ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT

BÁO CÁO TỔNG KẾT

ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP TRƯỜNG

DỰ BÁO CHI PHÍ XÂY DỰNG NHÀ Ở CAO TẦNG BẰNG MÔ HÌNH TÍCH HỢP DỰA TRÊN HỌC MÁY

Mã số: T2024-06-13

Chủ nhiệm đề tài: **ThS. Trương Thị Thu Hà**
Đơn vị: **Khoa Kỹ thuật Xây dựng**
Chương trình đào tạo: Công nghệ Kỹ thuật Xây dựng

Đà Nẵng, tháng 12/2025

ĐẠI HỌC ĐÀ NẴNG

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT

BÁO CÁO TỔNG KẾT

ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP TRƯỜNG

DỰ BÁO CHI PHÍ XÂY DỰNG NHÀ Ở CAO TẦNG BẰNG MÔ HÌNH TÍCH HỢP DỰA TRÊN HỌC MÁY

Mã số: T2024-06-13

Xác nhận của cơ quan chủ trì đề tài	Chủ nhiệm đề tài
KT. HIỆU TRƯỞNG	
PHÓ HIỆU TRƯỞNG	
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT	
PGS. TS. Võ Trung Hùng	ThS. Trương Thị Thu Hà

DANH SÁCH THÀNH VIÊN THAM GIA NGHIÊN CỨU ĐỀ TÀI

TT	Họ và tên	Đơn vị công tác và lĩnh vực chuyên môn
1	ThS. Trương Thị Thu Hà	Khoa Kỹ thuật Xây dựng - Trường Đại học Sư phạm Kỹ thuật, Đại học Đà Nẵng; chuyên ngành Quản lý xây dựng
2	TS. Ngô Ngọc Tri	Khoa Quản lý dự án - Trường Đại học Bách Khoa, Đại học Đà Nẵng; chuyên ngành chuyên ngành Quản lý xây dựng
3	ThS. Phạm Thị Phương Trang	Khoa Kỹ thuật Xây dựng - Trường Đại học Sư phạm Kỹ thuật, Đại học Đà Nẵng; chuyên ngành Quản lý xây dựng
4	ThS. Lê Thị Thùy Linh	Khoa Sư phạm Công nghiệp - Trường Đại học Sư phạm Kỹ thuật, Đại học Đà Nẵng; chuyên ngành Quản lý xây dựng

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

DANH MỤC CÁC TỪ VIẾT TẮT

AI	:	Trí tuệ nhân tạo
ANNs	:	Mạng nơ-ron nhân tạo
BIM	:	Mô hình thông tin công trình
CART	:	Cây phân loại và hồi quy
DT	:	Cây quyết định
HVAC	:	Hệ thống thống sưởi, thông gió và điều hòa không khí
KNN	:	K-Nearest Neighbor
LOO-FI	:	Leave-One-Out-Feature-Importance
LR	:	Hồi quy tuyến tính
MAE	:	Sai số trung bình tuyệt đối
MAPE	:	Sai số trung bình tuyệt đối phần trăm
MSE	:	Sai số bình phương trung bình
RMSE	:	Căn bậc hai của sai số bình phương trung bình
RF	:	Rừng ngẫu nhiên
R^2	:	Hệ số xác định
SEM	:	Phương pháp bao che tầng
SHAP	:	SHapley Additive exPlanations
SI	:	Chỉ số xếp hạng tổng hợp
SVR	:	Máy học véc-tơ hỗ trợ

ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐH SƯ PHẠM KỸ THUẬT

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

THÔNG TIN KẾT QUẢ NGHIÊN CỨU

1. Thông tin chung

- Mã số : T2024-06-13
- Tên đề tài : Dự báo chi phí xây dựng nhà ở cao tầng bằng mô hình tích hợp dựa trên học máy.
- Chủ nhiệm : ThS. Trương Thị Thu Hà
- Thành viên tham gia : Ngô Ngọc Tri, Phạm Thị Phương Trang, Lê Thị Thùy Linh
- Cơ quan chủ trì : Trường Đại học Sư phạm Kỹ thuật
- Thời gian thực hiện : 12 tháng (01/2025 – 12/2025)

2. Mục tiêu

- Tổng hợp các phương pháp ước lượng chi phí xây dựng công trình;
- Dự báo chi phí xây dựng nhà ở cao tầng trên địa bàn thành phố Hồ Chí Minh bằng các phương pháp học máy;
- Đề xuất mô hình tích hợp dựa trên học máy để dự đoán sơ bộ chi phí xây dựng nhà ở cao tầng.

3. Tính mới và sáng tạo

- Tích hợp phương pháp học máy tiên tiến trong ước tính chi phí xây dựng ở giai đoạn thiết kế sớm;
- Xây dựng khung so sánh toàn diện giữa mô hình đơn và mô hình tổ hợp.

4. Tóm tắt kết quả nghiên cứu

Nghiên cứu đã tiến hành so sánh và đánh giá hiệu quả dự báo chi phí xây dựng của các mô hình học máy đơn (ANNs, SVR, CART) và các mô hình học máy tích hợp (Voting, Bagging, Stacking). Đối với nhóm mô hình đơn lẻ, SVR chứng minh là phương pháp có hiệu suất dự báo vượt trội nhất, cho sai số dự báo thấp nhất trong ba mô hình. Ở nhóm mô hình tích hợp, Voting cho thấy kết quả dự báo khả quan nhất, đặc biệt với tổ hợp ANNs+SVR, và được xếp hạng cao nhất trong toàn bộ các mô hình.

5. Tên sản phẩm

5.1. Sản phẩm khoa học

Các sản phẩm khoa học đáp ứng theo yêu cầu theo đăng ký trong thuyết minh

[1] Truong, T.T.H. and N.T. Ngo. Predicting High-Rise Buildings Construction Cost Using Machine Learning-Based Ensemble Model. in 2025 10th International Conference on Applying New Technology in Green Buildings (ATiGB). 2025. Danang, Vietnam: IEEE.

Link bài báo: https://doi.org/10.1109/ATiGB66719.2025.11142151

6. Hiệu quả, phương thức chuyển giao kết quả nghiên cứu và khả năng áp dụng

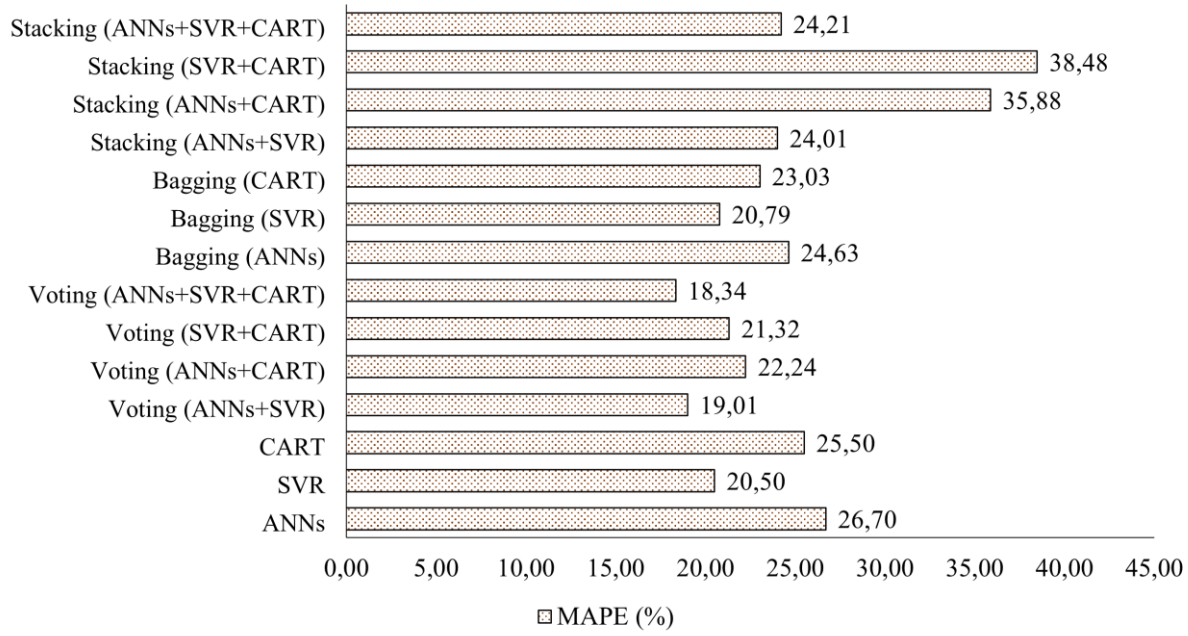
* Đối với tổ chức chủ trì và các cơ sở ứng dụng kết quả nghiên cứu:

- Các kết quả nghiên cứu của đề tài là tài liệu khoa học có giá trị phục vụ cho công tác nghiên cứu, đào tạo trong lĩnh vực xây dựng, …
- Đề tài góp phần nâng cao năng lực nghiên cứu của các cán bộ, là tài liệu tham khảo bổ ích cho giảng viên và sinh viên.

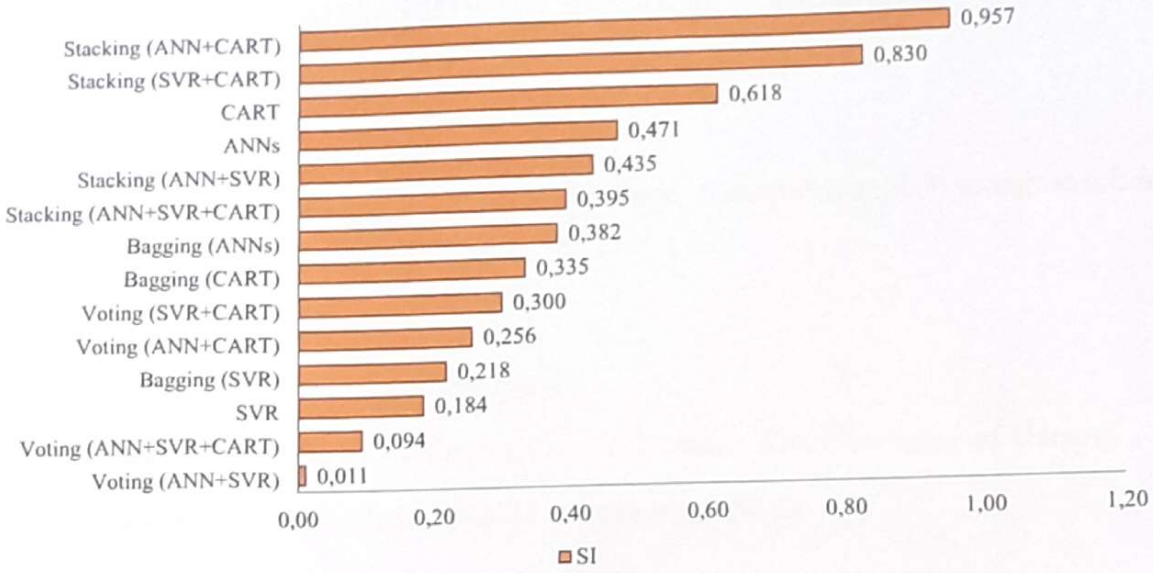
* Đối với kinh tế - xã hội và môi trường:

- Cung cấp cho chủ đầu tư và doanh nghiệp xây dựng một công cụ hỗ trợ thông tin khi đưa ra các quyết định liên quan đến tài chính;
- Giúp các bên liên quan phân bổ nguồn lực tài chính một cách hợp lý, đảm bảo tiến độ dự án;
- Giúp các bên liên quan tiết kiệm thời gian và chi phí trong quá trình chuẩn bị và triển khai dự án.

7. Hình ảnh, sơ đồ minh họa chính



Hình 1. *Giá trị MAPE của tất cả mô hình.*



Hình 2. *Tổng hợp xếp hạng của các mô hình.*

Ngày 20 tháng 01 năm 2026

TM. Hội đồng Khoa
Chủ tịch

Phan Tiến Vinh

Chủ nhiệm đề tài

ThS. Trương Thị Thu Hà

XÁC NHẬN CỦA TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT
KT. HIỆU TRƯỞNG
PHÓ HIỆU TRƯỞNG

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT
ĐẠI HỌC ĐÀ NẴNG

PGS. TS. Võ Trung Hùng

INFORMATION ON RESEARCH RESULTS

1. General information:

Project title: Predicting high-rise buildings construction cost using machine learning-based ensemble model

Code: T2024-06-13

Coordinator: MSc. Trương Thị Thu Hà

Sponsor: University of Technology and Education, The University of Danang

Duration: 12 months (January 2025 – December 2025).

2. Objectives:

- To review and synthesize existing cost estimation methods for construction projects;
- To predict the construction costs of high-rise residential buildings in Ho Chi Minh city using machine learning techniques;
- To propose an integrated machine learning model for preliminary estimation of construction costs of high-rise residential buildings.

3. Creativeness and innovativeness:

- The study integrates advanced machine learning algorithms into early-stage cost estimation;
- It develops a comprehensive comparative framework between single and hybrid models to enhance reliability and predictive accuracy.

4. Research results:

The research was conducted to compare and evaluate the performance of various machine learning approaches for predicting construction costs, including single models (ANNs, SVR, CART) and ensemble models (Voting, Bagging, Stacking). Among the single models, the SVR was demonstrated to be the highest predictive accuracy with the lowest error metrics. For ensemble models, the Voting produced stable and reliable results, especially the ensembler ANNs + SVR achieved the best performance among all tested models.

5. Products:

5.1. Scientific products

Scientific products meet the requirements registered in the research description.

[1] Truong, T.T.H. and N.T. Ngo. Predicting High-Rise Buildings Construction Cost Using Machine Learning-Based Ensemble Model. in 2025 10th International Conference on Applying New Technology in Green Buildings (ATiGB). 2025. Danang, Vietnam: IEEE.

Link bài báo: https://doi.org/10.1109/ATiGB66719.2025.11142151

6. Effects, transfer alternatives of reserach results and applicability:

* For the host institution and academic users:

- The research outputs serve as valuable academic references supporting teaching and research activities in the construction and cost estimation domains;
- The project contributes to improving the research capacity of lecturers and staff, and provides useful reference materials for both faculty members and students.

* For economic, social, and environmental sectors:

- Provides investors and construction enterprises with a practical decision-support tool for financial planning and project management;
- Assists stakeholders in optimizing financial resource allocation to ensure project efficiency and sustainability;
- Helps stakeholders save time and costs during project planning and implementation phases.

MỞ ĐẦU

1. LÝ DO CHỌN ĐỀ TÀI

Chi phí xây dựng là yếu tố quan trọng ảnh hưởng đến khả năng trúng thầu của các nhà thầu cũng như kế hoạch tài chính của nhà đầu tư. Việc ước lượng chính xác giúp giảm thiểu rủi ro tài chính, tránh thất thoát và gia tăng hiệu quả quản lý dự án. Trong giai đoạn hình thành dự án, ước tính chi phí xây dựng công trình đóng vai trò quan trọng. Nó giúp người quyết định đầu tư và chủ đầu tư bố trí vốn cho dự án, xác định giá gói thầu, từ đó lập kế hoạch lựa chọn nhà thầu và quản lý chi phí dự án. Đối với nhà thầu xây lắp, ước lượng chính xác chi phí xây dựng hỗ trợ lập chiến lược tranh thầu, tạo điều kiện để quản lý hiệu quả tiến độ và chi phí dự án.

Kinh nghiệm của người lập dự toán và thông tin dự án là các nhân tố thiết yếu để ước lượng chi phí bởi vì các thông tin về phạm vi dự án trong giai đoạn đầu còn hạn chế. Các phương pháp ước lượng chi phí truyền thống được sử dụng như phương pháp đơn vị, phương pháp thể tích, phương pháp diện tích sàn, phương pháp phân tích phần tử, hay phương pháp bóc tách khối lượng... Các phương pháp này tính toán thủ công, chủ yếu dựa vào kinh nghiệm nên tốn nhiều thời gian, công sức (đặc biệt là công trình có quy mô lớn).

Trong những thập niên gần đây, trí tuệ nhân tạo (AI) được ứng dụng để giải quyết các vấn đề trong lĩnh vực xây dựng nói riêng và các ngành khác nói chung. Ưu điểm của các mô hình học máy dựa trên trí tuệ nhân tạo là chúng không cần giả định trước các đặc tính của dữ liệu như các mô hình truyền thống. Một số mô hình học máy phổ biến như mạng nơ-ron nhân tạo (ANNs), máy học véc-tơ hỗ trợ hồi quy (SVR), cây quyết định (DT), rừng ngẫu nhiên (RF)…Mỗi mô hình học máy có lợi thế và hạn chế riêng, do đó việc kết hợp các mô hình đơn lẻ thành mô hình tích hợp nhằm tận dụng lợi thế và khắc phục hạn chế của từng mô hình, hứa hẹn cải thiện độ chính xác dự báo.

Việc ứng dụng các mô hình học máy không chỉ giúp nâng cao độ tin cậy trong dự đoán chi phí mà còn hỗ trợ ra quyết định ở giai đoạn sớm của dự án – khi dữ liệu đầu vào còn hạn chế. Ngoài ra, các mô hình này cho phép nhận diện được mức độ ảnh hưởng của từng yếu tố kỹ thuật (như diện tích sàn, loại móng, phương pháp thi công, thời gian thi công,…) đến chi phí tổng thể, giúp nhà đầu tư và nhà thầu có

chiến lược tối ưu hóa nguồn lực hiệu quả hơn. Hơn nữa, việc áp dụng học máy còn góp phần tự động hóa quy trình lập dự toán, giảm phụ thuộc vào kinh nghiệm cá nhân và tăng khả năng tiêu chuẩn hóa trong công tác quản lý chi phí.

Vì vậy, nghiên cứu dự báo chi phí xây dựng bằng mô hình tích hợp dựa trên học máy trên cơ sở dữ liệu nhà ở cao tầng thu thập từ Việt Nam có ý nghĩa khoa học và cấp thiết, là công cụ hỗ trợ nhà đầu tư và nhà thầu xác định sơ bộ chi phí đầu tư và chi phí xây dựng công trình.

2. MỤC TIÊU NGHIÊN CỨU

Mục tiêu tổng thể: Đề xuất mô hình học máy để dự báo chi phí thi công xây dựng nhà ở cao tầng trong giai đoạn hình thành dự án.

Mục tiêu cụ thể:

- Tổng hợp các phương pháp ước lượng chi phí xây dựng công trình;
- Dự báo chi phí xây dựng nhà ở cao tầng trên địa bàn thành phố Hồ Chí Minh bằng các phương pháp học máy;
- Đề xuất mô hình tích hợp dựa trên học máy để dự đoán sơ bộ chi phí xây dựng nhà ở cao tầng.

3. PHẠM VI NGHIÊN CỨU

Nhà ở cao tầng khu vực nội thành TP. Hồ Chí Minh.

4. Ý NGHĨA KHOA HỌC VÀ THỰC TIỄN CỦA ĐỀ TÀI

- Các kết quả nghiên cứu của đề tài là tài liệu khoa học có giá trị phục vụ cho công tác nghiên cứu, đào tạo trong lĩnh vực xây dựng;
- Đề tài góp phần nâng cao năng lực nghiên cứu của các cán bộ, là tài liệu tham khảo bổ ích cho giảng viên và sinh viên;
- Cung cấp cho chủ đầu tư và doanh nghiệp xây dựng một công cụ hỗ trợ thông tin khi đưa ra các quyết định liên quan đến tài chính;
- Giúp các bên liên quan phân bổ nguồn lực tài chính một cách hợp lý, đảm bảo tiến độ dự án;
- Giúp các bên liên quan tiết kiệm thời gian và chi phí trong quá trình chuẩn bị và triển khai dự án.

5. BỐ CỤC ĐỀ TÀI

Đề tài gồm những nội dung chính như sau:

MỞ ĐẦU

CHƯƠNG 1: TỔNG QUAN VỀ DỰ BÁO CHI PHÍ XÂY DỰNG

CHƯƠNG 2: CÁC PHƯƠNG PHÁP DỰ BÁO CHI PHÍ XÂY DỰNG

CHƯƠNG 3: DỰ BÁO CHI PHÍ XÂY DỰNG NHÀ Ở CAO TẦNG BẰNG MÔ HÌNH TÍCH HỢP DỰA TRÊN HỌC MÁY

CHƯƠNG 4: KẾT LUẬN VÀ KIẾN NGHỊ

TÀI LIỆU THAM KHẢO

CHƯƠNG 1: TỔNG QUAN VỀ DỰ BÁO CHI PHÍ XÂY DỰNG

1.1. VAI TRÒ CỦA DỰ BÁO CHI PHÍ XÂY DỰNG

Trong quá trình quản lý dự án, các công ty xây dựng đối mặt với những thách thức như vượt ngân sách, chi phí tăng do biến động giá vật liệu và áp lực ngày càng tăng để đáp ứng các yêu cầu về tính bền vững [15]. Các nhà quản lý và nhà ra quyết định thường bối rối và khó đưa ra các đánh giá chi phí nhanh chóng và đáng tin cậy ở giai đoạn thiết kế ban đầu, đặc biệt khi dữ liệu dự án chi tiết không đủ. Trong giai đoạn đầu của một dự án xây dựng thường thiếu bản vẽ xây dựng toàn diện, vì vậy việc hoàn thành dự đoán chi phí với thông tin hạn chế là một thách thức [16].

