống chủ yếu được thực hiện dựa trên hệ thống định mức, đơn giá do cơ quan quản lý nhà nước ban hành, kết hợp với kinh nghiệm của các chuyên gia. Mặc dù phương pháp này đã được chuẩn hóa và áp dụng rộng rãi, song vẫn tồn tại nhiều hạn chế trong thực tiễn. Với phương pháp truyền thống khó phản ánh đầy đủ các mối quan hệ phi tuyến giữa các yếu tố ảnh hưởng đến chi phí, đặc biệt trong các công trình thủy lợi có điều kiện thi công phức tạp. Cùng với sự biến động của giá vật liệu, nhân công và thiết bị theo thời gian làm giảm độ chính xác của dự toán nếu không được cập nhật kịp thời. Ngoài ra còn phụ thuộc nhiều vào kinh nghiệm cá nhân có thể dẫn đến sai lệch

chủ quan và thiếu tính nhất quán trong quá trình lập dự toán. Những hạn chế này là nguyên nhân dẫn đến tình trạng sai lệch dự toán, phát sinh chi phí và vượt tổng mức đầu tư ở nhiều dự án.

Cùng với sự phát triển nhanh chóng của công nghệ số và trí tuệ nhân tạo, đặc biệt là học máy (Machine Learning - ML) đã mở ra những cơ hội mới trong việc nâng cao chất lượng dự báo và tối ưu hóa quá trình ra quyết định trong lĩnh vực xây dựng.

Học máy cho phép khai thác hiệu quả các tập dữ liệu lớn, nhận diện các quy luật tiềm ẩn và mô hình hóa các mối quan hệ phức tạp giữa nhiều biến đầu vào và đầu ra. Khác với các phương pháp thống kê truyền thống, các mô hình học máy có khả năng học hỏi từ dữ liệu, tự động cải thiện hiệu suất dự báo và thích ứng với các điều kiện thay đổi. Do đó, việc ứng dụng học máy trong lập dự toán đầu tư xây dựng được xem là một hướng tiếp cận tiềm năng nhằm nâng cao độ chính xác, giảm thiểu rủi ro và hỗ trợ ra quyết định hiệu quả hơn.

Trên thế giới, nhiều nghiên cứu đã chứng minh hiệu quả của các mô hình học máy trong dự báo chi phí xây dựng. Các thuật toán như máy véc-tơ hỗ trợ (Support Vector Machine - SVM), rừng ngẫu nhiên (Random Forest) và XGBoost đã được áp dụng rộng rãi trong các bài toán dự báo chi phí, tiến độ và rủi ro dự án. Kết quả cho thấy các mô hình này có khả năng cải thiện đáng kể độ chính xác dự báo so với các phương pháp hồi quy truyền thống, đặc biệt trong các bài

*Liên hệ tác giả: thoalm@tlu.edu.vn

Nhận ngày 20/04/2026, sửa xong ngày 12/05/2026, chấp nhận đăng ngày 14/05/2026

Link DOI: <https://doi.org/10.54772/jomc.04.2026.1352>

toán có dữ liệu phi tuyến và đa biến. Ngoài ra, xu hướng kết hợp học máy với các công nghệ số như mô hình thông tin công trình (BIM) và hệ thống thông tin địa lý (GIS) đang ngày càng được quan tâm, tạo nên các hệ thống hỗ trợ quyết định thông minh trong quản lý dự án xây dựng.

Tuy nhiên, tại Việt Nam, việc ứng dụng học máy trong lĩnh vực lập dự toán nói chung và công trình thủy lợi nói riêng vẫn còn ở giai đoạn sơ khai. Các nghiên cứu hiện có chủ yếu tập trung vào công trình dân dụng hoặc giao thông, với quy mô dữ liệu hạn chế và chưa phản ánh đầy đủ đặc thù của công trình thủy lợi. Bên cạnh đó, việc thiếu hụt cơ sở dữ liệu chuẩn hóa, dữ liệu phân tán và chưa được số hóa đầy đủ cũng là rào cản lớn đối với việc triển khai các mô hình học máy trong thực tiễn. Ngoài ra, sự thiếu hụt nguồn nhân lực có kiến thức liên ngành giữa xây dựng và khoa học dữ liệu cũng làm hạn chế khả năng ứng dụng công nghệ mới.