Dự đoán chi phí cuối cùng của các dự án xây dựng ở giai đoạn đầu rất quan trọng để bàn giao dự án thành công [1]. Kết quả dự đoán chính xác cho phép chủ đầu tư xác định ngân sách dự án hợp lý trước khi bắt đầu và áp dụng các chiến lược kiểm soát chi phí phù hợp trong quá trình thực hiện dự án để tránh phát sinh tài chính [2]. Bên cạnh đó, dự đoán chi phí của các dự án xây dựng là rất quan trọng đối với quản lý dự án [[6], [7], [8]]. Dự đoán chi phí chính xác giảm thiểu rủi ro tài chính, từ đó tối ưu hóa kế hoạch tài chính và lựa chọn thiết kế, giúp đảm bảo hoàn thành suôn sẻ các dự án xây dựng và tránh đứt gãy chuỗi vốn trong quá trình xây dựng [9,10].

Bên cạnh yếu tố kỹ thuật, công tác dự báo chi phí còn có ý nghĩa chiến lược trong hoạch định đầu tư và ra quyết định tài chính. Một mô hình dự báo chi phí hiệu quả không chỉ giúp chủ đầu tư xác định quy mô vốn cần thiết mà còn hỗ trợ phân bổ ngân sách hợp lý giữa các hạng mục công trình. Dự báo chính xác còn tạo cơ sở cho việc lựa chọn phương án thiết kế, vật liệu, và giải pháp thi công tối ưu về mặt chi phí, góp phần nâng cao tính cạnh tranh của doanh nghiệp xây dựng. Hơn nữa, trong bối cảnh thị trường xây dựng biến động mạnh do yếu tố kinh tế vĩ mô, chính sách, và rủi ro chuỗi cung ứng, việc áp dụng các phương pháp dự báo hiện đại giúp doanh nghiệp chủ động hơn trong ứng phó và kiểm soát rủi ro.

Ngoài ra, công tác dự báo chi phí đóng vai trò quan trọng trong việc đánh giá tính khả thi của dự án ngay từ giai đoạn lập báo cáo đầu tư. Dự toán chi phí chính xác giúp đảm bảo cân đối vốn, hạn chế tình trạng điều chỉnh tổng mức đầu tư, đồng thời là căn cứ để cơ quan quản lý nhà nước thẩm định và phê duyệt dự án. Trong bối cảnh chuyển đổi số ngành xây dựng hiện nay, việc nâng cao năng lực dự báo chi phí thông qua dữ

liệu và mô hình hóa thông minh được xem là xu hướng tất yếu nhằm nâng cao tính minh bạch, hiệu quả và bền vững của các dự án đầu tư xây dựng.

Mặc dù tầm quan trọng của nó, dự đoán chi phí xây dựng vẫn là một nhiệm vụ đầy thách thức do thông tin có sẵn hạn chế và mức độ không chắc chắn cao ở giai đoạn đầu [3]. Phương pháp thông thường để ước tính chi phí xây dựng dựa vào kinh nghiệm của kỹ sư chi phí và các đánh giá chủ quan về các yếu tố chi phí [4], [5]. Tuy nhiên, các đánh giá chủ quan có thể không nhất quán và không đáng tin cậy, dẫn đến kết quả sai lệch và do đó gây ra tổn thất tài chính cho các bên liên quan của dự án [6], [7]. Ngoài ra, việc phụ thuộc nhiều vào chuyên môn của kỹ sư chi phí dẫn đến quy trình ước tính tốn nhiều công sức và thời gian, điều này không có lợi cho việc bàn giao các dự án xây dựng với các yêu cầu về tiến độ vốn đã chặt chẽ [8].

1.2. CÁC NGHIÊN CỨU QUỐC TẾ VỀ DỰ BÁO CHI PHÍ XÂY DỰNG

Phương pháp ước tính chi phí trong xây dựng đóng vai trò quan trọng trong việc dự đoán và kiểm soát chi phí dự án, ảnh hưởng trực tiếp đến thành công của dự án. Trong các nghiên cứu quốc tế, nhiều phương pháp dự báo truyền thống đã được áp dụng nhằm mô hình hóa và dự đoán xu hướng biến động chi phí xây dựng theo thời gian. Theo đó, có hai nhóm phương pháp ước tính chi phí chính dựa trên thống kê gồm phương pháp tham số (parametric cost estimating) và phương pháp phi tham số (nonparametric cost estimating) [1].

Phương pháp tham số dựa trên giả định rằng mối quan hệ giữa chi phí và các biến đầu vào có thể được mô tả bằng một hàm toán học xác định, thường sử dụng các mô hình hồi quy tuyến tính hoặc phi tuyến để ước lượng chi phí dự án. Các mô hình này có ưu điểm là dễ diễn giải và phù hợp với dữ liệu có cấu trúc rõ ràng, song hạn chế ở khả năng mô tả các mối quan hệ phi tuyến phức tạp giữa các biến. Ngược lại, phương pháp phi tham số không yêu cầu giả định dạng hàm cụ thể giữa biến độc lập và biến phụ thuộc, mà dựa trên dữ liệu thực nghiệm để xác định mối quan hệ chi phí. Các phương pháp phi tham số phổ biến như k-nearest neighbors (k-NN) và case-based reasoning (CBR), được sử dụng phổ biến trong giai đoạn ước lượng chi phí ban đầu khi dữ liệu dự án còn hạn chế.

Trong những năm gần đây, sự phát triển của trí tuệ nhân tạo (AI) và học máy đã mở rộng phạm vi của các phương pháp phi tham số. Các mô hình như mạng nơ-ron nhân

tạo (ANNs), máy học véc-tơ hỗ trợ (SVR), rừng ngẫu nhiên (RF) đã được chứng minh là có khả năng mô phỏng tốt hơn các mối quan hệ phi tuyến, phức tạp và đa chiều trong dữ liệu chi phí xây dựng [2, 3]. Những mô hình này không chỉ nâng cao độ chính xác của dự báo, mà còn hỗ trợ các nhà quản lý và nhà thầu trong việc ra quyết định tài chính ở giai đoạn sớm của dự án, góp phần tối ưu hóa phân bổ nguồn vốn và giảm thiểu rủi ro vượt chi phí. Do đó, xu hướng kết hợp giữa các phương pháp thống kê truyền thống và học máy đang được xem là hướng nghiên cứu tiềm năng nhằm cải thiện khả năng dự báo chi phí trong lĩnh vực xây dựng hiện nay.

1.2.1. Các mô hình dự báo hồi quy

Mô hình hồi quy (Regression Model) là một trong những phương pháp dự báo phổ biến và nền tảng nhất trong lĩnh vực ước lượng chi phí xây dựng. Phương pháp dựa trên tham số này mô tả mối quan hệ giữa biến phụ thuộc (thường là chi phí xây dựng) và một hoặc nhiều biến độc lập (như diện tích sàn, số tầng, thời gian thi công,…). Bằng cách xác định phương trình hồi quy, mô hình có thể ước lượng giá trị chi phí dựa trên các đặc trưng đầu vào của dự án. Mô hình hồi quy tuyến tính được sử dụng rộng rãi do tính đơn giản, dễ diễn giải và khả năng áp dụng cho dữ liệu có quan hệ tuyến tính rõ ràng. Tuy nhiên, trong các trường hợp mối quan hệ giữa các biến mang tính phi tuyến hoặc tương tác phức tạp, các mô hình hồi quy phi tuyến hoặc hồi quy đa biến nâng cao thường được sử dụng để cải thiện độ chính xác của dự báo. Nhờ đó, hồi quy trở thành nền tảng quan trọng cho các mô hình dự báo truyền thống và là cơ sở để phát triển các phương pháp hiện đại như học máy và trí tuệ nhân tạo.

Williams (2003) [4] đã nghiên cứu mối quan hệ giữa giá thầu thấp và chi phí hoàn thành thực tế của các dự án xây dựng đường bộ được đấu thầu cạnh tranh. Dữ liệu được thu thập từ năm cơ quan quản lý dự án đường bộ và dự án nạo vét tại Hoa Kỳ. Kết quả phân tích cho thấy việc biến đổi dữ liệu bằng logarit tự nhiên của giá thầu thấp và chi phí hoàn thành giúp xây dựng các mô hình hồi quy tuyến tính với hệ số tương quan cao, cho phép dự đoán chi phí hoàn thành dựa trên giá thầu thấp. Mô hình hồi quy còn được ứng dụng để dự đoán tổng chi phí cho nhóm dự án, hỗ trợ công tác lập ngân sách vốn. Mặc dù mô hình dự đoán khá chính xác với các dự án đường bộ, chúng kém hiệu quả hơn với dự án nạo vét do tính phức tạp và biến động cao. Tác giả chỉ ra cần phát triển

mô hình riêng biệt cho từng loại dự án và từng cơ quan quản lý, đồng thời gợi ý rằng chia nhỏ các dự án lớn thành nhiều phần nhỏ có thể giúp kiểm soát chi phí tốt hơn.

Skitmore và cộng sự (2003) [5] đã phát triển các mô hình dự báo thời gian và chi phí thực tế trong các dự án xây dựng dựa trên dữ liệu khảo sát 93 dự án xây dựng tại Úc. Các mô hình sử dụng phân tích hồi quy đa biến kết hợp với kỹ thuật hồi quy chéo kiểm định để nâng cao độ chính xác và giảm thiểu sai số dự báo. Biến phụ thuộc chính là thời gian và chi phí thực tế, trong khi các biến độc lập bao gồm thời gian và chi phí hợp đồng ước tính, loại khách hàng, loại dự án, phương pháp lựa chọn nhà thầu và hình thức hợp đồng. Kết quả cho thấy các mô hình có thể dự báo khá chính xác thời gian và chi phí thực tế khi biết các thông số hợp đồng và đặc điểm dự án, đồng thời mô hình cũng cho phép dự báo khi thời gian và chi phí hợp đồng chỉ được ước tính. Phân tích độ nhạy cho thấy sai số chi phí thực tế tương đối ổn định với các quy mô dự án khác nhau. Mô hình được trình bày dưới dạng các đường cong giúp khách hàng và nhà thầu dễ dàng lựa chọn phương án phù hợp với mức độ rủi ro và yêu cầu cụ thể của dự án. Tổng thể nghiên cứu cung cấp một công cụ thực tiễn hữu ích nhằm hỗ trợ quản lý xây dựng, giảm thiểu rủi ro và nâng cao hiệu quả dự án.

Lowe và cộng sự (2006) [6] đã phát triển các mô hình hồi quy tuyến tính để dự đoán chi phí xây dựng công trình dựa trên dữ liệu của 286 dự án tại Vương quốc Anh. Mô hình dự đoán chi phí trên mét vuông, logarit của chi phí, và logarit của chi phí trên mét vuông được xây dựng và đánh giá bằng cả phương pháp lựa chọn biến tiến (forward) và lùi (backward), tạo thành sáu mô hình khác nhau. Kết quả cho thấy năm biến độc lập xuất hiện trong tất cả các mô hình là diện tích sàn trong, chức năng sử dụng, thời gian thi công, lắp đặt cơ khí và cọc nền, cho thấy đây là những yếu tố chính ảnh hưởng đến chi phí xây dựng. Mô hình logarit chi phí với phương pháp biến tiến cho kết quả tốt nhất với hệ số xác định (R^2) là 0,661 và sai số trung bình tuyệt đối phần trăm (MAPE) là 19,3%, vượt trội hơn so với phương pháp ước tính truyền thống (MAPE khoảng 25%). Nhóm tác giả cũng so sánh hiệu quả của mô hình hồi quy với ANNs, trong đó ANNs cho kết quả tốt hơn nhưng sự khác biệt không lớn. Tóm lại, mô hình hồi quy phát triển mang lại độ chính xác cao hơn so với phương pháp truyền thống và có thể hỗ trợ đắc lực cho các chuyên gia định giá xây dựng trong giai đoạn đầu của dự án, đồng thời làm cơ sở để phát triển các mô hình mạng nơ-ron phức tạp hơn.

Jafarzadeh và cộng sự (2014) [7] đã phát triển mô hình hồi quy đa biến tuyến tính để dự đoán chi phí xây dựng cải tạo chống động đất cho các tòa nhà tại khu vực có nguy cơ động đất cao ở Iran. Dữ liệu được thu thập từ 158 công trình trường học có cấu trúc khung, với 14 biến độc lập được khảo sát. Phương pháp loại dần ngược được sử dụng để xác định các biến có ảnh hưởng đáng kể đến chi phí, đồng thời kiểm tra các giả định của mô hình hồi quy như tính chuẩn của sai số, không đa cộng tuyến, phương sai đồng nhất và không tự tương quan. Kết quả cho thấy các biến ảnh hưởng lớn nhất đến chi phí xây dựng là diện tích mặt bằng tổng thể và số tầng của tòa nhà, tiếp theo là loại kết cấu, mức độ động đất, loại đất nền, trọng lượng và sự bất thường trong mặt bằng. Đáng chú ý, tuổi tòa nhà và việc tuân thủ các tiêu chuẩn thiết kế động đất đầu tiên không có ảnh hưởng đáng kể. Mô hình đơn giản nhất nhưng hiệu quả nhất là mô hình log-log chỉ sử dụng biến diện tích mặt bằng có thể dự đoán chi phí với độ chính xác chấp nhận được (MAPE khoảng 22,9%). Tóm lại, nghiên cứu đã cung cấp một công cụ dự báo chi phí cải tạo chống động đất đơn giản nhưng hiệu quả, dựa trên dữ liệu thực tế và phân tích thống kê chặt chẽ. Mô hình này không chỉ giúp tiết kiệm thời gian và chi phí cho giai đoạn lên kế hoạch mà còn tạo nền tảng cho các nghiên cứu sâu hơn trong lĩnh vực dự báo chi phí xây dựng chống động đất và các công trình phức tạp khác.

Lshamrani (2017) [8] đã trình bày một mô hình hồi quy đa biến nhằm dự đoán chi phí ban đầu của các tòa nhà đại học truyền thống và bền vững tại Bắc Mỹ. Mô hình này được xây dựng dựa trên dữ liệu chi phí xây dựng trung bình quốc gia năm 2014. Các biến đầu vào của mô hình bao gồm diện tích tòa nhà, số tầng, chiều cao mỗi tầng, loại kết cấu (thép hoặc bê tông) và chỉ số bền vững (truyền thống hoặc bền vững). Mô hình được xây dựng qua ba giai đoạn chính: kiểm định chất lượng dữ liệu ban đầu, phát triển mô hình hồi quy và cuối cùng là xác thực mô hình với dữ liệu thực tế. Phân tích hồi quy cho thấy, tất cả các biến đầu vào đều có ảnh hưởng có ý nghĩa thống kê đến chi phí xây dựng ban đầu, với giá trị R^2 điều chỉnh đạt khoảng 87,3%, cho thấy mô hình có khả năng dự đoán tốt. Mô hình hồi quy cho kết quả dự báo chính xác đến 94,3%, cho thấy mô hình này có thể ứng dụng thực tiễn hiệu quả. Điểm nổi bật của nghiên cứu là việc cung cấp một công cụ đơn giản, dễ sử dụng để các trường đại học và các bên liên quan có thể ước tính chi phí xây dựng ban đầu cho các dự án xây dựng tòa nhà đại học (quy mô dưới 3 tầng), đồng thời so sánh được chi phí giữa các tòa nhà truyền thống và bền vững, cũng

như giữa các loại kết cấu bê tông và thép. Điều này hỗ trợ các bên liên quan ra quyết định đầu tư và lựa chọn thiết kế nhằm đạt hiệu quả kinh tế và bền vững.

Nhận xét:

Ước tính chi phí theo phương pháp hồi quy tuyến tính so sánh giả định mối quan hệ tuyến tính giữa chi phí cuối cùng và các biến thiết kế cơ bản của dự án. Tuy nhiên, trong thực tế, mối quan hệ giữa các biến đầu vào đầu ra đôi khi ở dạng phi tuyến tính, đa chiều và phụ thuộc lẫn nhau. Do đó, các mô hình hồi quy tuyến tính thường không đủ khả năng nắm bắt các mối quan hệ phức tạp giữa các biến trong dữ liệu xây dựng, dẫn đến độ chính xác dự báo và khả năng khái quát hóa dưới mức tối ưu. Trong bối cảnh đó, trí tuệ nhân tạo, đặc biệt là kỹ thuật học máy được xem như một giải pháp hứa hẹn để nâng cao độ chính xác và hiệu quả của việc dự đoán chi phí xây dựng.

1.2.2. Các mô hình dự báo học máy đơn

Để khắc phục những hạn chế của mô hình hồi quy, các mô hình học máy đơn như ANNs, SVR đã được ứng dụng rộng rãi. Các mô hình này có khả năng học trực tiếp từ dữ liệu, nhận diện các mẫu tiềm ẩn và mô hình hóa mối quan hệ phi tuyến giữa các biến đầu vào và chi phí xây dựng. Nhờ đó, chúng giúp nâng cao độ chính xác của dự báo, đặc biệt trong các giai đoạn thiết kế sớm khi dữ liệu còn hạn chế và biến động cao.

Mạng nơ-ron nhân tạo (ANNs) là công cụ trí tuệ nhân tạo có khả năng học, tổng quát hóa và thích nghi với dữ liệu, thích hợp cho các bài toán dự đoán và ước tính chi phí trong xây dựng, đặc biệt ở giai đoạn sớm. Các nghiên cứu trước đây đã ứng dụng thành công ANNs trong ước tính chi phí các dự án xây dựng khác nhau như trường học, cầu đường, công trình thể thao và nhà ở. Các mô hình ANNs được phát triển có tiềm năng ứng dụng rộng rãi trong các lĩnh vực quản lý dự án xây dựng, đặc biệt giúp các bên liên quan như nhà thầu, tư vấn, cơ quan nhà nước, nhà đầu tư và các nhà nghiên cứu có một công cụ nhanh chóng, hữu hiệu để ước lượng chi phí xây dựng ở giai đoạn đầu của dự án.

Juszczyk (2017) [9] đã ứng dụng mô hình ANNs, trong ước tính chi phí xây dựng theo phương pháp phi tham số. Theo nghiên cứu của tác giả, các nghiên cứu trước đó chủ yếu tập trung vào việc áp dụng ANNs để ước tính chi phí ở giai đoạn sơ bộ, trên phạm vi toàn bộ dự án hoặc công trình xây dựng như nhà ở, trường học, cầu đường. Kết

quả cho thấy các mô hình ANNs có độ chính xác cao, giá trị MAPE thường dưới 15%, phù hợp với yêu cầu ước tính sơ bộ. Từ đó, tác giả đề xuất mở rộng ứng dụng mô hình ANNs sang mức độ chi tiết hơn, tức là ước tính chi phí từng công việc xây dựng cụ thể trong quy trình sản xuất công trình. Để thực hiện điều này, cần có sự phân chia hợp lý công trình thành các công việc, xác định các biến dự báo chi phí phù hợp cho từng công việc và xây dựng nhiều mô hình ANNs tương ứng. Các thách thức chính gồm việc thu thập và tổ chức dữ liệu chi tiết, phát triển hệ thống phân loại công trình, đồng bộ hóa dữ liệu cũng như khả năng xử lý và huấn luyện mô hình ANN hiệu quả.