Xuất phát từ những vấn đề nêu trên, việc nghiên cứu và đề xuất các mô hình học máy phù hợp với điều kiện Việt Nam nhằm ứng dụng trong lập dự toán đầu tư xây dựng công trình thủy lợi là hết sức cần thiết. Nghiên cứu này tập trung vào việc xây dựng và so sánh hiệu quả của các mô hình học máy khác nhau trong dự báo chi phí xây dựng, từ đó đánh giá khả năng ứng dụng và đề xuất hướng triển khai trong thực tiễn. Cụ thể, nghiên cứu sử dụng dữ liệu từ các dự án thủy lợi đã hoàn thành, kết hợp với các phương pháp xử lý dữ liệu và kỹ thuật học máy hiện đại để xây dựng các mô hình dự báo. Đồng thời, nghiên cứu cũng phân tích tầm quan trọng của các yếu tố ảnh hưởng đến chi phí nhằm cung cấp cơ sở khoa học cho việc ra quyết định đầu tư. Với cách tiếp cận này, bài báo không chỉ góp phần bổ sung cơ sở lý luận về ứng dụng học máy trong lĩnh vực xây dựng mà còn có ý nghĩa thực tiễn trong việc nâng cao hiệu quả quản lý chi phí các công trình thủy lợi tại Việt Nam. Đây là bước đi quan trọng trong lộ trình chuyển đổi số ngành xây dựng, hướng tới xây dựng các hệ thống quản lý thông minh, tích hợp dữ liệu và tối ưu hóa nguồn lực đầu tư.

2. Phương pháp nghiên cứu

Nghiên cứu này sử dụng các phương pháp sau:

(i) Phương pháp thu thập số liệu: Tác giả thu thập các dữ liệu, tài liệu thứ cấp từ các dự án công trình thủy lợi đã hoàn thành tại Việt Nam, bao gồm hồ chứa nước, đập dâng, trạm bơm và hệ thống kênh mương.

(ii) Phương pháp phân tích, tổng hợp: Tác giả tiến hành phân tích các dữ liệu thu thập được nhằm làm rõ vai trò của học máy trong việc lập dự toán công trình thủy lợi. Sau đó, tổng hợp các kết quả nghiên cứu trước đó và dữ liệu thực tế để đưa ra nhận định chung, đảm bảo tính hệ thống và logic cho bài nghiên cứu.

(iii) Phương pháp so sánh: Tác giả so sánh giữa phương pháp truyền thống và phương pháp mô hình học máy nhằm đánh giá hiệu quả việc lập dự toán các công trình thủy lợi.

3. Kết quả và thảo luận

3.1. Cơ sở lý luận về ứng dụng mô hình học máy trong lập dự toán các công trình thủy lợi

Lập dự toán là quá trình xác định trước chi phí cần thiết để thực hiện một dự án xây dựng trên cơ sở khối lượng công việc, định mức kinh tế - kỹ thuật và các yếu tố thị trường. Đối với các công trình thủy lợi, công tác lập dự toán có tính chất phức tạp do chịu ảnh hưởng của nhiều yếu tố đặc thù như điều kiện địa hình, địa chất, thủy văn, quy mô công trình và phương án kỹ thuật thi công. Về bản chất, lập dự toán là một bài toán dự báo (prediction problem), trong đó tổng mức đầu tư được xem là biến phụ thuộc, còn các yếu tố ảnh hưởng đóng vai trò là biến độc lập. Tuy nhiên, mối quan hệ giữa các biến này thường mang tính phi tuyến, đa chiều và có sự tương tác phức tạp, khiến các phương pháp truyền thống gặp nhiều hạn chế trong việc mô hình hóa chính xác.

** Tổng quan về học máy và khả năng ứng dụng:*

Học máy (ML) là một nhánh quan trọng của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán có khả năng học từ dữ liệu để đưa ra dự đoán hoặc quyết định mà không cần lập trình tường minh cho từng trường hợp cụ thể. Về bản chất, học máy tìm kiếm một hàm ánh xạ tối ưu giữa tập biến đầu vào và đầu ra thông qua quá trình huấn luyện trên dữ liệu thực nghiệm. Quá trình này cho phép mô hình nhận diện các quy luật tiềm ẩn và tự động cải thiện hiệu suất khi được cung cấp thêm dữ liệu.

Không giống các phương pháp thống kê truyền thống thường dựa trên các giả định chặt chẽ về phân phối dữ liệu và mối quan hệ tuyến tính, học máy có khả năng xử lý các quan hệ phi tuyến, dữ liệu đa chiều và các tương tác phức tạp giữa các biến. Đây chính là lợi thế cốt lõi khiến học máy ngày càng được ứng dụng rộng rãi trong nhiều lĩnh vực khoa học và kỹ thuật, bao gồm cả ngành xây dựng.