Alrasheed và cộng sự (2015) [10] phát triển một mô hình ANNs nhằm cải thiện độ chính xác trong việc dự đoán chi phí giai đoạn đầu của các dự án xây dựng công cộng tại Kuwait. Dữ liệu huấn luyện được lấy từ 28 dự án công do các nhà thầu lớn thực hiện, với sáu biến đầu vào chính gồm loại công trình, năm hợp đồng được trao, chủ sở hữu, vị trí, khối lượng đào đất, và khối lượng bê tông; biến dự báo là chi phí dự toán được phê duyệt. Mô hình ANNs được xây dựng theo cấu trúc truyền thẳng gồm hai lớp ẩn và sử dụng thuật toán học lan truyền ngược. Kết quả cho thấy chỉ số MAPE của mô hình tốt nhất chỉ 0,72%, tương ứng độ chính xác 99,28%. Mô hình đề xuất cũng được kiểm thử trên 5 dự án mới, cho thấy độ tin cậy và khả năng ứng dụng thực tế cao. Bên cạnh đó, nghiên cứu chỉ ra khối lượng bê tông và năm trao hợp đồng là hai yếu tố quan trọng nhất ảnh hưởng đến chi phí dự toán. Tuy nhiên, nghiên cứu cũng thừa nhận hạn chế về kích thước dữ liệu nhỏ và tính đa dạng hạn chế, đồng thời đề xuất các hướng nghiên cứu mở rộng như mở rộng dữ liệu, áp dụng các kỹ thuật xác thực chéo, tích hợp các biến động giá nguyên vật liệu theo thời gian và liên kết mô hình với các nền tảng mô hình thông tin công trình (BIM) để nâng cao khả năng dự báo và ứng dụng thực tiễn trong quy trình xây dựng số hóa tại Kuwait.

Maya và cộng sự (2023) [11] đã phát triển mô hình ANNs để dự đoán hiệu suất các dự án xây dựng tại Syria dựa trên những yếu tố ảnh hưởng quan trọng. Từ khảo sát và phân tích ý kiến của các chuyên gia, tổng cộng 34 yếu tố ảnh hưởng đã được xác định và phân loại thành ba nhóm theo mức độ tác động lên hiệu suất dự án. Trong đó, 6 yếu tố có ảnh hưởng lớn nhất được lựa chọn làm đầu vào cho mô hình dự báo gồm phối hợp và cam kết của các bên dự án, ước tính tiến độ, kinh nghiệm và sự sẵn có của đội ngũ dự án, sự hỗ trợ từ ban lãnh đạo cấp cao, khả năng thanh toán đúng hạn, và sự hiện diện

của phần mềm quản lý dự án. Mô hình ANNs đề xuất được đào tạo và kiểm tra trên dữ liệu thực tế của 40 dự án xây dựng tại Syria, sử dụng bốn chỉ số đánh giá hiệu suất chính (thời gian, chi phí, chất lượng và lợi nhuận) để tính toán hiệu suất tổng thể. Kết quả cho thấy độ chính xác dự đoán của mô hình đề xuất rất cao (tới 96,1%). Phân tích tương quan Pearson chỉ ra mối liên hệ mạnh mẽ giữa các yếu tố đầu vào và hiệu suất đầu ra, xác nhận tính phù hợp và độ tin cậy của mô hình. Kết quả nghiên cứu cũng khẳng định vai trò quan trọng hàng đầu của yếu tố phối hợp và cam kết giữa các bên dự án (chiếm 30,9% ảnh hưởng), tiếp theo là ước tính tiến độ (25,4%), kinh nghiệm và sự sẵn có của đội ngũ dự án (24,5%), và sự hỗ trợ từ ban lãnh đạo cấp cao (14,3%). Nghiên cứu này, vì vậy, cung cấp một công cụ hữu ích giúp các nhà quản lý dự án tại Syria có thể dự đoán và cải thiện hiệu suất dự án ngay từ giai đoạn lập kế hoạch, giúp thiết lập các chiến lược chủ động để quản lý và kiểm soát hiệu quả các yếu tố ảnh hưởng.

Awad (2019) [2] đã sử dụng mô hình ANNs để ước tính chi phí xây dựng sơ bộ tại Yemen. Nghiên cứu sử dụng dữ liệu thực tế từ 136 dự án xây dựng đã hoàn thành trong giai đoạn 2011-2015 tại Yemen, thu thập thông tin về 17 biến độc lập ảnh hưởng đến chi phí dự án như diện tích sàn, số tầng, loại móng, loại gạch, độ phức tạp dự án, hệ thống HVAC,…Dữ liệu được mã hóa và xử lý thông qua phần mềm NeuroSolution 6 để xây dựng, huấn luyện và kiểm tra mô hình ANNs. Kết quả cho thấy mô hình ANNs có độ chính xác cao với sai số trung bình tuyệt đối (MAE) chỉ khoảng 720,97 USD trên tổng chi phí trung bình hơn 514.000 USD, sai số phần trăm trung bình tuyệt đối (MAPE) chỉ 0.14%, tương đương độ chính xác 99.86%. Phân tích độ nhạy cho thấy các biến như "loại gạch" và "độ phức tạp" có ảnh hưởng lớn nhất đến chi phí, trong khi các biến như "hệ thống HVAC" và "loại dự án" có ảnh hưởng nhỏ hơn. Nghiên cứu cũng chỉ ra, mô hình ANNs cho kết quả tốt hơn đáng kể so với các phương pháp truyền thống như mô hình hồi quy.

Karadimos và cộng sự (2025) [12] đã phát triển các mô hình ANNs nhằm dự đoán chi phí xây dựng các nhà máy xử lý nước thải tại Hy Lạp. Nghiên cứu sử dụng bộ dữ liệu gồm 31 nhà máy xử lý nước thải ở Hy Lạp, thu thập các biến định lượng và định tính liên quan như công suất xử lý, lưu lượng nước thải đầu vào, tải hữu cơ, tuyến xử lý nước thải và tuyến xử lý bùn, cũng như vị trí địa lý. Phân tích tương quan cho thấy các biến định lượng như công suất xử lý, dân số đỉnh điểm, tải hữu cơ đầu vào và lưu lượng

nước thải đầu vào có mối tương quan mạnh với chi phí xây dựng, trong khi các biến định tính không có mối tương quan đáng kể. Dựa trên kết quả phân tích, nhiều mô hình ANNs được xây dựng với các biến đầu vào được lựa chọn theo thứ tự mức độ tương quan giảm dần. Các mô hình này cho kết quả sai số bình phương trung bình (MSE) rất thấp trên dữ liệu huấn luyện, cho thấy mô hình ANNs đã học rất tốt mối quan hệ giữa các biến. Tuy nhiên, trên tập kiểm tra, MSE tăng lên, cảnh báo khả năng xảy ra hiện tượng quá khớp. Nghiên cứu cũng đề xuất một số giải pháp để giảm thiểu vấn đề này như tăng lượng dữ liệu huấn luyện, giảm độ phức tạp của mạng hoặc giới hạn thời gian huấn luyện. Nhóm tác giả cũng gợi mở hướng nghiên cứu trong tương lai về việc sử dụng thêm các biến đầu vào khác hoặc áp dụng ANNs trong dự báo các chỉ tiêu khác như phát thải khí nhà kính của các nhà máy xử lý nước thải.

Bên cạnh đó, các nghiên cứu đã đánh giá khả năng dự báo chi phí xây dựng của một loạt các mô hình học máy. Chẳng hạn, Chen và cộng sự (2025) [3] trình bày một khuôn khổ toàn diện ứng dụng các mô hình học máy tiên tiến nhằm dự đoán chi phí xây dựng một cách minh bạch và tin cậy. Nghiên cứu đánh giá hiệu quả của 10 mô hình học máy khác nhau, bao gồm các phương pháp hồi quy (Ridge, Lasso, Elastic Net), K-Nearest Neighbor (KNN), và các phương pháp tập hợp nâng cao như XGBoost, CatBoost, và HistGradient Boosting. Kết quả cho thấy HistGradient Boosting đạt hiệu suất tốt nhất trên tập dữ liệu kiểm tra. Ngoài các chỉ số R^2, RMSE nghiên cứu còn áp dụng phân tích khoảng tin cậy để đánh giá độ tin cậy của dự đoán và sử dụng kỹ thuật giải thích mô hình SHAP (SHapley Additive exPlanations) để làm rõ ảnh hưởng của các biến đầu vào đến kết quả dự đoán. Phương pháp luận này không chỉ nâng cao độ chính xác mà còn cung cấp cái nhìn sâu sắc về tính không chắc chắn và khả năng giải thích của mô hình, giúp các nhà quản lý dự án xây dựng đưa ra quyết định hiệu quả hơn. Nghiên cứu còn tổng hợp các công nghệ cách mạng trong cuộc Cách mạng Công nghiệp 4.0 và ứng dụng học máy trong lĩnh vực xây dựng, từ thiết kế kết cấu, giám sát sức khỏe công trình, đến tái chế vật liệu xây dựng nhằm thúc đẩy phát triển bền vững. Kết luận nhấn mạnh tiềm năng to lớn của học máy trong việc đổi mới phương pháp dự đoán chi phí xây dựng, đồng thời đề xuất hướng nghiên cứu tương lai tập trung vào việc tích hợp giải thích mô hình và phân tích độ tin cậy.

Chakraborty và cộng sự (2020) [13] đã so sánh khả năng dự đoán xác suất chi phí xây dựng trong giai đoạn thiết kế ban đầu của 6 thuật toán học máy khác nhau, gồm – hồi quy tuyến tính (LR), ANNs, rừng ngẫu nhiên (RF), tăng cường gradient cực đoan (extreme gradient boosting – EGB), tăng cường gradient nhẹ (light gradient boosting – LGB) và tăng cường gradient tự nhiên (natural gradient boosting – NGB). Kết quả phân tích cho thấy mô hình LGB và NGB vượt trội các mô hình khác về độ chính xác và tốc độ dự báo. So sánh giữa chi phí dự đoán và chi phí thực tế xác nhận sự phù hợp tốt với hệ số xác định R^2 là 0,99, và RMSE bằng 0,5. Bên cạnh đó, mô hình kết hợp được đề xuất có thể cung cấp ước tính độ không chắc chắn thông qua các dự đoán xác suất cho các đầu ra giá trị thực.

Việc ước tính chính xác chi phí xây dựng từ giai đoạn đầu của dự án bệnh viện là rất quan trọng để lập ngân sách hiệu quả và quản lý rủi ro trong phát triển cơ sở hạ tầng y tế. Jezeh và cộng sự (2025) [14] dự đoán chi phí xây dựng cuối cùng của các dự án bệnh viện dựa trên các thuộc tính ban đầu của dự án bằng cách sử dụng các phương pháp hồi quy tiên tiến. Nhóm tác giả đã áp dụng bốn kỹ thuật hồi quy chính, gồm hồi quy tuyến tính (LR), hồi quy SVR, hồi quy ANNs, và hồi quy rừng ngẫu nhiên (Random Forest Regression - RFR). Dữ liệu nghiên cứu là một bộ dữ liệu giả lập gồm 100 dự án bệnh viện với các biến số như diện tích xây dựng, số giường bệnh, vùng chịu động đất, loại hợp đồng, phương pháp chế tạo, và chứng nhận bền vững. Mỗi mô hình được huấn luyện và đánh giá qua các chỉ số RMSE, MAPE, và R^2. Kết quả cho thấy mô hình RFR vượt trội nhất với R^2 đạt 0,65, RMSE và MAPE thấp nhất, cho thấy khả năng nắm bắt các mối quan hệ phi tuyến và đa chiều trong dữ liệu tốt hơn các mô hình khác. Trong khi đó SVR và ANNs thể hiện hiệu quả kém do hiện tượng quá khớp và dữ liệu hạn chế. Mô hình LR mặc dù có tính giải thích cao nhưng không thể nắm bắt được các tương tác phức tạp giữa các biến đầu vào. Nghiên cứu nhấn mạnh rằng các phương pháp hồi quy tiên tiến, đặc biệt là mô hình RFR, có thể cải thiện đáng kể độ chính xác dự đoán chi phí xây dựng bệnh viện ở giai đoạn đầu, hỗ trợ công tác lập kế hoạch, quản lý ngân sách và giảm thiểu rủi ro tài chính trong lĩnh vực cơ sở hạ tầng y tế.

1.2.3. Các mô hình dự báo hỗn hợp

Mặc dù các mô hình học máy đơn như ANNs, SVR hay RF đã chứng minh được khả năng dự báo chi phí xây dựng hiệu quả hơn so với các phương pháp hồi quy truyền

thống, song mỗi mô hình đều tồn tại những hạn chế nhất định. Chẳng hạn, ANNs có thể gặp khó khăn trong việc xác định cấu trúc mạng tối ưu, SVR phụ thuộc mạnh vào việc lựa chọn hàm nhân, trong khi RF dễ bị quá khớp với dữ liệu huấn luyện. Để tận dụng ưu điểm và khắc phục điểm yếu riêng của từng mô hình, các nghiên cứu gần đây đã phát triển mô hình dự báo hỗn hợp (hybrid models) và mô hình tích hợp (ensemble models). Các mô hình này kết hợp kết quả của nhiều thuật toán học máy nhằm cải thiện độ chính xác, tính ổn định và khả năng khái quát hóa của dự báo chi phí. Nhờ đó, phương pháp dự báo hỗn hợp ngày càng được xem là hướng tiếp cận tiên tiến và tiềm năng trong lĩnh vực ước tính chi phí xây dựng hiện nay.

Trong mô hình dự báo hỗn hợp này, các thuật toán được tích hợp với mô hình dự báo cơ sở để tối ưu tham số nhằm cải thiện khả năng dự báo. Chẳng hạn, Mahmoodzadeh và cộng sự [15] đã ứng dụng bốn phương pháp học máy, gồm LR, hồi quy tiến trình Gaussian (Gaussian Process Regression - GPR), SVR, và cây quyết định (Decision Tree - DT) để dự đoán chi phí và thời gian thi công các dự án đào hầm. Dữ liệu huấn luyện gồm 350 bộ dữ liệu từ 34 dự án đào hầm thực tế trên thế giới với 16 tham số đầu vào ảnh hưởng đến chi phí và thời gian, trong khi dữ liệu kiểm tra lấy từ một dự án hầm thử nghiệm với 181 bộ dữ liệu. Thuật toán tối ưu hóa bầy sói xám (Grey Wolf Optimization - GWO) được sử dụng để tinh chỉnh siêu tham số của các mô hình học máy nhằm nâng cao hiệu suất dự đoán. Kết quả dự đoán của các mô hình được so sánh với dữ liệu thực tế cho thấy, mô hình LR có hiệu quả dự đoán tốt nhất, tiếp theo là GPR, SVR và cuối cùng là DT. Độ chính xác của mô hình được đánh giá qua các chỉ số thống kê như hệ số xác định, MAPE, sai số trung bình căn bậc hai (RMSE) và được xác thực qua kỹ thuật kiểm định chéo 5 lần. Phân tích độ nhạy cho thấy mực nước ngầm có ảnh hưởng lớn nhất đến chi phí. Tóm lại, nghiên cứu cung cấp một công cụ dự báo tin cậy giúp các nhà quản lý dự án và kỹ sư đưa ra quyết định tối ưu hóa kế hoạch thi công, giảm thiểu chi phí và thời gian trong các dự án đào hầm phức tạp.

Chen và cộng sự (2012) [16] đã trình bày một mô hình dự báo chi phí hoàn thành dự án (Estimate at Completion - EAC) xây dựng thông qua việc kết hợp ba kỹ thuật trí tuệ nhân tạo tiên tiến gồm logic mờ (fuzzy logic), máy vector hỗ trợ có trọng số (wSVM), và thuật toán di truyền nhanh dạng hỗn tạp (fast messy Genetic Algorithm - fmGA). Mô hình đề xuất sử dụng wSVM để xử lý dữ liệu chuỗi thời gian và ưu tiên các

điểm dữ liệu gần đây hơn bằng các hàm trọng số thời gian khác nhau (hàm tuyến tính, hàm bậc hai, hàm mũ). Logic mờ giúp mô hình xử lý sự bất định và mơ hồ trong dữ liệu chi phí. Thuật toán fmGA được ứng dụng để tối ưu hoá đồng thời các tham số quan trọng của mô hình. Dữ liệu thực nghiệm được thu thập từ 13 dự án xây dựng thực tế tại Đài Loan với 10 yếu tố đầu vào quan trọng ảnh hưởng đến chi phí hoàn thành dự án như tiến độ xây dựng, chi phí thực tế, chỉ số hiệu suất chi phí, chỉ số hiệu suất tiến độ, và các yếu tố môi trường như số ngày mưa. Mô hình được huấn luyện trên 11 dự án và kiểm tra trên 2 dự án còn lại. Kết quả so sánh cho thấy mô hình dự báo EAC đề xuất vượt trội hơn hẳn các các công thức EAC truyền thống (chỉ đạt khoảng 36-50% độ chính xác dự báo), và các mô hình trí tuệ nhân tạo trước đó. Tóm lại, mô hình không những nâng cao độ chính xác mà còn giảm thiểu sự can thiệp của con người trong việc thiết lập cấu hình hàm thành viên và lựa chọn tham số, giúp việc ứng dụng trong thực tế trở nên dễ dàng và hiệu quả hơn cho các nhà quản lý dự án xây dựng.

Jin và cộng sự (2026) [17] đề xuất một khung dự đoán thông minh tích hợp kết hợp Quy trình phân tích thứ bậc mờ (Fuzzy Analytic Hierarchy Process - FAHP) để lựa chọn đặc trưng với mạng thần kinh truyền ngược được tối ưu hóa bằng thuật toán di truyền (GA-BPNN). FAHP định lượng một cách có hệ thống đánh giá của chuyên gia để xác định các đặc trưng xây dựng phù hợp nhất, trong khi GA cải thiện sự hội tụ và mạnh mẽ của mô hình BPNN bằng cách tối ưu các tham số của nó. Mô hình đề xuất được kiểm chứng bằng cách sử dụng tập dữ liệu thực tế gồm 4.552 dự án xây dựng nhà ở. Kết quả cho thấy khung FAHP-GA-BPNN vượt trội đáng kể so với các mô hình chuẩn, đạt RMSE là 79,5, MAE là 65,2 và R^2 là 0,93 trên tập dữ liệu kiểm chứng. Nghiên cứu này, vì vậy, không chỉ đóng góp một phương pháp luận có thể mở rộng và thích ứng để ước tính chi phí thông minh mà còn cung cấp những hiểu biết thực tế để nâng cao việc ra quyết định trong lập kế hoạch và lập ngân sách dự án nhà ở.

Gần đây, các mô hình học sâu (deep learning) được ứng dụng để dự đoán chi phí trong lĩnh vực xây dựng. Nó học hỏi từ dữ liệu lịch sử để liên tục cải thiện độ chính xác của kết quả mà không cần sự can thiệp của con người, và đã được chứng minh là đáng tin cậy cho phân tích và dự đoán dữ liệu trong ngành xây dựng. Chẳng hạn, Liu và cộng sự (2025) [18] đề xuất một khung học sâu siêu đồ thị để dự đoán chi phí thực tế của các dự án xây dựng ở giai đoạn đầu. Trước tiên, một công thức siêu đồ thị kết hợp hệ thống

phân cấp và mối quan hệ tương hỗ giữa các yếu tố chi phí được thiết lập. Một mô hình học sâu siêu đồ thị sau đó được phát triển dựa trên siêu đồ thị đã được xây dựng để dự đoán chi phí xây dựng từ đầu đến cuối. Tiếp theo, mô hình được giải thích định lượng về tầm quan trọng của yếu tố chi phí từ kết quả huấn luyện. Khung đề xuất được xác thực bằng cách sử dụng một tập dữ liệu chi phí xây dựng thực tế của các dự án trường học. Kết quả cho thấy độ chính xác cao trong dự đoán chi phí (MAPE đạt 10,72%) mà không cần sự can thiệp của con người. Tổng hợp các nghiên cứu quốc tế về dự báo chi phí xây dựng được trình bày ở Bảng 1.1 dưới đây.

Bảng 1.1. Tổng hợp các nghiên cứu quốc tế về dự báo chi phí xây dựng.