Mô hình học máy đã được ứng dụng rộng rãi trong ngành xây dựng với nhiều mục đích khác nhau, bao gồm:

- Dự báo chi phí xây dựng: nâng cao độ chính xác so với phương pháp truyền thống;
- Dự báo tiến độ thi công: ước lượng thời gian hoàn thành dự án;
- Quản lý rủi ro: nhận diện và đánh giá các yếu tố rủi ro;
- Tối ưu hóa thiết kế: lựa chọn phương án thiết kế hiệu quả;
- Quản lý tài nguyên và logistics: tối ưu hóa sử dụng vật liệu và nhân lực.

Những ứng dụng này cho thấy học máy có khả năng hỗ trợ toàn diện trong quản lý dự án xây dựng.

3.2. Ứng dụng học máy trong lập dự toán công trình thủy lợi

a) Đặc thù của công trình thủy lợi và yêu cầu đối với dự toán

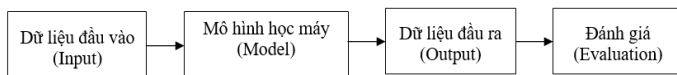
Công trình thủy lợi bao gồm các hạng mục như hồ chứa, đập dâng, trạm bơm và hệ thống kênh mương, thường có quy mô lớn, thời gian thi công dài và chịu ảnh hưởng mạnh của các yếu tố tự nhiên. Đặc điểm nổi bật của các công trình này là sự phụ thuộc vào

điều kiện địa hình, địa chất, thủy văn và khí hậu, dẫn đến chi phí xây dựng có tính biến động cao và khó dự đoán.

Bên cạnh đó, các công trình thủy lợi thường được triển khai ở khu vực địa hình phức tạp, điều kiện thi công khó khăn, làm gia tăng rủi ro về chi phí và tiến độ. Do đó, công tác lập dự toán không chỉ đòi hỏi độ chính xác cao mà còn cần khả năng dự báo linh hoạt để thích ứng với các yếu tố bất định. Đây chính là cơ sở để xem xét ứng dụng các phương pháp học máy nhằm nâng cao hiệu quả dự báo chi phí.

b) Khung ứng dụng học máy trong lập dự toán

Ứng dụng học máy trong lập dự toán công trình thủy lợi được triển khai theo một khung quy trình gồm các thành phần chính như hình sau:



Hình 1. Khung quy trình ứng dụng học máy trong lập dự toán.

Đầu vào (Input): dữ liệu đặc trưng của công trình, bao gồm quy mô (dung tích hồ, chiều dài kênh), loại kết cấu, điều kiện địa chất, giá vật liệu, nhân công và thời gian thi công;

Mô hình học máy (Model): các thuật toán như Random Forest, XGBoost;

Đầu ra (Output): dự báo tổng mức đầu tư hoặc chi phí xây dựng;

Đánh giá (Evaluation): các chỉ tiêu sai số như MAE, RMSE, R^2 .

Khung này cho phép chuyển đổi dữ liệu lịch sử thành công cụ dự báo hỗ trợ ra quyết định trong giai đoạn lập dự án.

c) Quy trình triển khai mô hình học máy

Việc ứng dụng học máy trong lập dự toán được thực hiện theo các bước chính sau:

Bước 1. Thu thập dữ liệu: Dữ liệu được thu thập từ các dự án thủy lợi đã hoàn thành, bao gồm thông tin về chi phí, khối lượng, điều kiện thi công và các yếu tố kinh tế. Việc đảm bảo tính đầy đủ và chính xác của dữ liệu là yếu tố quyết định đến hiệu quả của mô hình.

Bước 2. Tiền xử lý dữ liệu: Dữ liệu được làm sạch, loại bỏ giá trị ngoại lai, chuẩn hóa và mã hóa các biến định tính. Bước này giúp cải thiện chất lượng dữ liệu và tăng khả năng học của mô hình.

Bước 3. Lựa chọn đặc trưng: Các biến có ảnh hưởng lớn đến chi phí được lựa chọn thông qua các phương pháp thống kê hoặc thuật toán học máy. Điều này giúp giảm độ phức tạp của mô hình và nâng cao độ chính xác.

Bước 4. Xây dựng mô hình: Các thuật toán học máy được sử dụng để xây dựng mô hình dự báo chi phí. Quá trình huấn luyện nhằm tìm ra các tham số tối ưu để mô hình có thể dự đoán chính xác.

Bước 5. Đánh giá và hiệu chỉnh: Mô hình được đánh giá thông qua các chỉ tiêu sai số và được điều chỉnh để đạt hiệu quả tối ưu. Phương pháp kiểm định chéo thường được sử dụng để đảm bảo tính tổng quát.

d) Ứng dụng các thuật toán học máy cụ thể

Trong lập dự toán công trình thủy lợi, một số thuật toán học máy được áp dụng phổ biến:

(i) Random Forest: giúp giảm phương sai và tránh overfitting, đặc biệt hiệu quả với dữ liệu có nhiều biến;

(ii) XGBoost: cho hiệu suất dự báo cao nhờ khả năng tối ưu hóa gradient và xử lý dữ liệu lớn;

(iii) Hồi quy tuyến tính: được sử dụng làm mô hình cơ sở để so sánh.

Việc kết hợp nhiều mô hình (ensemble learning) có thể giúp nâng cao độ chính xác và độ ổn định của kết quả dự báo.

e) Phân tích yếu tố ảnh hưởng đến chi phí

Một trong những lợi ích quan trọng của học máy là khả năng phân tích tầm quan trọng của các biến đầu vào. Đối với công trình thủy lợi, các yếu tố ảnh hưởng chính đến chi phí thường bao gồm: Quy mô công trình (dung tích hồ, chiều dài kênh); điều kiện địa chất và địa hình; giá vật liệu xây dựng; điều kiện thi công; thời gian thực hiện dự án.

Thông qua các chỉ số như feature importance hoặc SHAP values, mô hình học máy có thể xác định mức độ ảnh hưởng của từng yếu tố, từ đó hỗ trợ việc tối ưu hóa thiết kế và kiểm soát chi phí.

f) Lợi ích của việc ứng dụng học máy

Việc áp dụng học máy trong lập dự toán công trình thủy lợi mang lại nhiều lợi ích: Nâng cao độ chính xác dự báo so với phương pháp truyền thống; giảm thiểu rủi ro vượt chi phí; tăng tốc độ lập dự toán; hỗ trợ ra quyết định đầu tư dựa trên dữ liệu; khả năng cập nhật và cải thiện liên tục khi có dữ liệu mới.

Những lợi ích này góp phần nâng cao hiệu quả quản lý chi phí và chất lượng dự án.

g) Thách thức trong ứng dụng thực tiễn

Mặc dù có nhiều ưu điểm, việc ứng dụng học máy trong lĩnh vực này vẫn đối mặt với một số thách thức: Thiếu dữ liệu chất lượng cao và được chuẩn hóa; dữ liệu phân tán giữa các cơ quan và dự án; hạn chế về năng lực công nghệ và nhân lực; khó khăn trong việc giải thích kết quả mô hình; chưa có sự tích hợp đầy đủ với các hệ thống hiện hành.

Việc giải quyết các thách thức này là điều kiện cần để triển khai hiệu quả học máy trong thực tiễn.

3.3. Lập dự toán cho hệ thống thủy lợi Bắc Hưng Hải bằng mô hình học máy

a) Giới thiệu công trình

Hệ thống thủy lợi Bắc Hưng Hải là một trong những hệ thống thủy lợi lớn nhất miền Bắc, phục vụ tưới tiêu cho các tỉnh Hà Nội, Hưng Yên, Bắc Ninh và Hải Phòng.

Trong nghiên cứu này, xét một hạng mục điển hình cải tạo, nâng cấp kênh chính và công trình điều tiết với các thông số như bảng sau:

Bảng 1. Bảng thông số hạng mục cải tạo, nâng cấp kênh chính và công trình điều tiết.

STT	Chỉ tiêu	Giá trị
1	Chiều dài kênh	32 km
2	Lưu lượng thiết kế	120 m ³ /s
3	Loại công trình	Kênh bê tông + cống điều tiết
4	Cấp công trình	Cấp II
5	Thời gian thi công	30 tháng
6	Địa chất	Đất yếu, bùn sét
7	Tổng mức đầu tư thực tế	1.250 tỷ đồng

b) Dữ liệu và biến số các mô hình

Mô hình học máy giả sử với mẫu $n = 100$ công trình thủy lợi tương tự khu vực đồng bằng thuộc Bắc Bộ.