STT	Tên bài báo	Tác giả	Năm	Phương pháp sử dụng
1	Predicting final cost for competitively bid construction projects using regression models	Williams, T.P.	2003	Mô hình hồi quy
2	Forecast models for actual construction time and cost	Martin Skitmore, R. & Thomas Ng, S.	2003	Mô hình dự báo thống kê
3	Predicting Construction Cost Using Multiple Regression Techniques	Lowe, D.J. và cộng sự	2006	Hồi quy đa biến (Multiple Regression)
4	Predicting Seismic Retrofit Construction Cost for Buildings with Framed Structures Using Multilinear Regression Analysis	Jafarzadeh, R. và cộng sự	2014	Hồi quy đa tuyến tính (Multilinear Regression)
5	Construction cost prediction model for conventional and sustainable college buildings in North America	Alshamrani, O.S.	2017	Mô hình hồi quy đa biến
6	The Challenges of Nonparametric Cost Estimation of Construction Works with the use of Artificial Intelligence Tools	Juszczyk, M.	2017	Trí tuệ nhân tạo
7	Artificial Neural Network-based cost estimation for	Alrasheed, K. et al.	2025	Mạng nơ-ron nhân tạo (ANNs)

STT	Tên bài báo	Tác giả	Năm	Phương pháp sử dụng
	public construction projects in Kuwait			
8	Develop an artificial neural network (ANN) model to predict construction projects performance in Syria	Maya, R., Hassan, B., & Hassan, A.	2023	ANNs
9	Preliminary Construction Cost Estimate in Yemen by Artificial Neural Network	Hassan, A.	2019	ANNs
10	Development of Artificial Neural Networks for Predicting the Construction Costs of WWTPs in Greece	Karadimos, P. & Anthopoulos, L.	2025	ANNs
11	Transparent and reliable construction cost prediction using advanced machine learning and explainable AI	Chen, L. và cộng sự	2025	Học máy nâng cao (Advanced ML, Explainable AI)
12	A novel construction cost prediction model using hybrid natural and light gradient boosting	Chakraborty, D. và cộng sự	2020	Mô hình lai – LightGBM, Gradient Boosting
13	Predicting Final Construction Costs of Hospitals Based on Initial Project Attributes: An Advanced Regression Approach	Jezeh, M.V.A. & Shayegan, D.S.	2025	Hồi quy nâng cao (Advanced Regression)
14	Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects	Mahmoodzadeh, A. và cộng sự	2022	Học máy tối ưu hóa (Optimized ML)
15	A novel time-depended evolutionary fuzzy SVM inference model for estimating construction project at completion	Cheng, M.-Y. và cộng sự	2012	Hệ mờ và máy vector hỗ trợ (Fuzzy SVM)
16	A systematic intelligent prediction model for residential construction cost	Jin, G. & Yang, C.	2026	Mô hình lai: Fuzzy AHP +

STT	Tên bài báo	Tác giả	Năm	Phương pháp sử dụng
	based on fuzzy AHP and GA-BP neural network			GA-BP Neural Network
17	Actual construction cost prediction using hypergraph deep learning techniques	Liu, H. và cộng sự	2025	Học sâu (Deep Learning – Hypergraph Neural Network)

1.3. CÁC NGHIÊN CỨU Ở VIỆT NAM VỀ DỰ BÁO CHI PHÍ XÂY DỰNG

Tại Việt Nam, các nghiên cứu về dự báo chi phí xây dựng được triển khai muộn hơn so với các quốc gia phát triển, chủ yếu tập trung vào việc xây dựng cơ sở dữ liệu chi phí, phát triển mô hình ước lượng ban đầu, và đánh giá các yếu tố ảnh hưởng đến biến động giá. Các công trình nghiên cứu ban đầu thường áp dụng phương pháp hồi quy tuyến tính hoặc phi tuyến để xác định mối quan hệ giữa chi phí và các biến kỹ thuật của công trình như diện tích sàn, số tầng, kết cấu chịu lực, hoặc loại vật liệu. Một số nghiên cứu trong nước đã thử nghiệm các mô hình thống kê cải tiến hoặc phương pháp tham số nhằm nâng cao độ chính xác của dự báo ở giai đoạn thiết kế sơ bộ, tuy nhiên kết quả vẫn còn hạn chế do thiếu cơ sở dữ liệu đầy đủ, đồng bộ và có tính đại diện cao cho thị trường xây dựng Việt Nam.

Trong những năm gần đây, xu hướng ứng dụng trí tuệ nhân tạo và học máy đã bắt đầu được các nhà nghiên cứu Việt Nam quan tâm trong lĩnh vực dự báo chi phí. Một số công trình đã bước đầu sử dụng mô hình ANNs để mô phỏng mối quan hệ phi tuyến giữa các yếu tố thiết kế và chi phí thi công. Nhằm nâng cao độ chính xác trong giai đoạn thiết kế sơ bộ, góp phần quản lý chi phí hiệu quả và giảm thiểu tranh chấp, Van và cộng sự (2009) [19] đã nghiên cứu ứng dụng mô hình ANNs trong dự báo tổng chi phí xây dựng (Total Construction Cost - TCC) các dự án căn hộ tại Việt Nam. Qua khảo sát các chuyên gia xây dựng, 7 biến đầu vào quan trọng được lựa chọn, gồm diện tích sàn, số tầng, giá thép, xi măng, xăng dầu, cùng với hạng dự án. Mô hình ANNs ba lớp với một tầng ẩn gồm 5 nút được xây dựng và huấn luyện trên 14 bộ dữ liệu dự án căn hộ hoàn thành tại thành phố Hồ Chí Minh, sử dụng thuật toán Levenberg-Marquardt cùng hàm truyền tín hiệu. Mô hình được kiểm thử trên 5 dự án thực tế cho kết quả MAPE dưới 10%, tốt hơn so với các phương pháp khác như hồi quy đa biến và thuật toán di truyền. Mặc dù mô hình chưa được kiểm định nghiêm ngặt, nó mang lại giá trị thực tiễn cho

nhà quản lý, nhà đầu tư và các nhà nghiên cứu trong ngành xây dựng tại Việt Nam cũng như các nước đang phát triển trong khu vực Đông Nam Á, đặc biệt là Hàn Quốc – nhà đầu tư lớn tại Việt Nam. Nghiên cứu kết luận mô hình ANNs là công cụ hữu ích giúp dự báo chi phí xây dựng căn hộ hiệu quả ngay từ giai đoạn đầu, đồng thời đề xuất cập nhật và mở rộng dữ liệu để nâng cao độ chính xác trong tương lai.

Phạm Hoàng (2025) [20] đã đề xuất dự báo chi phí biện pháp thi công khoan tạo lỗ cọc khoan nhồi theo lý thuyết độ tin cậy. Theo đó, phương pháp Monte - Carlo được sử dụng để mô phỏng toàn bộ quá trình thi công dựa theo công nghệ thi công. Các số liệu về thời gian thi công các giai đoạn được đo đạc, thu thập và ghi chép đầy đủ để làm cơ sở cho việc xác định luật phân phối và thông số đầu vào cho quá trình mô phỏng. Thông số đầu ra là nhóm thông số về chất lượng bao gồm thời gian thi công các công đoạn, thể tích bentonitte cần sử dụng trong mỗi công đoạn, tốc độ cấp bentonitte trong mỗi công đoạn, độ sâu khoan được sau mỗi công đoạn thi công, độ sâu tổng cộng đã khoan được sau mỗi công đoạn, tổng thời gian thi công. Kết quả cho thấy, phương pháp đề xuất có khả năng dự báo chi phí thi công khoan tạo lỗ cọc khoan nhồi khi xem xét thời gian là yếu tố ngẫu nhiên, và có khả năng xây dựng các phân phối dẫn suất -phân phối tổng của một số công đoạn thành phần làm cơ sở mô phỏng các quá trình thi công khác.

Phong và cộng sự (2022) [21] đã ứng dụng mô hình ANNs để ước lượng chi phí xây dựng nhà xưởng trong giai đoạn đấu thầu. Mô hình sử dụng bộ dữ liệu thực tế gồm 35 dự án nhà xưởng tại Việt Nam với 11 yếu tố đầu vào ảnh hưởng đến chi phí xây dựng. Mô hình ANNs trong nghiên cứu gồm 3 lớp (lớp vào 26 nút, lớp ẩn 4 nút, lớp ra 1 nút), sử dụng hàm truyền Sigmoid và thuật toán lan truyền ngược. Qua quá trình huấn luyện và thử nghiệm, mô hình ANNs đạt được hệ số tương quan là 0,91 và giá trị MAPE là 22,49%, thể hiện khả năng dự báo chính xác và ổn định hơn so với các mô hình truyền thống như hồi quy tuyến tính và máy véc tơ hỗ trợ (SVM). Tóm lại, nghiên cứu cung cấp một phương pháp ước lượng chi phí xây dựng nhà xưởng hiện đại, dựa trên trí tuệ nhân tạo, mang lại hiệu quả cao và khả năng áp dụng rộng rãi trong điều kiện công nghiệp hóa nhanh chóng của Việt Nam.

Dang và Long [22] phát triển một số mô hình dự đoán để ước tính chi phí xây dựng kết cấu và thiết lập ước tính khoảng chi phí xây dựng kết cấu bằng cách sử dụng thông

tin thiết kế có sẵn trong giai đoạn đầu của các dự án xây dựng nhà ở. Dữ liệu về các dự án xây dựng nhà ở được thu thập dựa trên các tài liệu dự án từ các công ty xây dựng liên quan đến các thông số thiết kế và chi phí xây dựng kết cấu thực tế khi hoàn thành. Phương pháp bao che tầng (SEM) là cơ sở để xác định các thông số thiết kế tòa nhà, hình thành các biến tiềm năng và phát triển các mô hình ước tính chi phí bằng cách sử dụng phân tích hồi quy. Phương pháp bootstrap phi tham số được sử dụng để thiết lập ước tính khoảng cho chi phí xây dựng kết cấu. Nghiên cứu này có thể cung cấp cho các chuyên gia thực hành sự hiểu biết tốt hơn về sự không chắc chắn và biến động có trong ước tính chi phí. Do đó, họ có thể thực hiện những cải tiến hiệu quả về các phương pháp quản lý liên quan đến chi phí để nâng cao hiệu suất chi phí dự án.

Trong những năm gần đây, mức độ cạnh tranh giữa các nhà thầu xây dựng tại Việt Nam ngày càng lớn, đặc biệt ở phân khúc các dự án nhà cao tầng. Để doanh nghiệp tồn tại và phát triển, các nhà thầu cần đề ra chiến lược kinh doanh và chiến lược đấu thầu. Vũ và cộng sự (2019) [23] đã tìm ra 10 yếu tố chính ảnh hưởng đến chi phí xây dựng công trình nhà ở cao tầng trên địa bàn thành phố Hồ Chí Minh thông qua khảo sát các chuyên gia ngành xây dựng. Mô hình hồi quy tuyến tính được sử dụng để dự báo chi phí xây dựng từ dữ liệu của các dự án nhà ở cao tầng. Kết quả phân tích cho thấy mô hình hồi quy theo thủ thuật chọn biến stepwise hoặc forward cho sai số dự báo nhỏ nhất với MAPE bằng 17,5% và hệ số tương quan R^2 bằng 0,885. Tóm lại, nghiên cứu đã cung cấp một công cụ hỗ trợ thông tin cho doanh nghiệp khi đưa ra các quyết định tới giá dự thầu và kế hoạch dòng tiền trong giai đoạn thi công.

Đức và cộng sự (2019) [24] đã đề xuất mô hình gồm các mô-đun tối ưu hóa lợi nhuận và mô phỏng lợi nhuận của nhà thầu xây dựng. Thuật toán tiến hóa vi phân được đề xuất để tối ưu đồng thời thời gian và chi phí xây dựng. Sau đó, mô phỏng Monte Carlo được ứng dụng để đánh mức độ rủi ro của lợi nhuận tối ưu nhận được của nhà thầu xây dựng. Mô hình đề xuất được khảo sát qua dự án công trình xây dựng cao tầng tại Đà Nẵng, nhằm giúp các nhà quản lý xây dựng có một công cụ hỗ trợ trong việc ra quyết định, và hỗ trợ việc quản lý dự án xây dựng một cách hiệu quả hơn, tối ưu hơn.

Việc nâng cao hiệu quả công tác quản lý ở các dự án nhằm tối ưu chi phí thi công và lợi nhuận đạt được của từng gói thầu đóng vai trò quan trọng. Tự và cộng sự (2023) [25] đã xếp hạng 24 yếu tố gây vượt chi phí thi công trong các dự án nhà cao tầng. Dữ

liệu được thu thập từ các cá nhân liên quan đến quản lý dự án, quản lý chi phí và các cá nhân tham gia quản lý các hạng mục an toàn, chất lượng, vật tư liên quan mật thiết đến chi phí tại công trình xây dựng cao tầng, bao gồm cả tư vấn và nhà thầu. Qua tham khảo ý kiến chuyên gia, nghiên cứu đã phân tích mở rộng các giải pháp ứng phó với rủi ro đối với 10 yếu tố gây vượt chi phí mạnh nhất. Theo đó, bỏ giá đấu thầu không chính xác, sai sót trong thi công và làm lại, và sai thiếu trong thiết kế là 3 yếu tố gây vượt chi phí thi công lớn nhất. Bằng những kết quả thu được, nghiên cứu như một đề xuất nhằm hỗ trợ các nhà quản lý, các ban chỉ huy dự án trong việc ra các quyết định quan trọng trong quản lý chi phí hiệu quả tại các dự án xây dựng hiện nay.

Sự ra đời của mô hình thông tin công trình (BIM) đã tạo nên bước tiến mạnh mẽ trong lĩnh vực xây dựng. BIM được triển khai từ thiết kế đến gia công, thi công sản xuất, vận hành và bảo trì dựa trên mức độ tích hợp thông tin. Phước và cộng sự (2019) [26] đã ứng dụng BIM 5D để tự động hóa lập dự toán công trình xây dựng. Nghiên cứu này tạo tiền đề cho việc ứng dụng BIM và trí thông minh nhân tạo AI để quản lý chi phí.

Nhận xét*:*

Kết quả ứng dụng các mô hình học máy trong dự báo chi phí cho thấy các mô hình này có khả năng cải thiện độ chính xác dự báo so với các phương pháp truyền thống, đặc biệt trong các dự án nhà cao tầng và công trình dân dụng có quy mô lớn. Tuy nhiên, các nghiên cứu này vẫn còn mang tính thử nghiệm, quy mô dữ liệu nhỏ và chưa khai thác hiệu quả tiềm năng của các mô hình tích hợp (ensemble) hoặc mô hình lai (hybrid). Điều này cho thấy cần có thêm các nghiên cứu toàn diện hơn, sử dụng tập dữ liệu lớn, quy trình chuẩn hóa và phương pháp đánh giá khách quan để hình thành hệ thống dự báo chi phí đáng tin cậy, phù hợp với đặc thù của thị trường xây dựng Việt Nam. Các nghiên cứu ở Việt Nam về dự báo chi phí xây dựng được tổng hợp ở Bảng 1.2 dưới đây.

Bảng 1.2. Tổng hợp các nghiên cứu trong nước về dự báo chi phí xây dựng.

STT	Tên bài báo	Tác giả	Năm	Phương pháp sử dụng
1	Neural Network Model for Construction Cost Prediction of Apartment Projects in Vietnam	Van và cộng sự	2009	Mô hình ANNs
2	Nghiên cứu dự báo chi phí biện pháp thi công xây dựng theo lý thuyết độ tin cậy	Phạm Hoàng	2025	Phương pháp Monte - Carlo
3	Ước lượng chi phí xây dựng nhà xưởng trong giai đoạn đấu thầu ứng dụng mạng Neural nhân tạo (ANN)	Phong và cộng sự	2022	Mô hình ANNs
4	Revisiting storey enclosure method for early estimation of structural building construction cost	Dang và Long	2019	Phương pháp bootstrap phi tham số
5	Nghiên cứu mô hình tối ưu hóa lợi nhuận của nhà thầu xây dựng trong triển khai thi công các dự án nhà cao tầng	Đức và cộng sự	2019	Thuật toán tiến hóa vi phân, mô phỏng Monte Carlo
6	Xác định các yếu tố gây vượt chi phí thi công các dự án nhà cao tầng xảy ra tại các thầu ở Việt Nam	Tự và cộng sự	2023	Điều tra xã hội học
7	Phát triển chương trình ứng dụng mô hình thông tin (BIM) trong việc tự động hóa lập dự toán công trình xây dựng	Phước và cộng sự	2019	BIM 5D

CHƯƠNG 2: CÁC PHƯƠNG PHÁP DỰ BÁO CHI PHÍ XÂY DỰNG

Chương này trình bày các nhóm phương pháp ước tính chi phí xây dựng, gồm: phương pháp số học, phương pháp học máy đơn lẻ, và phương pháp học máy tích hợp.

2.1. PHƯƠNG PHÁP SỐ HỌC

Phương pháp số học gồm phương pháp diện tích sàn, phương pháp thể tích, phương pháp đơn vị, phương pháp bao che tầng, phương pháp phân tích phần tử, phương pháp ước lượng thừa số, và phương pháp bóc tách khối lượng.

2.1.1. Phương pháp diện tích sàn

Đây là phương pháp ước tính chi phí sơ bộ phổ biến dựa trên dựa trên chi phí xây dựng trung bình cho mỗi mét vuông diện tích sàn xây dựng và tổng diện tích sàn của công trình. Tổng diện tích sàn bằng tổng diện tích của tất cả các sàn của các tầng (tính diện tích lọt lòng, không trừ diện tích tường ngăn bên trong, thang máy và cầu thang). Phương pháp thường được áp dụng ở giai đoạn đầu của thiết kế và lập kế hoạch, khi đã xác định được quy mô cơ bản của công trình nhưng chưa có bản vẽ chi tiết hoàn chỉnh.

$$\mathbf{CPXD = Ssàn \times Cđv} \quad (2.1)$$

Trong đó:

Ssàn: Tổng diện tích sàn (m^2)

Cđv: Chi phí đơn vị theo m^2 sàn

* Ưu điểm:

- Dễ áp dụng ngay cả khi thông tin còn hạn chế;
- Hữu ích cho việc so sánh nhanh giữa các công trình có tính chất tương đồng;
- Giúp chủ đầu tư hình dung được chi phí phát sinh khi diện tích hoặc quy mô công trình thay đổi.

* Nhược điểm:

- Không phản ánh được các yếu tố phức tạp về thiết kế hoặc mức độ hoàn thiện;
- Độ chính xác thấp đối với các công trình đặc thù hoặc phi tiêu chuẩn;
- Thường bỏ qua các yếu tố đặc thù của địa điểm xây dựng.

2.1.2. Phương pháp thể tích

Tương tự phương pháp diện tích sàn, phương pháp thể tích ước tính chi phí xây dựng sơ bộ thông qua đơn vị khối tích. Phương pháp này dựa trên việc nhân thể tích xây dựng của công trình với chi phí xây dựng trung bình tính cho 1 m^3 thể tích.

CPXD = V x Cđvtt (2.2)

Trong đó:

V: Thể tích công trình (m^3)

Cđvtt: Chi phí đơn vị thể tích ($1m^3$)

2.1.3. Phương pháp đơn vị

Còn gọi là phương pháp tính đơn giá chi phí đến từng đơn vị sử dụng của công trình, phương pháp này có thể áp dụng cho giai đoạn ước tính và dự toán, tùy thuộc vào độ chi tiết của khối lượng. Đối với khái toán chi phí, phương pháp giá được áp dụng cho các trường hợp như: trường học (chi phí/học sinh hay chi phí/$1m^2$ sàn xây dựng), bệnh viện (chi phí/giường bệnh), nhà hát (chi phí/ghế), bãi đậu xe (chi phí/chỗ đậu xe),…Đối với lập dự án theo đơn giá tổng hợp, phương pháp được áp dụng để lập giá dự thầu.

CPXD = Psd x Cđv (2.3)

Trong đó:

Psd: Năng lực sử dụng (còn gọi là năng lực thiết kế trong các báo cáo thống kê Việt Nam)

Cđv: Chi phí đơn vị

* Ưu điểm:

- Chính xác hơn so với phương pháp tính theo mét vuông;
- Cho phép kiểm soát chi phí chi tiết ở cấp độ từng hạng mục/cấu kiện;
- Dễ dàng điều chỉnh khi giá vật liệu hoặc nhân công thay đổi.

* Nhược điểm:

- Cần cập nhật thường xuyên các đơn giá để đảm bảo tính chính xác;
- Sự khác biệt về nhân công và giá cả theo vùng miền có thể làm sai lệch kết quả;
- Có thể mất nhiều thời gian nếu không có phần mềm hỗ trợ.