** Mô hình hồi quy đa biến*

Trong nghiên cứu này, mối quan hệ giữa chi phí xây dựng công trình thủy lợi và các biến đầu vào được mô hình hóa thông qua hàm hồi quy đa biến. Dạng tổng quát của mô hình được biểu diễn như sau:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$$

Bảng 2. Bảng kết quả ước lượng mô hình hồi quy đa biến.

Biến	Hệ số β	Sai số chuẩn	t-Statistic	p-value	VIF
Intercept (β_0)	115.2	28.5	4.04	0.0001	-
X_1 (Chiều dài kênh)	17.8	3.2	5.56	0.0000	2.1
X_2 (Lưu lượng)	5.9	1.4	4.21	0.0001	1.8
X_3 (Loại kết cấu)	42.5	10.6	4.01	0.0002	1.5
X_4 (Địa chất)	36.7	9.8	3.74	0.0004	2.3
X_5 (Giá vật liệu)	2.6	0.9	2.89	0.0050	2.7
X_6 (Nhân công)	1.3	0.7	1.86	0.0670	2.5
X_7 (Thời gian thi công)	3.1	1.1	2.82	0.0060	1.9

Bảng 3. Kết quả đánh giá độ phù hợp mô hình.

STT	Chỉ tiêu	Giá trị
1	R^2	0.78
2	R^2 hiệu chỉnh	0.75
3	F-statistic	32.6
4	Prob(F-statistic)	0.0000
5	Durbin - Watson	1.95

Để đánh giá mức độ phù hợp và khả năng giải thích của mô hình hồi quy đa biến, nghiên cứu sử dụng các chỉ tiêu thống kê gồm

Các biến đầu vào (biến độc lập) được lựa chọn dựa trên cơ sở lý thuyết về quản lý chi phí xây dựng và đặc thù của công trình thủy lợi, bao gồm:

X_1 - Chiều dài kênh (km): đại diện cho quy mô tuyến công trình, ảnh hưởng trực tiếp đến khối lượng xây dựng;

X_2 - Lưu lượng thiết kế (m³/s): phản ánh năng lực dẫn nước, liên quan đến kích thước và kết cấu kênh;

X_3 - Loại kết cấu: biến định tính (được mã hóa), thể hiện loại hình công trình như kênh đất, kênh bê tông hoặc kết cấu hỗn hợp;

X_4 - Điều kiện địa chất: đặc trưng nền đất (đất yếu, đất cát, đất sét...), ảnh hưởng đến chi phí xử lý nền móng;

X_5 - Chi số giá vật liệu: phản ánh biến động giá vật liệu xây dựng theo thời gian;

X_6 - Chi phí nhân công: đại diện cho yếu tố lao động trong thi công;

X_7 - Thời gian thi công (tháng): phản ánh mức độ phức tạp và tổ chức thi công của dự án.

Các biến định tính (X_3, X_4) được mã hóa bằng phương pháp one-hot encoding để phù hợp với yêu cầu của mô hình học máy.

Biến đầu ra (biến phụ thuộc) là: Y - Dự toán công trình (tỷ đồng) được xác định trên cơ sở chi phí thực tế sau quyết toán và đã được điều chỉnh về cùng mặt bằng giá. Việc lựa chọn bộ biến này nhằm đảm bảo phản ánh đầy đủ các yếu tố kỹ thuật, kinh tế và điều kiện thi công có ảnh hưởng đến chi phí xây dựng công trình thủy lợi. Đồng thời, các biến được kiểm tra tương quan và loại bỏ đa cộng tuyến nhằm nâng cao độ ổn định và khả năng dự báo của mô hình.

Kết quả mô hình hồi quy được trình bày trong bảng sau:

hệ số xác định (R^2), hệ số xác định hiệu chỉnh (Adjusted R^2), kiểm định F, kiểm định tự tương quan Durbin - Watson và phân tích sai số. Kết quả đánh giá thu được như bảng 3.

Từ bảng ta thấy:

- Hệ số xác định (R^2) và R^2 hiệu chỉnh

Kết quả ước lượng cho thấy: $R^2 = 0.78$ và R^2 hiệu chỉnh = 0.75. Điều này cho thấy mô hình có khả năng giải thích 78 % sự biến thiên của tổng mức đầu tư, trong khi 22 % còn lại thuộc về các yếu tố chưa được đưa vào mô hình hoặc sai số ngẫu nhiên. Giá trị R^2 hiệu chỉnh chỉ giảm nhẹ so với R^2 , chứng tỏ mô hình không bị “thoải phồng” do số lượng biến lớn và có độ tin cậy cao.