2.1.4. Phương pháp bao che tầng

Phương pháp bao che tầng (Story Enclosure Method - SEM) là một phương pháp ước tính chi phí xây dựng sơ bộ, được phát triển để cải thiện các phương pháp đơn giản hơn như phương pháp diện tích sàn và phương pháp thể tích. Phương pháp này đo lường các thành phần bao che của tòa nhà (sàn, tường và mái) và áp dụng hệ số điều chỉnh để phản ánh những đặc điểm thiết kế như chiều cao tầng, tầng hầm hay hình dạng mặt bằng.

$$CPXD = \left(\sum_{i=0}^{n} (2 + 0.15i) f_i + \sum_{i=0}^{n} p_i s_i + (2 \div 3) \sum_{j=0}^{m} f'_j + 2 \sum_{j=0}^{m} p'_j s'_j + r \right) R \quad (2.4)$$

Trong đó:

n: số tầng trên mặt đất

f_i: diện tích sàn tầng thứ *i* (trên mặt đất) trừ trường

p_i: chu vi tường phía bên ngoài của tầng thứ i (trên mặt đất)

s_i: chiều cao tầng thứ *i* (trên mặt đất)

m: số tầng hầm

f'_j: diện tích sàn tầng hầm thứ *j* trừ tường

p'_j: chu vi tường bên ngoài của tầng hầm thứ *j*

s'_j: chiều cao tầng hầm thứ *j*

r: diện tích sàn mái

R: chi phí xây dựng theo diện tích quy đổi.

* Ưu điểm:

- Có khả năng phản ánh sự khác biệt về hình dạng mặt bằng, tổng diện tích sàn, vị trí tầng trong chiều cao công trình, chiều cao tầng, cũng như chi phí bổ sung để tạo diện tích sử dụng ở phần ngầm;
- Cho phép đưa ra một mức đơn giá tổng hợp duy nhất;
- Có thể bổ sung chi phí cho các hạng mục ngoại vi.

* Nhược điểm:

- Dữ liệu lịch sử phục vụ so sánh, hiệu chỉnh không có sẵn;
- Ít hữu ích khi cần đáp ứng các yêu cầu chi tiết của chủ đầu tư và kiến trúc sư;
- Khó đánh giá ảnh hưởng khi thay đổi các thông số kỹ thuật;
- Chưa có phương pháp đo lường thống nhất, chuẩn hóa.

2.1.5. Phương pháp phân tích phần tử

Phương pháp này áp dụng phân tích kết quả chi phí phần tử của những dự án tương tự đã thực hiện trước đó làm cơ sở cho ước lượng chi phí. Theo đó, chi phí được tính toán dựa vào một diện tích bề mặt hoặc một diện tích sàn cơ sở nhưng chi phí đơn vị bề mặt toàn bộ thì được phân chia thành các phần tử chính và những phần tử phụ. Tại mức thấp hơn của sự phân chia, nó trở nên dễ hiệu chỉnh cho các sự khác biệt trong thiết kế của các dự án mới như là sự so sánh với các dự án cũ mà dữ liệu là có sẵn.

$$C_1 = \frac{QF_1}{QF_0} \times C_0$$
$$QF = \frac{n_{dv}}{S} \quad (2.5)$$

Trong đó:

C_1: Chí phí/m^2 công trình mới

C_o: Chí phí/m^2 công trình hiện hữu

QF_1: Thừa số khối lượng công trình mới

QF_o: Thừa số khối lượng công trình hiện hữu

n_{dv}: Số đơn vị phần tử của công trình

S: Diện tích sàn

Thực hiện bước làm như trên cho mỗi loại phần tử và tổng chi phí của dự án là tổng tất cả các chi phí của các loại phần tử.

2.1.6. Phương pháp ước lượng thừa số

Phương pháp này ưu tiên áp dụng cho các dự án với những thành phần chi phí nổi trội như nhà máy lọc dầu, nhà máy tinh chế kim loại,… Các thừa số được tính cho mỗi thành phần như là hàm số của chi phí nổi trội (predominant cost). Thông thường chi phí nổi trội là chi phí mua sắm thiết bị cho dự án. Người ta xem dự án mới sẽ có tỷ lệ giữa từng chi phí thành phần và chi phí nổi trội giống như dự án hiện hữu. Sử dụng dữ liệu của các dự án hiện hữu tương tự sẽ ước lượng sơ bộ được chi phí của một dự án công nghiệp khá nhanh với độ chính xác chấp nhận được.

$$\mathbf{CPXD_{TP} = k_{hh} \times C_m} \quad (2.6)$$

Trong đó:

$CPXD_{TP}$: chi phí xây dựng thành phần

k_{hh} : thừa số tương ứng của dự án hiện hữu

C_m : chi phí mua sắm thiết bị của dự án mới

* Ưu điểm:

- Có thể lặp lại và điều chỉnh;
- Hiệu quả trong việc xử lý các dự án lớn.

* Nhược điểm:

- Yêu cầu dữ liệu lịch sử chất lượng;
- Không phù hợp với các dự án có tính đặc thù, mới mẻ.

2.1.7. Phương pháp bóc tách khối lượng

Còn được gọi ước tính chi tiết theo bóc tách khối lượng hay dự toán chi tiết, phương pháp này bao gồm việc đo bóc và tính giá cho từng vật liệu, nhân công và hạng mục trong dự án. Nó dựa trên bản vẽ thi công, hồ sơ thiết kế kỹ thuật và dữ liệu thị trường hiện hành. Phương pháp này thường được áp dụng ở giai đoạn cuối của quá trình đấu thầu và đàm phán hợp đồng, khi toàn bộ tài liệu thiết kế và thông số kỹ thuật đã đầy đủ.

* Ưu điểm:

- Độ chính xác cao nhất trong tất cả các phương pháp ước tính;
- Cung cấp tính minh bạch trong lập ngân sách và kiểm soát chi phí;
- Rất phù hợp cho báo giá của nhà thầu, hồ sơ dự thầu, và khi cần sự phê duyệt cuối cùng của chủ đầu tư.

* Nhược điểm:

- Tốn nhiều thời gian và nguồn lực để thực hiện;
- Yêu cầu trình độ chuyên môn cao và dữ liệu thị trường luôn được cập nhật;
- Sai sót nhỏ trong bóc tách hoặc đơn giá có thể dẫn đến rủi ro lớn về ngân sách.

Nhận xét chung:

- Các phương pháp truyền thống như ước tính theo diện tích sàn, thể tích, đơn vị, hay ước lượng thừa số đóng vai trò quan trọng trong giai đoạn đầu của thiết kế và lập kế hoạch. Điểm mạnh nổi bật của chúng là dễ áp dụng, ít tốn dữ liệu, hỗ trợ chủ đầu tư trong việc ra quyết định và có thể được sử dụng như một cơ sở tham chiếu để điều chỉnh cho các dự toán chi tiết hơn về sau;
- Ở giai đoạn hồ sơ thiết kế đã hoàn thiện, phương pháp bóc tách khối lượng được xem là toàn diện và đáng tin cậy nhất, điều này hoàn toàn dễ hiểu do mức độ thông tin đầy đủ mà nó dựa vào. Tuy nhiên, ước tính chi phí truyền thống sử dụng bản vẽ và thông số kỹ thuật rất tốn thời gian. Thách thức đặt ra đối với nhà thầu xây dựng là cần có một phương pháp ước tính chi phí ngay từ giai đoạn thiết kế sơ bộ nhưng vẫn đạt độ chính xác chấp nhận được. Trong bối cảnh đó, sự phát triển của trí tuệ nhân tạo, đặc biệt là các phương pháp học máy, đã mở ra hướng đi mới nhằm khắc phục những hạn chế của các phương pháp truyền thống.

2.2. PHƯƠNG PHÁP HỌC MÁY ĐƠN LẺ

Trong các mô hình học máy đơn, mạng nơ-ron nhân tạo (Artificial Neural Networks – ANNs), máy véc tơ hỗ trợ hồi quy (Support Vector Regression – SVR) và cây phân loại và hồi quy (Classification and Regression Tree – CART) đại diện cho ba hướng tiếp cận khác nhau trong lĩnh vực học máy. ANNs có khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp thông qua cấu trúc mạng nhiều lớp, phù hợp với bản chất đa yếu tố và không tuyến tính của chi phí xây dựng. SVR dựa trên nguyên lý tối ưu hóa cấu trúc và sử dụng hàm nhân để xử lý các mối quan hệ phi tuyến trong không gian đặc trưng, cho phép đạt được hiệu suất dự báo ổn định ngay cả khi quy mô dữ liệu hạn chế. Trong khi đó, CART xây dựng mô hình dựa trên các quy tắc phân tách dạng cây, giúp diễn giải rõ ràng ảnh hưởng của các biến đầu vào và hỗ trợ phân tích cấu trúc chi phí. Việc lựa chọn ba mô hình này cho phép đánh giá và so sánh toàn diện hiệu quả dự báo của các phương pháp học máy đơn trong bài toán dự báo chi phí xây dựng.

2.2.1. Mạng nơ-ron nhân tạo

Mạng nơ-ron nhân tạo (Artificial Neural Networks- ANNs) là mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, được giới thiệu lần đầu tiên vào năm 1943 bởi hai nhà nghiên cứu McCulloch và Pitts [27]. ANNs bao gồm số lượng lớn các nơ-ron được gắn kết để xử lý thông tin nên có thể xử lý hiệu quả các bài toán phi tuyến dựa trên cơ chế xấp xỉ hàm tùy ý 'học' được từ các dữ liệu quan sát. Trong các dạng ANNs, mô hình mạng nhiều tầng truyền thẳng (Multilayer perceptron - MLP) là một mạng nơ-ron truyền tới (feedforward) kết nối các dữ liệu đầu vào và cho ra tập dữ liệu đầu ra. MLP bao gồm một lớp vào (input layer) chứa các nút cảm biến, một hoặc nhiều lớp ẩn (hidden layer) chứa các các nút tính toán, và một lớp ra (output layer) chứa một nút tính toán.

Một trong những thuật toán phổ biến để huấn luyện mạng nơ-ron MLP là thuật toán lan truyền ngược (Backpropagation Algorithm - BP). Thuật toán này điều chỉnh các trọng số liên kết và thành phần lỗi độc lập (bias value) trong quá trình huấn luyện. Hình 3.1. minh họa cấu trúc của một mô hình ANNs. Công thức 1 mô tả sự kích hoạt của một nơ-ron trong lớp ẩn như sau

$$net_j = \sum w_{ji} x_i \text{ and } y_j = f(net_j) \tag{2.7}$$

Trong đó, net_j thể hiện sự kích hoạt của nơ-ron thứ j, i thể hiện các nơ-ron trong lớp trước, w_{ji} thể hiện trọng số liên kết giữa nơ-ron j và i, y_j thể hiện hàm chuyển đổi sigmoid còn gọi là logistic.

Với $f(net_j) = \frac{1}{1+e^{\lambda net_j}}$ (2.8)

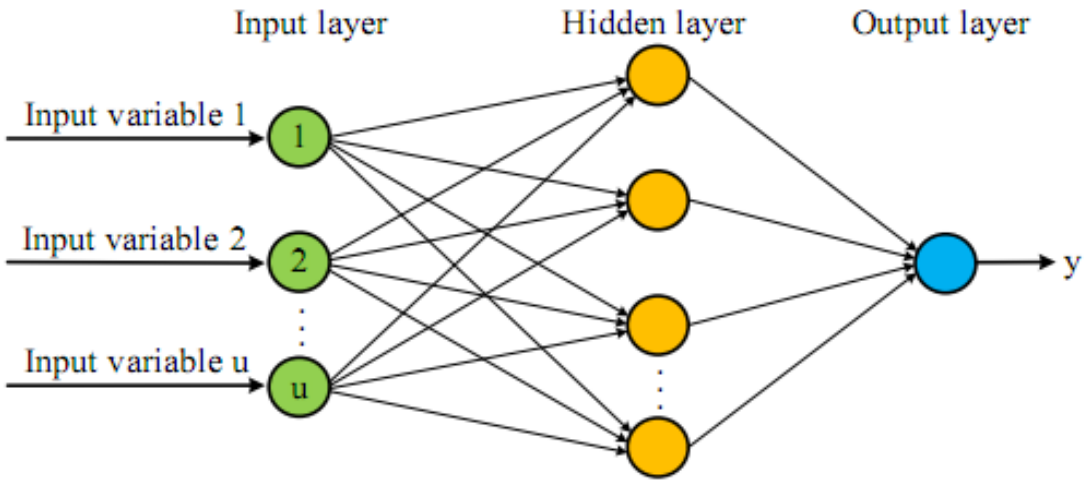
Trong đó, hệ số λ kiểm soát chức năng của hàm chuyển đổi. Công thức để huấn luyện và cập nhật các trọng số kết nối w_{ji} trong mỗi vòng h được xác định bởi

$$w_{ji}(h) = w_{ji}(h-1) + \Delta_{ji}(h) \quad (2.9)$$

Với $\Delta_{ji}(h)$ là độ lệch.

$$\Delta_{ji}(h) = \eta\delta_{pi}\chi_{pi} + \alpha\Delta w_{ji}(h-1) \quad (2.10)$$

Trong đó, η là tham số tỉ trọng học máy, δ_p là sai số lan truyền, χ_{pi} là nơ-ron đầu ra i của record p, α là tham số mô-men và $\Delta w_{ji}(h\text{-}1)$ là chênh lệch của trọng số w_{ji} trong vòng trước.



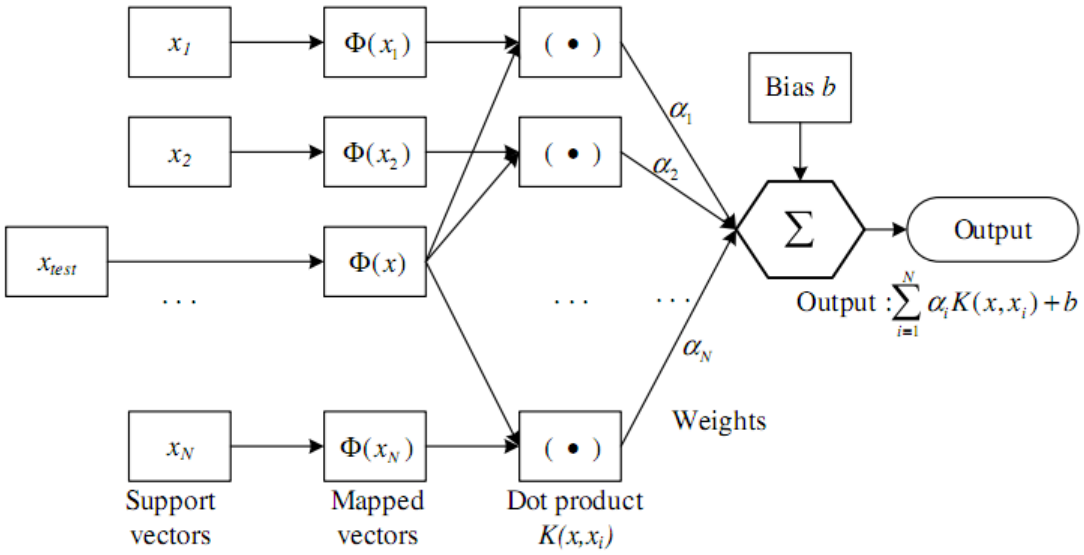
Hình 2.1. Cấu trúc của một mô hình ANNs.

2.2.2. Máy véc-tơ hỗ trợ hồi quy

Máy véc tơ hỗ trợ (support vector machines – SVMs) là một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. Thuật toán SVMs được phát triển bởi Vapnik (1995) [28], ban đầu là một thuật toán dạng nhị phân. SVMs xây dựng một siêu phẳng hoặc một tập hợp các siêu phẳng trong một không gian nhiều chiều hoặc vô hạn chiều. Để phân loại tốt nhất thì các siêu phẳng nằm ở càng xa

các điểm dữ liệu của tất cả các lớp (hàm lề) càng tốt, vì lề càng lớn thì sai số tổng quát hóa của thuật toán phân loại càng bé.

Để sử dụng cho mục đích hồi quy, máy véc tơ hỗ trợ hồi quy ra đời (Support vector regresion - SVR). SVR tìm cách tối thiểu giới hạn trên của sai số tổng quát hóa thay vì tối thiểu sai số thực nghiệm như mô hình mạng nơ-ron. Hình 2.2 mô tả cấu trúc của một máy học véc-tơ hồi quy điển hình.



Hình 2.2. Cấu trúc điển hình của máy học véc-tơ hồi quy.

SVR tìm một siêu phẳng đi qua tất cả các điểm dữ liệu với độ lệch chuẩn ε nhằm tìm một hàm $f(x)$ có sai số nhỏ nhất ε so với y_i.

$$f(x) = \omega\Phi(x) + b \tag{2.11}$$

Trong đó, ω là véc tơ trọng số của hàm tuyến tính, $\omega \subset R^M$; $\Phi(x)$ biểu thị một hàm phi tuyến được chuyển từ không gian R^M vào không gian nhiều chiều; b là độ dịch.

Từ đó dẫn đến bài toán tối ưu hóa như sau

$$\min \Phi(\omega, b, \xi, \xi^*) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.12}$$

Với điều kiện $y_i - f(x_i, \omega) \le \varepsilon + \xi_i^*;\ f(x_i, \omega) - y_i \le \varepsilon + \xi_i^*;\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots, n$

Trong đó, $C \ge 0$ là hằng số chuẩn hóa thể hiện sự cân bằng giữa sai số thực nghiệm và độ phẳng của hàm $f(x_i)$; ξ, ξ^* là các biến bù không âm; x_i là các đặc tính đầu vào; y_i là các nhãn dự báo liên quan đến x_i; n là kích thước dữ liệu.

Hàm tối ưu 2.12 có thể được chuyển đổi thành hàm đối ngẫu như sau

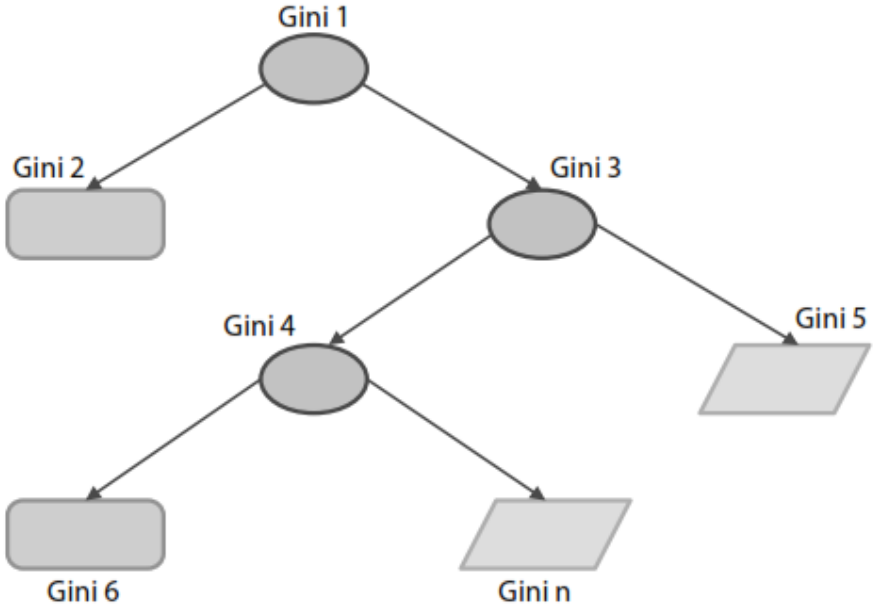
$$f(x) = \sum_{i=1}^{n_{SV}} (\alpha_i - \alpha_i^*) K(x, x_i) \quad (2.13)$$

Với điều kiện $0 \le \alpha_i^* \le C;\ 0 \le \alpha_i \le C$

Trong đó, α_i , α_i^* là các hệ số Lagrange, n_{sv} là các điểm support vectors và $K(x, x_i)$ là hàm nhân (hàm kernel). Trong quá trình huấn luyện, hàm nhân được sử dụng để nhận dạng các support vector dọc theo bề mặt hàm số. Trong số các hàm nhân được sử trong không gian phi tuyến nhiều chiều, hàm radial basis (RBF) cho kết quả tốt hơn cả [29].

2.2.3. Cây phân loại và hồi quy

Được giới thiệu bởi Breiman et al. [30], cây phân loại và hồi quy (Classification and regression tree – CART) đề cập đến các thuật toán cây quyết định (DT) có thể sử dụng cho vấn đề phân loại hoặc hồi quy. Cây phân loại (classification tree) được thiết kế cho các biến phụ thuộc có giá trị là tên thể loại với sai số dự báo được đo bằng giá trị của phân lớp dự báo sai lệch. Cây hồi quy (regression tree) được sử dụng cho các biến phụ thuộc có giá trị là số thực với sai số dự báo được đo bằng sự chênh lệch bình phương giữa các giá trị quan sát được và giá trị dự báo [31]. Ưu điểm nổi bật của CART là mô hình được thiết kế dễ hiểu và có thể xử lý tốt một lượng lớn dữ liệu trong thời gian ngắn. Hình 2.3 minh họa cấu trúc của một mô hình CART điển hình.