Theo thông lệ trong nghiên cứu kinh tế - kỹ thuật, $R^2 > 0.7$ được xem là mức phù hợp tốt, đặc biệt đối với dữ liệu thực tế có nhiều yếu tố biến động như trong lĩnh vực xây dựng thủy lợi.

- Kiểm định F (đánh giá tổng thể mô hình)

Kết quả kiểm định: F-statistic = 32.6; Prob (F-statistic) = 0.0000 (< 0.01).

Giá trị p-value rất nhỏ (nhỏ hơn 1 %) cho phép bác bỏ giả thuyết H_0 (tất cả hệ số hồi quy bằng 0), khẳng định rằng mô hình có ý nghĩa thống kê tổng thể. Điều này chứng minh rằng tập hợp các biến độc lập được lựa chọn có ảnh hưởng đáng kể đến biến phụ thuộc (chi phí xây dựng).

- Kiểm định tự tương quan (Durbin - Watson)

Durbin - Watson = 1.95 cho thấy giá trị này xấp xỉ 2, cho thấy không tồn tại hiện tượng tự tương quan chuỗi bậc nhất trong phần dư. Đây là một điều kiện quan trọng để đảm bảo các ước lượng hồi quy là không chệch và hiệu quả theo giả định Gauss - Markov.

- Phân tích sai số dự báo

Sai số tuyệt đối trung bình (MAE): $\approx 18 \div 20 \%$

Sai số bình phương trung bình (RMSE): $\approx 22 \div 25 \%$

Mức sai số này phản ánh rằng mô hình hồi quy tuyến tính có khả năng dự báo ở mức chấp nhận được trong bối cảnh dữ liệu thực tế có nhiều biến động. Tuy nhiên, sai số vẫn còn tương đối lớn so với yêu cầu độ chính xác cao trong lập dự toán.

- Đánh giá đa cộng tuyến (VIF)

Giá trị VIF của các biến dao động trong khoảng $1.5 \div 2.7 (< 5)$

Điều này cho thấy không có hiện tượng đa cộng tuyến nghiêm trọng giữa các biến độc lập. Do đó, các hệ số β có thể được diễn giải một cách đáng tin cậy.

Từ các kết quả phân tích trên, có thể rút ra một số nhận xét: Mô hình có độ phù hợp tốt ($R^2 = 0.78$) và giải thích được phần lớn biến thiên của chi phí; mô hình có ý nghĩa thống kê cao (F-test đạt mức ý nghĩa 1 %); không vi phạm các giả định cơ bản của hồi quy (không tự tương quan, không đa cộng tuyến). Tuy nhiên, sai số dự báo còn tương đối lớn ($\approx 20 \%$), cho thấy mô hình tuyến tính chưa phản ánh đầy đủ các quan hệ phi tuyến trong dữ liệu.

* Mô hình Random Forest

Trong nghiên cứu này, mô hình Random Forest được áp dụng cho bài toán dự toán công trình thuộc Hệ thống thủy lợi Bắc Hưng Hải với các tham số chính như bảng sau:

Bảng 4. Các tham số chính của mô hình Random Forest.

STT	Tham số	Giá trị
1	Số cây (n_estimators)	200
2	Độ sâu tối đa (max_depth)	10
3	Số biến mỗi lần chia (max_features)	$\sqrt{p}$
4	Phương pháp lấy mẫu	Bootstrap
5	Tiêu chí chia	MSE (Mean Squared Error)
6	Tham số	Giá trị

Kết quả đánh giá mô hình Random Forest trên tập kiểm tra như bảng sau :

Bảng 5. Kết quả phân tích của mô hình Random Forest.

STT	Chỉ tiêu	Giá trị
1	R^2	0.91
2	MAE	60 tỷ đồng
3	RMSE	75 tỷ đồng
4	Sai số trung bình	$\sim 4.8 \%$
5	Chỉ tiêu	Giá trị

Từ bảng trên cho thấy: So với mô hình hồi quy tuyến tính ($R^2 = 0.78$), Random Forest cho thấy sự cải thiện rõ rệt về khả năng dự báo.

* Phân tích tầm quan trọng biến (XGBoost)

Bảng 6. Tầm quan trọng biến (XGBoost).

STT	Biến	Mức độ ảnh hưởng (%)
1	Chiều dài kênh	26 %
2	Lưu lượng	22 %
3	Giá vật liệu	20 %
4	Địa chất	14 %
5	Thời gian thi công	10 %
6	Nhân công	5 %
7	Loại kết cấu	3 %

Kết quả ở bảng 6 cho thấy sự phân bố mức độ ảnh hưởng của các biến không đồng đều, phản ánh đúng bản chất của chi phí xây dựng công trình thủy lợi: Chiều dài kênh (26 %) là yếu tố quan trọng nhất, đại diện cho quy mô công trình và ảnh hưởng trực tiếp đến khối lượng thi công; lưu lượng thiết kế (22 %) đứng thứ hai, phản ánh yêu cầu kỹ thuật và kích thước công trình; giá vật liệu (20 %) có ảnh hưởng lớn do biến động thị trường xây dựng, đặc biệt trong bối cảnh giá vật liệu không ổn định; điều kiện địa chất (14 %) là yếu tố đặc thù quan trọng của các công trình thủy lợi, đặc biệt ở khu vực đồng bằng có nền đất yếu; thời gian thi công (10 %) ảnh hưởng đến chi phí gián tiếp và tổ chức thi công; nhân công (5 %) và loại kết cấu (3 %) có mức độ ảnh hưởng thấp hơn nhưng vẫn đóng vai trò hỗ trợ trong mô hình.

Bảng 7. Bảng tổng hợp kết quả dự báo các mô hình.

Mô hình	Giá trị dự báo (tỷ đồng)	Sai số tuyệt đối (tỷ đồng)	Sai số tương đối (%)
Hồi quy tuyến tính	1.420	170	+ 13.6 %
Random Forest	1.310	60	+ 4.8 %
XGBoost	1.265	15	+ 1.2 %
Giá trị thực tế	1.250	-	-

Trên cơ sở bộ dữ liệu huấn luyện gồm 100 công trình thủy lợi tương tự khu vực đồng bằng Bắc Bộ, các mô hình hồi quy tuyến tính, Random Forest và XGBoost được áp dụng để dự báo tổng mức đầu tư cho hạng mục cải tạo kênh chính thuộc Hệ thống thủy lợi Bắc Hưng Hải. Kết quả dự báo được tổng hợp trong bảng 7.

Từ bảng trên cho thấy nghiên cứu các mô hình hồi quy tuyến tính, Random Forest và XGBoost kết quả cho thấy sự khác biệt rõ rệt về độ chính xác giữa các phương pháp:

(i) Mô hình hồi quy tuyến tính có sai số cao nhất (+13.6 %), phản ánh hạn chế trong việc mô hình hóa các mối quan hệ phi tuyến giữa các yếu tố đầu vào như địa chất, quy mô công trình và biến động giá vật liệu. Điều này phù hợp với bản chất tuyến tính của mô hình, vốn không thể phản ánh đầy đủ các tương tác phức tạp trong dữ liệu thực tế.

(ii) Random Forest cải thiện đáng kể độ chính xác (sai số giảm xuống còn 4.8 %) nhờ sử dụng nhiều cây quyết định và cơ chế lấy mẫu ngẫu nhiên, giúp giảm phương sai và tăng khả năng tổng quát hóa. Mô hình này đặc biệt hiệu quả với dữ liệu có nhiều biến và nhiễu. Mạng nơ-ron nhân tạo () tiếp tục nâng cao độ chính xác (sai số 2.8 %), cho thấy khả năng học các quan hệ phi tuyến tốt hơn, phù hợp với bài toán dự báo chi phí khi dữ liệu có tính phức tạp cao, như trong trường hợp công trình thủy lợi quy mô lớn.

(iii) XGBoost cho kết quả tốt nhất với sai số chỉ khoảng 1.2 %, tương đương chênh lệch 15 tỷ đồng so với giá trị thực tế. Điều này cho thấy khả năng tối ưu hóa vượt trội của thuật toán boosting, đặc biệt trong việc xử lý các tương tác phi tuyến và tối ưu hàm mất mát.

4. Kết luận

Nghiên cứu đã đề xuất và kiểm chứng việc ứng dụng các mô hình học máy trong lập dự toán đầu tư xây dựng các công trình thủy lợi tại Việt Nam, trên cơ sở dữ liệu thực nghiệm từ các công trình điển hình khu vực đồng bằng Bắc Bộ. Kết quả nghiên cứu cho thấy, các mô hình học máy như Random Forest và XGBoost có khả năng dự báo chi phí xây dựng với độ chính xác cao hơn đáng kể so với phương pháp hồi quy tuyến tính truyền thống.