Hình 2.3. Cấu trúc của mô hình CART.

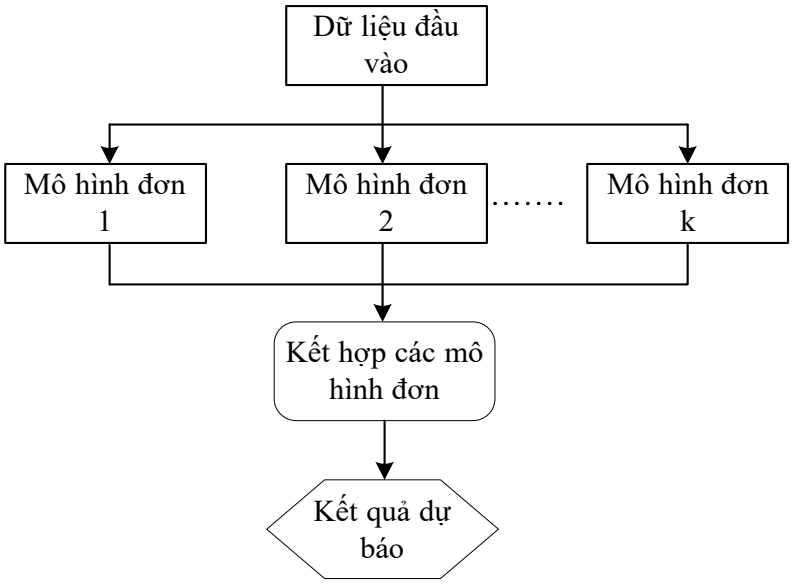
2.3. PHƯƠNG PHÁP TÍCH HỢP

Khi nhu cầu dự báo chi phí ngày càng đòi hỏi độ chính xác cao hơn ở giai đoạn thiết kế sớm, các phương pháp học máy đơn lẻ như ANNs, SVR hay CART đã chứng minh được ưu thế nhờ khả năng xử lý dữ liệu đa chiều và nhận diện quan hệ phi tuyến giữa các biến số. Tuy nhiên, nhược điểm chung của các mô hình này là độ ổn định và tính khái quát hóa vẫn còn hạn chế khi áp dụng cho các bộ dữ liệu đa dạng trong thực tế. Trong bối cảnh đó, các mô hình học máy tích hợp (ensemble learning) ra đời như một giải pháp tiềm năng, với nguyên lý kết hợp nhiều mô hình đơn lẻ để khai thác ưu điểm và bù đắp nhược điểm của từng mô hình. Những kỹ thuật như Voting, Bagging hay Stacking không chỉ cải thiện độ chính xác và tính ổn định trong dự báo mà còn làm tăng khả năng ứng dụng rộng rãi trong các dự án xây dựng có quy mô và mức độ phức tạp khác nhau. Chính vì vậy, phương pháp tích hợp đang dần trở thành xu hướng mới trong ước tính chi phí xây dựng, thay thế dần sự phụ thuộc vào các phương pháp truyền thống vốn chỉ mang tính tham chiếu.

2.3.1. Voting

Mô hình Voting (bỏ phiếu) sử dụng sức mạnh của nhiều mô hình đơn lẻ để tạo ra một kết quả dự đoán tốt hơn [32]. Các mô hình đơn lẻ ở đây chính là ANNs, SVR, và CART. Hình 2.4 mô tả cấu trúc của mô hình Voting, theo đó các mô hình cơ sở lần lượt sử dụng dữ liệu đầu vào để dự đoán kết quả. Kết quả dự đoán cuối cùng dựa trên giá trị trung bình các giá trị đầu ra của các mô hình đơn lẻ. Voting có ưu điểm là dễ triển khai và giúp giảm sai số ngẫu nhiên của từng mô hình riêng lẻ.

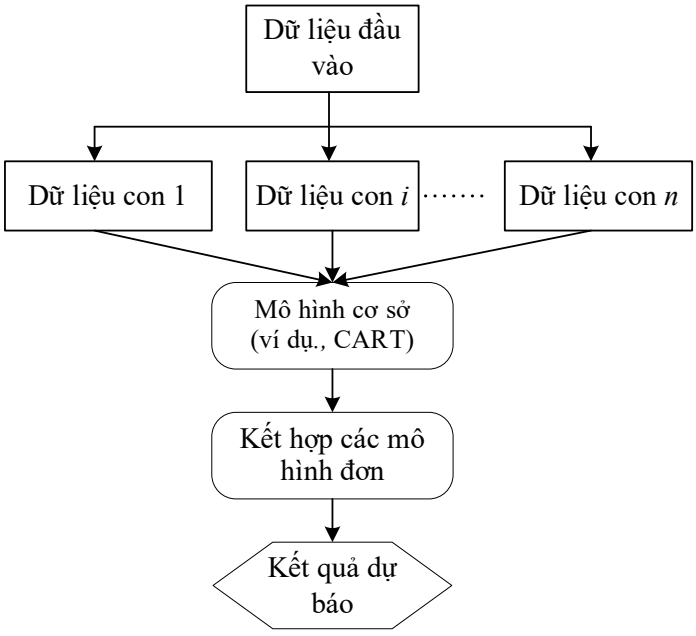
Có 4 mô hình Voting được tạo ra từ sự kết hợp của từ 2 đến 3 mô hình học máy đơn lẻ. Các mô hình Voting được tạo ra từ sự kết hợp của 2 mô hình đơn lẻ gồm: ANNs+CART, ANNs+SVR, và CART+SVR. Mô hình Voting hình thành từ 3 mô hình đơn lẻ là ANNs+CART+SVR.



Hình 2.4. Cấu trúc mô hình Voting.

2.3.2. Bagging

Mô hình Bagging (đóng gói) sao chép các mẫu dữ liệu một cách ngẫu nhiên thay thế tập dữ liệu ban đầu và mỗi mô hình hồi quy dự đoán các giá trị từ các mẫu dữ liệu một cách độc lập (Hình 2.5) [33]. Kỹ thuật này được sử dụng để xây dựng nhiều mô hình dự báo độc lập trên các tập dữ liệu con được chọn ngẫu nhiên từ tập dữ liệu gốc. Mỗi mô hình được xây dựng trên một tập dữ liệu khác nhau, do đó các mô hình có thể đưa ra dự đoán khác nhau cho cùng một tập dữ liệu đầu vào. Bagging giúp giảm phương sai (variance), cải thiện độ ổn định và tránh hiện tượng quá khớp (overfitting).

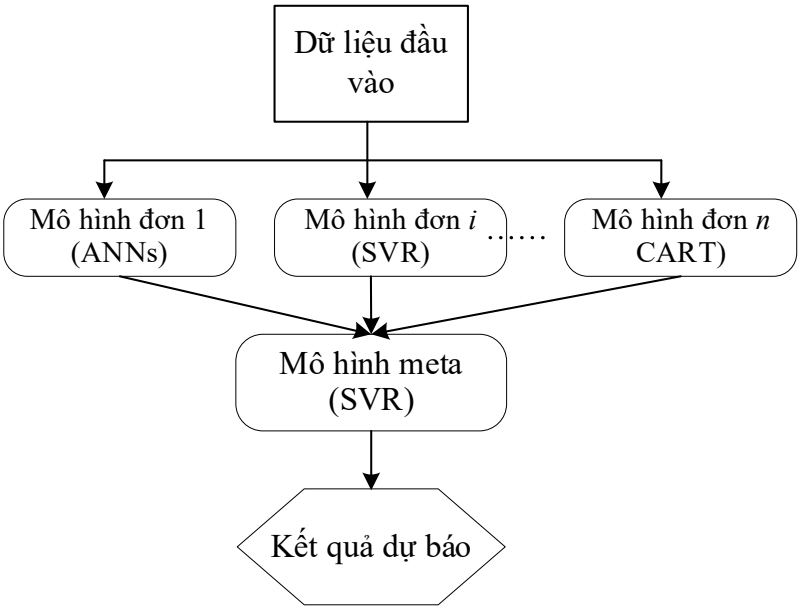


Hình 2.5. Cấu trúc mô hình Bagging.

2.3.3. Stacking

Mô hình Stacking (xếp chồng) là một phương pháp phân loại theo cấu trúc phân cấp nhiều tầng [34]. Ở tầng thứ nhất, nhiều mô hình phân loại cơ sở ((base classifiers) được huấn luyện song song, sau đó đầu ra của chúng được sử dụng làm dữ liệu đầu vào cho một mô hình phân loại ở tầng thứ hai, thường được gọi là bộ phân loại "stacked" hoặc "meta". Dự đoán cuối cùng được xác định bởi mô hình học máy ở tầng thứ 2 này, vốn đóng vai trò như một bộ kết hợp có khả năng học, điều chỉnh và sửa lỗi từ các mô hình ở tầng đầu tiên trên cùng một tập dữ liệu.

Trong nghiên cứu này, mô hình Stacking được thiết kế theo cấu trúc hai tầng: tầng thứ nhất bao gồm hai hoặc ba mô hình máy học đơn lẻ (tương tự phương pháp voting), trong khi tầng thứ hai sử dụng SVR làm meta-classifier. Hình 2.6 minh họa cấu trúc của khung stacking này. Có bốn mô hình xếp chồng bao gồm: ANNs+SVR, ANNs+CART, SVR+CART, và ANNs+SVR+CART.



Hình 2.6. Cấu trúc mô hình Stacking.

Nhận xét:

So với các phương pháp truyền thống (số học), các phương pháp học máy đơn lẻ mang đến một cách tiếp cận hoàn toàn khác. Mạng nơ-ron nhân tạo có khả năng mô hình hóa các quan hệ phi tuyến tính phức tạp, giúp nắm bắt được những yếu tố mà phương pháp truyền thống khó định lượng. Chúng thường được coi là "hộp đen", khó giải thích và yêu cầu dữ liệu lớn để huấn luyện. Máy vector hỗ trợ lại có ưu thế về khả năng dự

báo với dữ liệu nhỏ và tính ổn định cao, nhưng nhược điểm là hiệu năng giảm khi xử lý tập dữ liệu rất lớn hoặc có nhiều nhiễu. Cây hồi quy và phân loại (CART) cung cấp kết quả trực quan, dễ hiểu, và có thể xử lý các đặc trưng phức tạp; tuy nhiên, chúng dễ bị quá khớp (overfitting) nếu không được điều chỉnh hợp lý. Như vậy, so với các phương pháp truyền thống, các kỹ thuật học máy đơn tuy đòi hỏi dữ liệu và tính toán cao hơn, nhưng lại cho phép dự báo chính xác và linh hoạt hơn ở giai đoạn thiết kế sớm, nơi mà nhu cầu ước tính nhanh và chính xác là rất quan trọng.

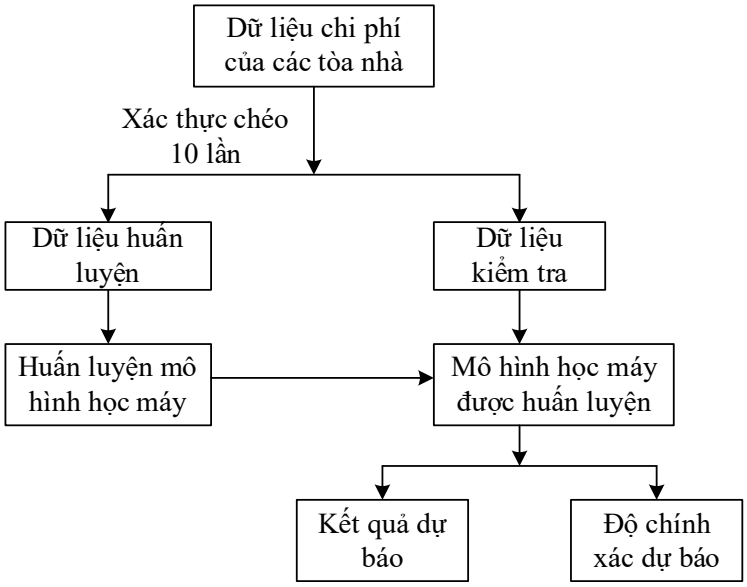
Tuy nhiên, nhược điểm chung của các mô hình học máy đơn là độ ổn định và tính khái quát hóa vẫn còn hạn chế khi áp dụng cho các bộ dữ liệu đa dạng trong thực tế. Trong bối cảnh đó, các mô hình học máy tích hợp ra đời như một giải pháp tiềm năng, với nguyên lý kết hợp nhiều mô hình đơn lẻ để khai thác ưu điểm và bù đắp nhược điểm của từng mô hình. Những kỹ thuật như Voting, Bagging hay Stacking không chỉ cải thiện độ chính xác và tính ổn định trong dự báo mà còn làm tăng khả năng ứng dụng rộng rãi trong các dự án xây dựng có quy mô và mức độ phức tạp khác nhau. Chính vì vậy, phương pháp tích hợp đang dần trở thành xu hướng mới trong ước tính chi phí xây dựng, thay thế dần sự phụ thuộc vào các phương pháp truyền thống vốn chỉ mang tính tham chiếu.

CHƯƠNG 3: DỰ BÁO CHI PHÍ XÂY DỰNG NHÀ Ở CAO TẦNG BẰNG MÔ HÌNH TÍCH HỢP DỰA TRÊN HỌC MÁY

3.1. QUY TRÌNH XÂY DỰNG MÔ HÌNH NGHIÊN CỨU

Dữ liệu ban đầu được chia thành 2 tập dữ liệu dùng để huấn luyện mô hình (training data) và kiểm tra mô hình (test data). Để tránh sự 'thiên vị' trong quá trình lựa chọn 2 tập dữ liệu trên, phương pháp xác thực chéo *k* lần (K-fold cross validation) được sử dụng. Phương pháp này phân chia toàn bộ dữ liệu thành *k* tập con có cùng kích thước. Quá trình huấn luyện cho mỗi mô hình có *k* lần. Trong mỗi lần, một tập con được dùng để kiểm tra và (*k*-1) tập còn lại dùng để huấn luyện. Kohavi (1995) chỉ ra rằng *k*=10 là tối ưu [35]; do đó, xác thực chéo 10 lần được sử dụng nhằm đánh giá khả năng dự báo tổng thể của các mô hình học máy. Hình 3.1 mô tả về phương pháp xác thực chéo 10 lần. Sơ đồ huấn luyện và kiểm tra các mô hình học máy được thể hiện ở Hình 3.2.

Hình 3.1. Phương pháp xác thực chéo 10 lần.



Hình 3.2. Sơ đồ huấn luyện và kiểm tra các mô hình.

3.2. THIẾT LẬP THÔNG SỐ CÁC MÔ HÌNH HỌC MÁY

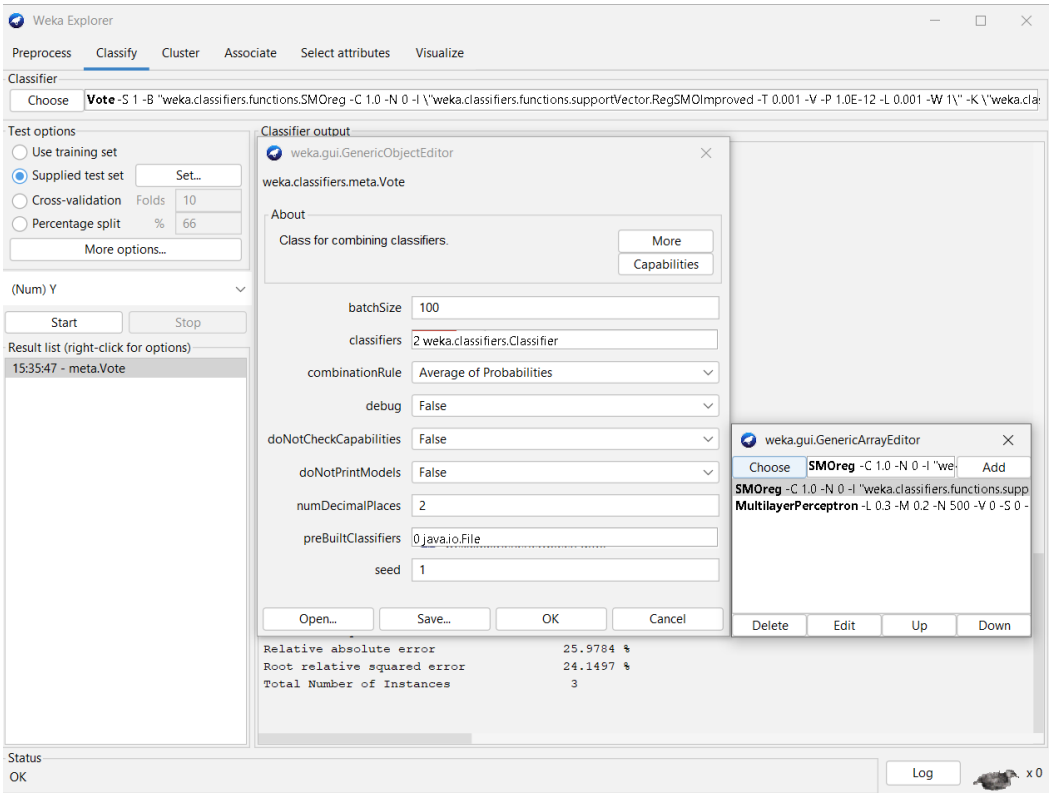
Các mô hình học máy trong nghiên cứu này được chạy trên phần mềm mã nguồn mở WEKA (phiên bản 3.8.6). Đây là phần mềm học máy do Đại học Waikato (New Zealand) phát triển nhằm mục đích khai phá dữ liệu. Bảng 3.1 trình bày các tham số được thiết lập cho các mô hình trên phần mềm WEKA. Các thông số của các mô hình học máy tích tích hợp Voting, Bagging, Stacking lần lượt được minh họa ở các Hình 3.3, Hình 3.4, và Hình 3.5. Các thông số này được xác định dựa trên kinh nghiệm từ các nghiên cứu trước [36, 37].

Bảng 3.1. Thiết lập tham số cho các mô hình học máy đơn lẻ.

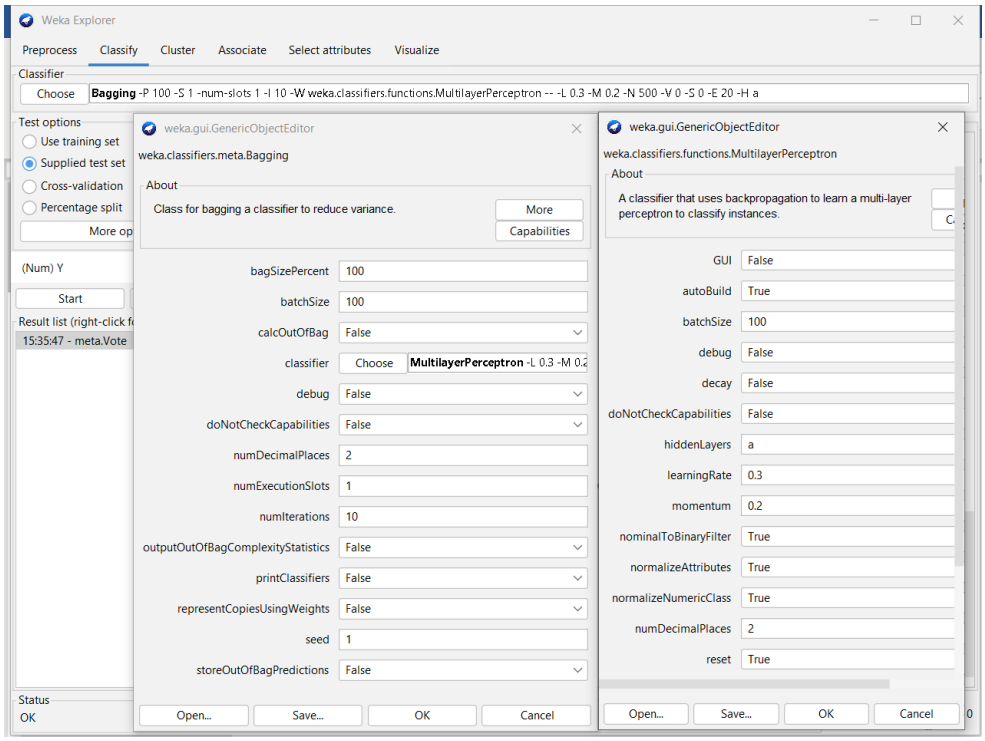
Mô hình	Tham số	Sự thiết lập
ANNs	Hidden layer	3
	Leaning rate	0.3
	Momentum	0.2
	Training/time	500
	Seed	0
SVR	*C*	1.0
	Kernel	RBF
CART	Initial count	0.0
	Max depth	-1
	MinNum	2.0
	MinVarianceProp	0.001
	NoPruning	False
	NumFolds	3

Bảng 3.2. Thiết lập tham số cho các mô hình học máy tích hợp.