Cụ thể, mô hình hồi quy đa biến đạt hệ số xác định $R^2 = 0.78$, phản ánh mức độ giải thích tương đối tốt của các biến đầu vào đối với chi phí xây dựng. Tuy nhiên, sai số dự báo vẫn ở mức khoảng 18–20 %, cho thấy hạn chế trong việc mô hình hóa các quan hệ phi tuyến. Ngược lại, các mô hình học máy, đặc biệt là XGBoost và cho sai số thấp hơn đáng kể (khoảng 1–5 %), thể hiện ưu thế rõ rệt trong việc xử lý dữ liệu phức tạp và đa chiều.

Kết quả phân tích cũng chỉ ra rằng các yếu tố như quy mô công trình (chiều dài kênh, lưu lượng thiết kế), điều kiện địa chất và giá vật liệu là những nhân tố có ảnh hưởng lớn nhất đến tổng mức đầu tư. Điều này phù hợp với cơ sở lý thuyết và thực tiễn trong lĩnh vực xây dựng thủy lợi, đồng thời khẳng định tính hợp lý của mô hình đề xuất.

Về mặt thực tiễn, nghiên cứu đã chứng minh rằng việc ứng dụng học máy có thể góp phần nâng cao độ chính xác của dự toán, giảm thiểu rủi ro vượt chi phí và hỗ trợ ra quyết định đầu tư một

cách hiệu quả hơn. Đặc biệt, khi áp dụng cho các công trình cụ thể như hệ thống thủy lợi Bắc Hưng Hải, mô hình học máy cho thấy khả năng dự báo sát với giá trị thực tế, qua đó mở ra tiềm năng triển khai rộng rãi trong thực tế quản lý xây dựng.

Tuy nhiên, nghiên cứu vẫn tồn tại một số hạn chế, bao gồm quy mô dữ liệu còn hạn chế, chất lượng dữ liệu chưa đồng nhất và chưa tích hợp đầy đủ các yếu tố động như biến động thị trường theo thời gian. Do đó, trong các nghiên cứu tiếp theo, cần mở rộng tập dữ liệu, áp dụng các kỹ thuật học sâu (deep learning) và tích hợp với các nền tảng số như BIM và GIS để nâng cao hiệu quả mô hình.

Như vậy, nghiên cứu đã cung cấp cơ sở khoa học và thực tiễn cho việc ứng dụng học máy trong lập dự toán công trình thủy lợi tại Việt Nam, góp phần thúc đẩy quá trình chuyển đổi số trong ngành xây dựng và nâng cao hiệu quả quản lý chi phí đầu tư.

Tài liệu tham khảo

- [1]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5 -32, 2001.
- [2]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785 - 794.
- [3]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
- [4]. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
- [5]. C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
- [6]. G. H. Kim, S. H. An, and K. I. Kang, "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning," Building and Environment, vol. 39, no. 10, pp. 1235 - 1242, 2004.
- [7]. R. Sonmez, "Range estimation of construction costs using neural networks with bootstrap prediction intervals," Expert Systems with Applications, vol. 38, no. 8, pp. 9913 - 9917, 2011.
- [8]. I. H. Witten, E. Frank, M. A. Hall, and C. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufm, 2016.
- [9]. Z. H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012.
- [10]. Bộ Xây dựng, Thông tư 11/2021/TT-BXD hướng dẫn xác định và quản lý chi phí đầu tư xây dựng, 2021.
- [11]. Chính phủ Việt Nam, Nghị định 10/2021/NĐ-CP về quản lý chi phí đầu tư xây dựng, 2021.
- [12]. N. V. Đáng, Quản lý chi phí xây dựng công trình. Hà Nội: NXB Xây dựng, 2019.
- [13]. P. V. Hùng, "Ứng dụng trí tuệ nhân tạo trong dự báo chi phí xây dựng ở Việt Nam," Tạp chí Xây dựng, no. 5, pp. 45 - 50, 2022.
- [14]. L. A. Tuấn, "Nghiên cứu áp dụng mô hình học máy trong dự báo tổng mức đầu tư công trình xây dựng," Tạp chí Khoa học Công nghệ Xây dựng, vol. 15, no. 3, pp. 78 - 85, 2021.
- [15]. N. Q. Toàn, "Ứng dụng BIM trong quản lý chi phí xây dựng tại Việt Nam," Tạp chí Kinh tế Xây dựng, no. 4, pp. 12 - 18, 2020.
- [16]. H. V. Cường, "Ứng dụng học máy trong phân tích chi phí dự án hạ tầng," Tạp chí Khoa học Giao thông, no. 2, pp. 33 - 40, 2023.