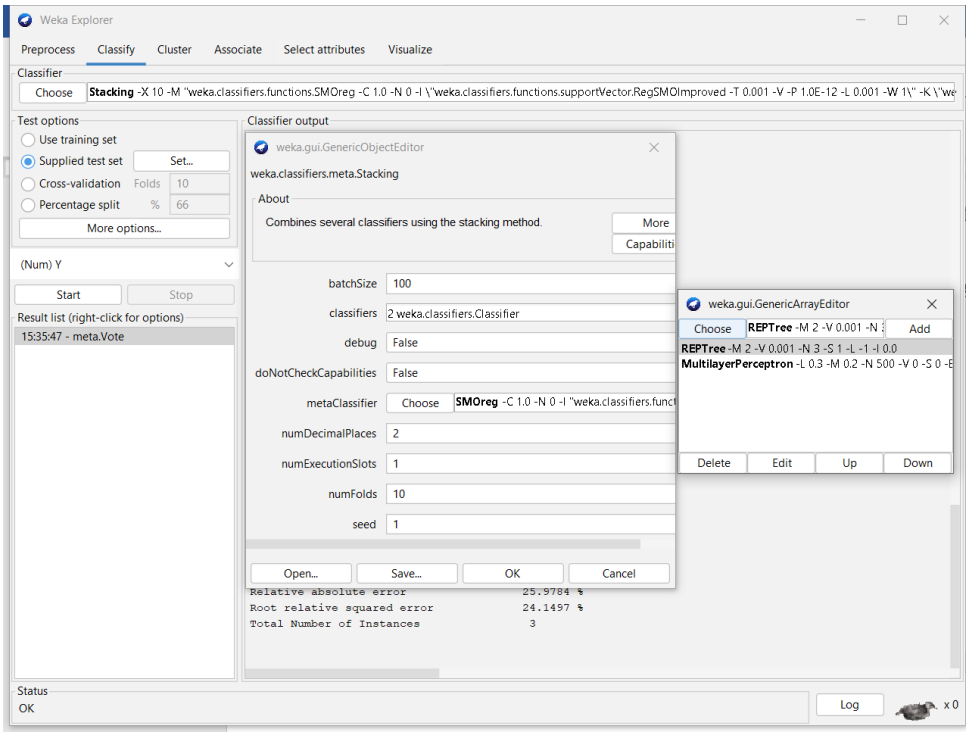
Mô hình	Tham số	Sự thiết lập
Voting	Classifiers	2−3 weka.classifiers. Classifier
	Combination Rule	Average
	Seed	1
Bagging	BatchSizePercent	100
	BatchSize	100
	Classifier	ANNs/SVR/CART
	Seed	1
Stacking	BatchSize	100
	Classifiers	2−3 weka.classifiers. Classifier
	MetaClassifier	SVR
	NumFolds	10
	Seed	1



Hình 3.3. Các thông số của mô hình Voting (ANNs+SVR).



Hình 3.4. Các thông số của Bagging (ANNs).



Hình 3.5. Các thông số của Stacking (ANNs+CART).

3.3. CÁC CHỈ SỐ ĐÁNH GIÁ ĐỘ CHÍNH XÁC DỰ BÁO

Để đánh giá độ chính xác kết quả dự đoán của các mô hình đề xuất, các số chỉ số được sử dụng, gồm: căn bậc hai của sai số bình phương trung bình (root mean square error – RMSE), sai số tuyệt đối trung bình (mean absolute error – MAE), phần trăm sai

số tuyệt đối trung bình (mean absolute percentage error – MAPE) và chỉ số xếp hạng tổng hợp (synthesis index – SI). Công thức tính toán của các chỉ số thể hiện ở bên dưới.

RMSE là một chỉ số đánh giá độ chính xác thường dùng trong phân tích hồi quy. Nó cho biết mức độ sai lệch trung bình giữa giá trị dự đoán và giá trị thực tế trong mô hình hồi quy. RMSE được xác định bằng cách lấy giá trị căn bậc hai của trung bình của tổng bình phương chênh lệch giữa giá trị thực tế và giá trị dự đoán. RMSE đo lường sự khác biệt giữa giá trị mẫu (thực tế) với giá trị dự báo bằng mô hình. RMSE luôn có giá trị không âm, và giá trị càng nhỏ thì độ chính xác dự báo của mô hình càng cao.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'-y)^2} \tag{3.1}$$

Trong đó, y là giá trị thực tế; y' là giá trị dự đoán, n là số lượng dữ liệu dự đoán.

MAE là một thước đo sai số trong các bài toán hồi quy, xem xét giá trị tuyệt đối của chênh lệch giữa dự đoán của mô hình và giá trị thực, sau đó lấy trung bình trên toàn bộ dữ liệu. MAE phản ánh sai số giữa giá trị thực tế và giá trị dự đoán mà không quan tâm đó là sai số vượt quá hay sai số thiếu hụt. Chỉ số này hữu ích khi mục tiêu là đánh giá chất lượng dự báo dựa trên độ lệch tuyệt đối, thay vì tỷ lệ hay mức độ tương đối của sai lệch. Tương tự RMSE, MAE có giá trị càng nhỏ thì độ chính xác dự báo của mô hình càng cao.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y-y'| \tag{3.2}$$

MAPE là một đại lượng thống kê phổ biến dùng để để đánh giá độ chính xác của mô hình dự đoán. MAPE đo lường độ lệch trung bình (tính theo phần trăm) giữa giá trị dự đoán và giá trị thực tế, hay nói cách khác, nó cho biết mức độ sai lệch trung bình của dự đoán. Khi so sánh độ chính xác dự đoán của các mô hình khác nhau, MAPE là chỉ số hữu ích vì nó không bị ảnh hưởng bởi kích thước mẫu và đơn vị của giá trị dự đoán. MAPE càng nhỏ chứng tỏ mô hình có độ chính xác dự đoán càng cao.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y-y'}{y}\right| \tag{3.3}$$

Chỉ số SI dùng để xếp hạng các mô hình (gồm 11 tổ hợp của mô hình tích hợp và 3 mô hình đơn). Giá trị của SI thuộc [0,1], mô hình có SI càng tiến về 0 chứng tỏ kết

quả dự đoán của mô hình đó càng chính xác. SI được xác định dựa vào 3 chỉ số thống kê trên (RMSE, MAE, và MAPE).

$$SI = \frac{1}{m}\sum_{i=1}^{m}\frac{P_i - P_{min,i}}{P_{max,i} - P_{min,i}} \quad (3.4)$$

Trong đó, m là số lượng chỉ số đánh giá (ở đây m=3), P_i là giá trị chỉ số đánh giá thứ i.

3.4. PHÂN TÍCH VÀ ĐÁNH GIÁ KẾT QUẢ

3.4.1. Thu thập và xử lý dữ liệu

Bộ dữ liệu chi phí được thu thập từ các dự án xây dựng nhà ở cao tầng trên địa bàn thành phố Hồ Chí Minh. Sau khi phân tích và xử lý các mẫu ngoại lai, thu được 32 mẫu. Bảng 3.3 thống kê các biến số trong bộ dữ liệu này. Theo đó, có 6 biến đầu vào ảnh hưởng đến chi phí xây dựng của nhà thầu (Y) gồm: tổng diện tích sàn xây dựng (X1), số tầng hầm (X2), số tầng cao (X3), loại kết cấu móng (X4), biện pháp thi công tầng hầm (X5), và thời gian thi công (X6). Trong 6 biến đầu vào, ngoài 2 biến định tính là X4 và X5, thì các biến còn lại ở dạng định lượng. Trong số các biến định lượng, X1và X6 ở dạng liên lục; còn X2 và X3 ở dạng rời rạc. Loại kết cấu móng (X4) gồm móng có đài trên cọc ép bê tông cốt thép, móng có đài trên cọc nhồi, và móng khác. Biện pháp thi công tầng hầm (X5) gồm đào hở và đào không hở.

Để huấn luyện mô hình, biến giả (dummy) được tạo ra để chuyển biến định tính thành định lượng. Theo đó, biến X4 được phân thành X41 và X42 với X41=1 tương ứng móng có đài trên cọc ép bê tông cốt thép, X42=1 tương ứng móng có đài trên cọc nhồi. Tương tự, X51=1 tương ứng biện pháp thi công tầng hầm là đào hở. Tổng hợp các biến số được sử dụng để huấn luyện mô hình được thể hiện ở Bảng 3.4. Hình 3.6 thể hiện phân phối thống kê của các biến liên tục và rời rạc trong bộ dữ liệu.

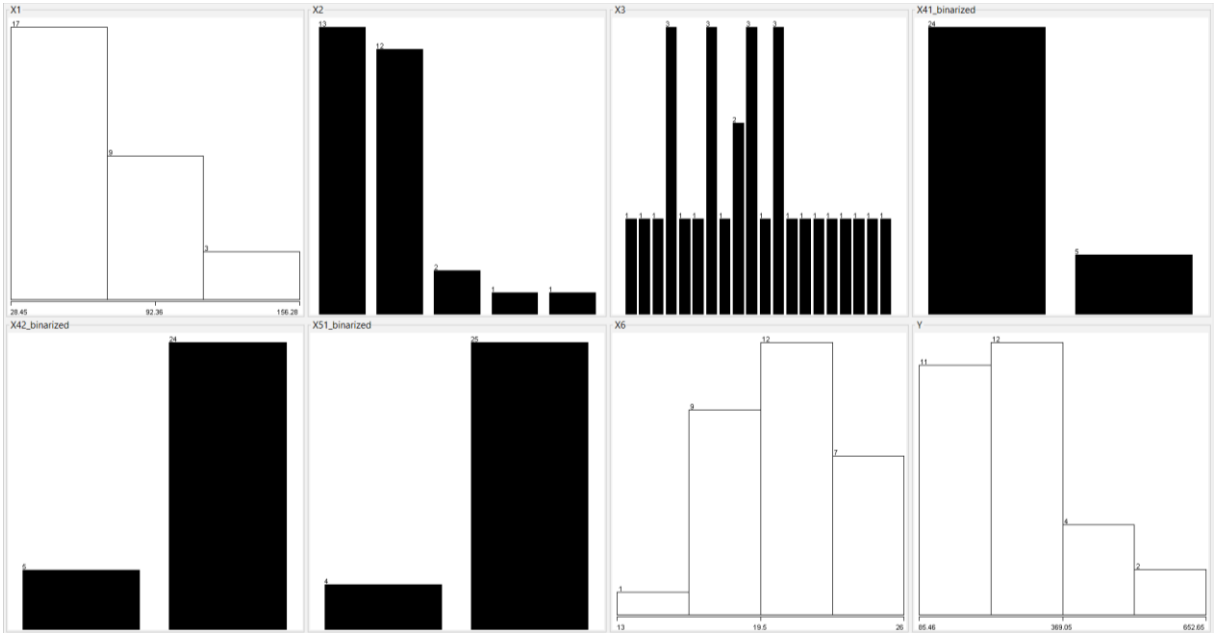
Với xác thực chéo 10 lần, bộ dữ liệu được huấn luyện và kiểm chứng 10 lần. Trong mỗi lần, 29 mẫu được sử dụng để huấn luyện mô hình, 3 mẫu còn lại được dùng để kiểm chứng mô hình. Độ chính xác dự báo của mô hình trong mỗi lần được đánh giá thông qua 3 mẫu kiểm chứng.

Bảng 3.3. Mô tả thống kê các biến số trong bộ dữ liệu.

STT	Diễn giải nội dung	Kí hiệu	Kiểu biến	Đơn vị	Giá trị nhỏ nhất	Giá trị lớn nhất
1	Tổng diện tích sàn xây dựng	X1	Định lượng	1000m2	28,447	156,281
2	Số tầng hầm	X2	Định lượng	Tầng	1	5
3	Số tầng cao (không tính tầng hầm)	X3	Định lượng	Tầng	10	50
4	Loại kết cấu móng	X4	Định tính	-	-	-
5	Biện pháp thi công tầng hầm	X5	Định tính	-	-	-
6	Thời gian thi công	X6	Định lượng	tháng	13	26
7	Chi phí xây dựng nhà thầu bỏ ra	Y	Định lượng	Tỷ đồng	85,459	652,649

Bảng 3.4. Định dạng các biến trong bộ dữ liệu.

STT	Diễn giải nội dung	Kí hiệu	Định dạng
1	Tổng diện tích sàn xây dựng	X1	Liên tục
2	Số tầng hầm	X2	Rời rạc
3	Số tầng cao (không tính tầng hầm)	X3	Rời rạc
4	Móng có đài trên cọc ép bê tông cốt thép	X41	Nhị phân
5	Móng có đài trên cọc nhồi	X42	Nhị phân
6	Biện pháp thi công hầm là đào hở	X51	Nhị phân
7	Thời gian thi công	X6	Liên tục
8	Chi phí xây dựng nhà thầu bỏ ra	Y	Liên tục



Hình 3.6. Phân phối thống kê của các biến đầu vào và đầu ra trong bộ dữ liệu.

3.4.2. Kết quả và đánh giá các mô hình đơn lẻ

Bảng 3.5. Độ chính xác dự báo của các mô hình học máy đơn lẻ.

Tên mô hình	MAE (tỷ đồng)	RMSE (tỷ đồng)	MAPE (%)
ANNs	67,524	77,025	26,70
SVR	**54,535**	**62,746**	**20,50**
CART	77,704	91,661	25,50

Độ chính xác dự báo của các mô hình học máy thể hiện qua so sánh giá trị thực tế và giá trị dự báo của tập dữ liệu kiểm chứng (Bảng 3.5). Trong số các mô hình đơn, mô hình SVR thể hiện khả năng dự báo vượt trội ở tất cả các chỉ số. Sai số tuyệt đối trung bình (MAE) của mô hình SVR là 54,535 tỷ đồng, thấp hơn nhiều so với ANNs (67,524 tỷ đồng) và CART (77,704 tỷ đồng). SVR giảm MAE khoảng 19,24% so với ANNs và 29,82% so với CART. Tương tự, căn bậc hai của sai số bình phương trung bình (RMSE) của SVR là nhỏ nhất, tiếp theo là ANNs và CART. SVR giảm RMSE 14,279 tỷ đồng (tương đương 18,54%) so với ANNs và giảm RMSE 28,915 tỷ đồng (tương đương 31,55%). Phần trăm sai số tuyệt đối trung bình (MAPE) của SVR đạt mức độ chính xác chấp nhận được với 20,5%. Trong khi đó, ANNs và CART với MAPE trên 25% cho thấy sai số dự báo vẫn còn đáng kể. Lợi thế của SVR có thể xuất phát từ cơ chế tối ưu

khoảng cách biên và khả năng xử lý quan hệ phi tuyến qua kernel. Do đó, trong ba mô hình đơn lẻ, SVR là lựa chọn tiềm năng nhất cho dự báo chi phí xây dựng nhà ở cao tầng.

3.4.3. Kết quả và đánh giá các mô hình tích hợp

Bảng 3.6. Độ chính xác dự báo của các mô hình học máy tích hợp.

Tên mô hình	MAE (tỷ đồng)	RMSE (tỷ đồng)	MAPE (%)
Mô hình Voting			
ANNs+SVR	**43,771**	**51,479**	**19,01**
ANNs+CART	56,577	67,072	22,24
SVR+CART	60,647	71,791	21,32
ANNs+SVR+CART	49,525	59,675	18,34
Mô hình Bagging			
Bagging (ANNs)	63,752	72,697	24,63
Bagging (SVR)	**55,228**	**66,325**	**20,79**
Bagging (CART)	62,183	71,177	23,03
Mô hình Stacking			
ANNs+SVR	68,543	77,298	24,01
ANNs+CART	92,603	101,402	35,88
SVR+CART	79,062	89,754	38,48
ANNs+SVR+CART	**65,749**	**73,592**	**24,21**

Bảng 3.6 tổng hợp kết quả dự báo của ba cơ chế tích hợp gồm Voting, Bagging và Stacking. Các tổ hợp theo cơ chế Voting thể hiện kết quả dự báo khá khả quan, đặc biệt khi kết hợp mô hình ANNs và mô hình SVR. Cặp mô hình này đạt MAE khoảng 43,771 tỷ đồng, RMSE khoảng 51,479 tỷ đồng và MAPE khoảng 19,01%**,** thấp nhất trong tất cả các cấu hình Voting, cho thấy sự bổ trợ lẫn nhau về năng lực dự báo. Tổ hợp ba mô hình ANNs+SVR+CART cũng cho kết quả đáng chú ý với MAPE đạt 18,34%**,** mặc dù MAE và RMSE cao hơn một chút so với ANNs+SVR. Trong khi đó, các cặp ANNs+CART và SVR+CART cho sai số dự báo cao hơn (MAPE lần lượt là 22,24% và

21,32%), chứng tỏ CART chưa mang lại hiệu quả cao khi kết hợp với một mô hình khác. Tóm lại, với cơ chế tích hợp Voting thì ANNs+SVR là phương án nổi trội nhất.

Trong nhóm Bagging, kết quả cho thấy SVR tiếp tục vượt trội hơn cả với MAE đạt 55,228 tỷ đồng, RMSE đạt 66,325 tỷ đồng và MAPE khoảng 20,79%**,** tốt hơn so với ANNs và CART. ANNs có sai số cao hơn với MAPE khoảng 24,63%, trong khi CART đạt MAPE khoảng 23,03%. Như vậy, cơ chế Bagging dựa trên SVR thể hiện khả năng dự báo tốt nhất trong nhóm Bagging, tiếp theo là dựa trên CART và ANNs.

Kết quả của các mô hình Stacking không đạt được sự cải thiện rõ rệt như Voting. Các tổ hợp ANNs+SVR, ANNs+CART và SVR+CART đều cho sai số cao (MAPE lần lượt là 24,01%, 35,88% và 38,48%), trong đó ANNs+CART và SVR+CART thể hiện mức sai số cao, phản ánh sự kết hợp chưa tối ưu. Tổ hợp ba mô hình ANNs+SVR+CART cho kết quả khả quan hơn (MAPE = 24,21%), song vẫn kém xa so với Voting.

Bảng 3.7 sử dụng chỉ số SI như một thước đo tổng hợp để đánh giá khả năng dự báo của tất cả các mô hình, trong đó giá trị SI càng thấp thì hiệu quả dự báo càng tốt. Kết quả xếp hạng cho thấy sự khác biệt rõ rệt giữa các nhóm mô hình đơn lẻ và mô hình tích hợp. Trong nhóm mô hình đơn lẻ, SVR nổi bật nhất với giá trị SI là 0,184 và xếp hạng 3, trong khi ANNs và CART có SI lần lượt là 0,471 (xếp hạng 11) và 0,618 (xếp hạng 12). Điều này khẳng định SVR là mô hình mạnh nhất trong ba mô hình đơn.

Trong ba cơ chế tích hợp, nhóm Voting cho kết quả khả quan nhất. Cặp ANNs+SVR đạt SI bằng 0,011, thấp nhất trong toàn bộ các mô hình và được xếp hạng 1, cho thấy đây là cấu hình dự báo tối ưu. Tổ hợp ba mô hình ANNs+SVR+CART xếp thứ 2 với SI = 0,094. Các tổ hợp Voting khác như ANNs+CART và SVR+CART có SI cao hơn (0,256 và 0,3), lần lượt xếp hạng 5 và 6, nhưng vẫn tốt hơn nhiều so với các mô hình đơn lẻ như ANNs hoặc CART.

Nhóm mô hình Bagging duy trì độ chính xác dự báo khá ổn định. Trong đó, SVR có SI đạt 0,218 và xếp hạng 4, thể hiện rằng phương pháp này giúp giữ vững hiệu năng của SVR. Ngược lại, ANNs và CART có SI cao hơn (lần lượt là 0,382 và 0,335), tương ứng các vị trí 8 và 7, cho thấy hiệu quả cải thiện còn hạn chế.

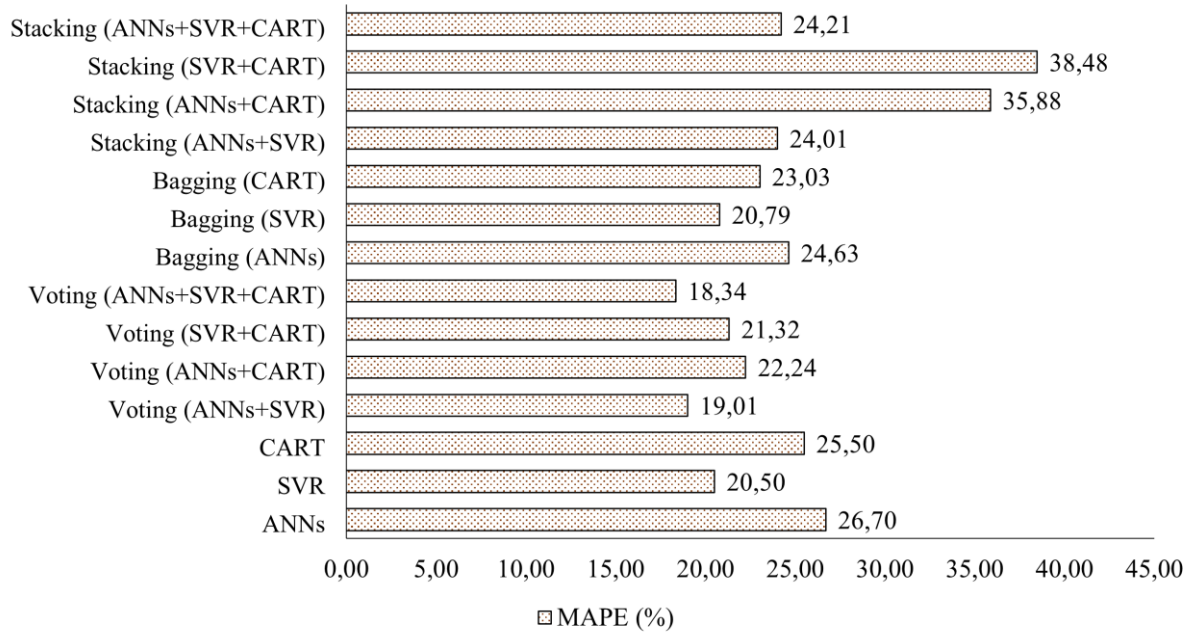
Nhóm mô hình Stacking nhìn chung không đem lại kết quả tích cực. Các mô hình ANNs+SVR, ANNs+CART và SVR+CART đều có SI cao (lần lượt là 0,435; 0,957;

0,830), lần lượt xếp hạng 10, 14 và 13. Mặc dù tổ hợp ANNs+SVR+CART có kết quả tốt hơn (SI = 0,395, xếp vị trí 9), nhưng vẫn kém xa so với các mô hình Voting và thậm chí một số mô hình Bagging. Điều này cho thấy phương pháp Stacking chưa phù hợp trong bối cảnh dữ liệu nghiên cứu.

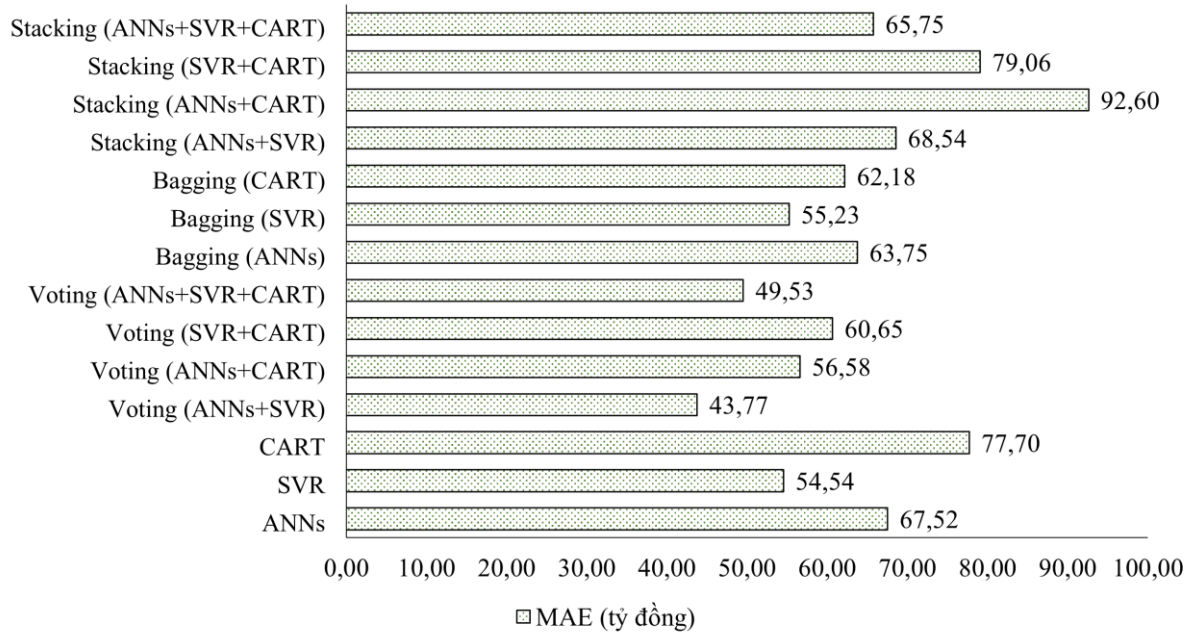
Bảng 3.7. Tổng hợp xếp hạng khả năng dự báo của tất cả mô hình.

Tên mô hình	MAE (tỷ đồng)	RMSE (tỷ đồng)	MAPE (%)	SI	Rank
Mô hình đơn lẻ					
ANNs	67,524	77,025	26,70	0,471	11
SVR	**54,535**	**62,746**	**20,50**	**0,184**	**3**
CART	77,704	91,661	25,50	0,618	12
Mô hình tích hợp - Voting					
ANNs+SVR	**43,771**	**51,479**	**19,01**	**0,011**	**1**
ANNs+CART	56,577	67,072	22,24	0,256	5
SVR+CART	60,647	71,791	21,32	0,300	6
ANNs+SVR+CART	49,525	59,675	18,34	0,094	2
Mô hình tích hợp - Bagging					
Bagging (ANNs)	63,752	72,697	24,63	0,382	8
Bagging (SVR)	**55,228**	**66,325**	**20,79**	**0,218**	**4**
Bagging (CART)	62,183	71,177	23,03	0,335	7
Mô hình tích hợp - Stacking					
ANNs+SVR	68,543	77,298	24,01	0,435	10
ANNs+CART	92,603	101,402	35,88	0,957	14
SVR+CART	79,062	89,754	38,48	0,830	13
ANNs+SVR+CART	**65,749**	**73,592**	**24,21**	**0,395**	**9**

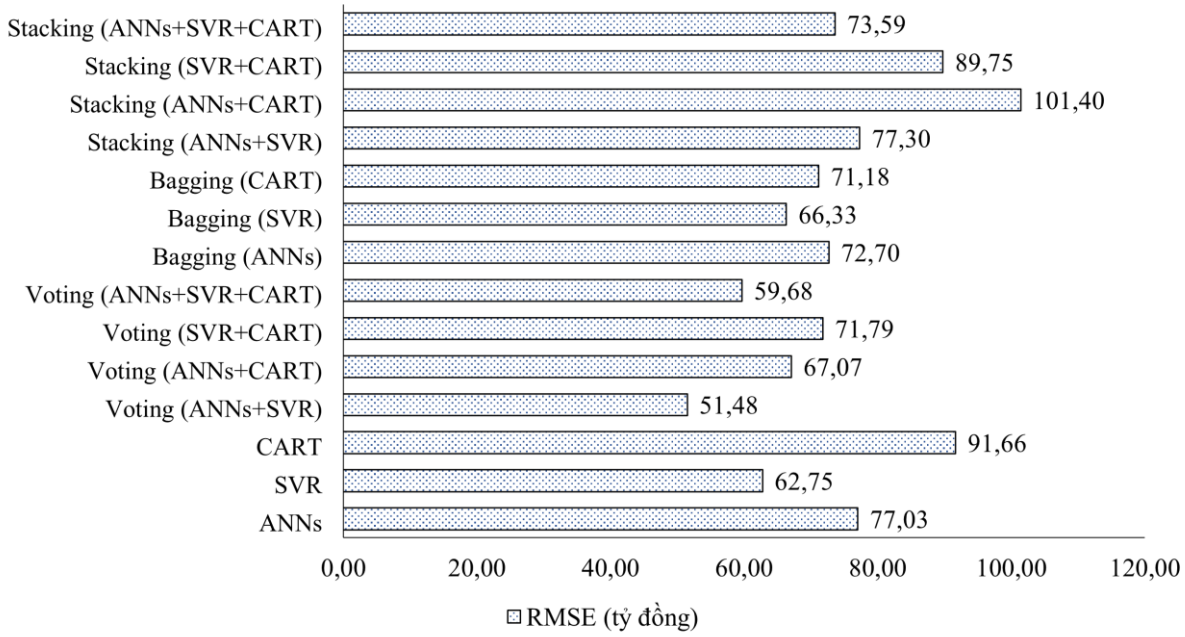
Tóm lại, kết quả xếp hạng cho thấy Voting là chiến lược tích hợp hiệu quả nhất, đặc biệt với tổ hợp ANNs+SVR. Mô hình SVR đơn lẻ vẫn duy trì vị trí cao, chứng tỏ tiềm năng làm mô hình nền tảng mạnh. Ngược lại, Bagging chỉ cải thiện hạn chế, còn Stacking không đem lại hiệu quả, thậm chí làm gia tăng sai số dự báo. Hình 3.7, Hình 3.8, và Hình 3.9 lần lượt thể hiện giá trị MAPE, MAE, và RMSE của các mô hình dự báo. Hình 3.10 tổng hợp xếp hạng của tất cả các mô hình dựa trên chỉ số SI.



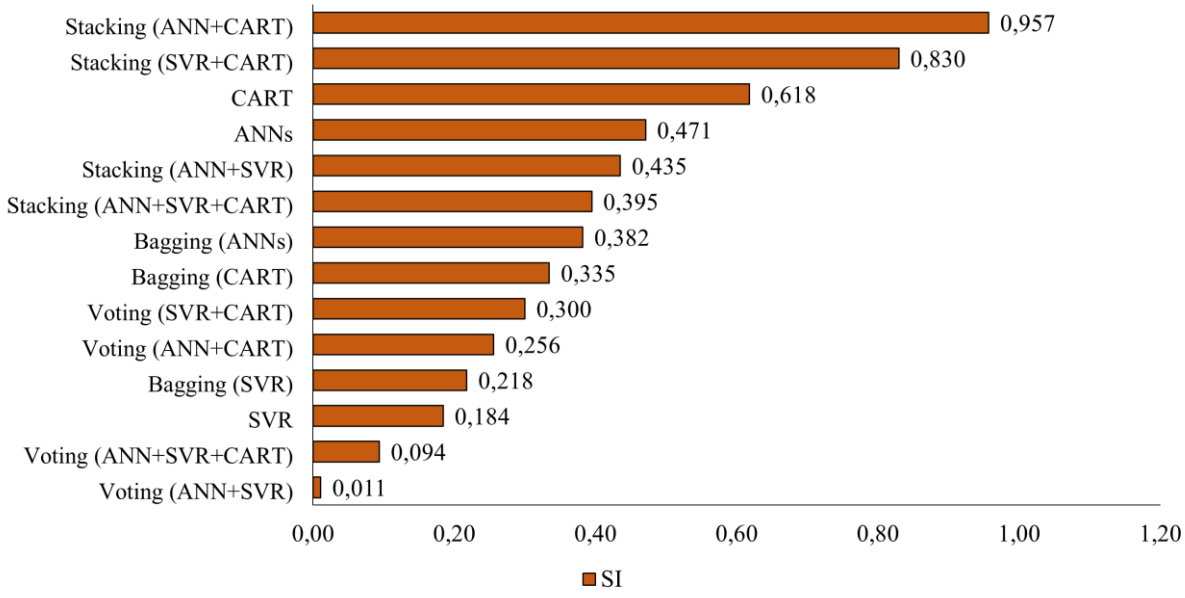
Hình 3.7. Giá trị MAPE của tất cả mô hình.



Hình 3.8. Giá trị MAE của tất cả mô hình.



Hình 3.9. Giá trị RMSE của tất cả mô hình.



Hình 3.10. Tổng hợp xếp hạng của các mô hình.

KẾT LUẬN VÀ KIẾN NGHỊ

3.5. KẾT LUẬN

Nghiên cứu đã tiến hành so sánh và đánh giá hiệu quả dự báo chi phí xây dựng của các mô hình học máy đơn (ANNs, SVR, CART) và các mô hình học máy tích hợp (Voting, Bagging, Stacking). Kết quả phân tích định lượng cho thấy mỗi nhóm mô hình thể hiện đặc trưng riêng về độ chính xác, khả năng ổn định và mức độ phù hợp trong bối cảnh dữ liệu ngành xây dựng.

Đối với nhóm mô hình đơn lẻ, SVR chứng minh là phương pháp có hiệu suất dự báo vượt trội nhất, đạt giá trị MAE, RMSE và MAPE thấp nhất trong ba mô hình. Ưu thế này bắt nguồn từ khả năng xử lý tốt các mối quan hệ phi tuyến giữa các biến đầu vào và đầu ra thông qua hàm kernel. Do đó, SVR được xem là mô hình nền tảng mạnh, có tiềm năng cao để phát triển và kết hợp trong các mô hình tích hợp.

Ở nhóm mô hình tích hợp, Voting cho thấy kết quả dự báo khả quan nhất, đặc biệt với tổ hợp ANNs+SVR, đạt giá trị SI thấp nhất (0,011) và được xếp hạng cao nhất trong toàn bộ các mô hình. Điều này khẳng định cơ chế bầu chọn của Voting giúp tận dụng ưu điểm của từng mô hình thành phần, giảm sai số và nâng cao độ tin cậy của kết quả dự báo. Tổ hợp ANNs+SVR+CART cũng đạt độ chính xác đáng kể, phản ánh hiệu quả của việc đa dạng hóa mô hình trong cùng cơ chế Voting.

Ngược lại, Bagging chỉ cải thiện hạn chế so với mô hình đơn lẻ. Mặc dù giúp duy trì độ ổn định dự báo, Bagging chưa tạo ra sự bứt phá rõ rệt về độ chính xác, trong khi Stacking thậm chí cho kết quả kém hiệu quả hơn do cơ chế kết hợp phức tạp và khả năng lan truyền sai số giữa các tầng mô hình. Như vậy, trong phạm vi nghiên cứu này, Voting được xác định là chiến lược tích hợp tối ưu cho bài toán ước tính chi phí xây dựng, trong khi SVR là mô hình đơn lẻ có tiềm năng cao nhất.

3.6. KIẾN NGHỊ

Từ các kết quả trên, kiến nghị rằng việc áp dụng mô hình học máy tích hợp, đặc biệt là Voting giữa ANNs và SVR, có thể mang lại độ chính xác và ổn định cao hơn cho quá trình dự báo chi phí xây dựng ở giai đoạn thiết kế sơ bộ. Các nhà nghiên cứu và kỹ sư thực hành có thể sử dụng phương pháp này như một công cụ hỗ trợ ra quyết định, góp phần nâng cao hiệu quả quản lý chi phí và hạn chế rủi ro trong đầu tư xây dựng.

Trong tương lai, cần mở rộng nghiên cứu theo hướng: (i) tích hợp thêm các biến đặc trưng về kỹ thuật và thị trường để nâng cao khả năng tổng quát hóa của mô hình; (ii) áp dụng các kỹ thuật giải thích mô hình như SHAP hoặc LOO-FI nhằm phân tích sâu hơn ảnh hưởng của từng biến đầu vào đến kết quả dự báo; và (iii) thử nghiệm các mô hình học sâu (deep learning) hoặc các thuật toán tối ưu lai (hybrid optimization) để cải thiện hơn nữa hiệu suất dự báo. Việc kết hợp giữa phương pháp học máy tiên tiến và dữ liệu thực tế chất lượng cao hứa hẹn mở ra hướng tiếp cận mới trong lĩnh vực ước tính chi phí xây dựng, hướng tới mô hình hóa chính xác và minh bạch hơn trong tương lai.

Quá trình thu thập và xử lý dữ liệu đóng vai trò quyết định trong việc đảm bảo độ tin cậy và khả năng khái quát hóa của mô hình dự báo chi phí xây dựng. Mặc dù bộ dữ liệu gồm 32 mẫu có độ đa dạng nhất định về quy mô và đặc điểm công trình, tuy nhiên quy mô này vẫn còn hạn chế so với yêu cầu của các mô hình học máy. Do đó, trong các nghiên cứu tiếp theo, cần mở rộng kích thước mẫu bằng cách thu thập dữ liệu từ nhiều dự án hơn. Bên cạnh đó, việc xử lý biến định tính bằng phương pháp biến giả (dummy variables) là phù hợp, song cần xem xét thêm các kỹ thuật mã hóa nâng cao nếu số lượng mẫu được mở rộng. Điều này sẽ giúp giảm hiện tượng mất cân bằng giữa các nhóm dữ liệu và cải thiện khả năng học của mô hình.

DANH MỤC TÀI LIỆU THAM KHẢO

1. Kim, G.-H., 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, 2004. **39**(10): p. 1235-1242.
2. Hassan, A., *Preliminary Construction Cost Estimate in Yemen by Artificial Neural Network.* Baltic Journal of Real Estate Economics and Construction Management, 2019. **7**: p. 110-122.
3. Chen, L., et al., *Transparent and reliable construction cost prediction using advanced machine learning and explainable AI.* Engineering Science and Technology, an International Journal, 2025. **70**: p. 102159.
4. Williams, T.P., *Predicting final cost for competitively bid construction projects using regression models.* International Journal of Project Management, 2003. **21**(8): p. 593-599.
5. Martin Skitmore, R. and S. Thomas Ng, *Forecast models for actual construction time and cost.* Building and Environment, 2003. **38**(8): p. 1075-1083.
6. Lowe David, J., W. Emsley Margaret, and A. Harding, *Predicting Construction Cost Using Multiple Regression Techniques.* Journal of Construction Engineering and Management, 2006. **132**(7): p. 750-758.
7. Jafarzadeh, R., et al., *Predicting Seismic Retrofit Construction Cost for Buildings with Framed Structures Using Multilinear Regression Analysis.* Journal of Construction Engineering and Management, 2014. **140**(3): p. 04013062.
8. Alshamrani, O.S., *Construction cost prediction model for conventional and sustainable college buildings in North America.* Journal of Taibah University for Science, 2017. **11**(2): p. 315-323.
9. Juszczyk, M., *The Challenges of Nonparametric Cost Estimation of Construction Works with the use of Artificial Intelligence Tools.* Procedia Engineering, 2017. **196**: p. 415-422.
10. Alrasheed, K., et al., *Artificial Neural Network-based cost estimation for public construction projects in Kuwait.* Journal of Engineering Research, 2025.
11. Maya, R., B. Hassan, and A. Hassan, *Develop an artificial neural network (ANN) model to predict construction projects performance in Syria.* Journal of King Saud University - Engineering Sciences, 2023. **35**(6): p. 366-371.
12. Karadimos, P. and L. Anthopoulos, *Development of Artificial Neural Networks for Predicting the Construction Costs of WWTPs in Greece.* Procedia Computer Science, 2025. **263**: p. 285-292.
13. Chakraborty, D., et al., *A novel construction cost prediction model using hybrid natural and light gradient boosting.* Advanced Engineering Informatics, 2020. **46**: p. 101201.

14. Jezeh, M.V., A. Amirkardoust, and D.S. Shayegan, *Predicting Final Construction Costs of Hospitals Based on Initial Project Attributes: An Advanced Regression Approach.* The International Journal of Multiphysics, 2025. **19**(1): p. 849 - 857.
15. Mahmoodzadeh, A., H.R. Nejati, and M. Mohammadi, *Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects.* Automation in Construction, 2022. **139**: p. 104305.
16. Cheng, M.-Y., et al., *A novel time-depended evolutionary fuzzy SVM inference model for estimating construction project at completion.* Engineering Applications of Artificial Intelligence, 2012. **25**(4): p. 744-752.
17. Jin, G. and C. Yang, *A systematic intelligent prediction model for residential construction cost based on fuzzy AHP and GA-BP neural network.* Advanced Engineering Informatics, 2026. **69**: p. 103858.
18. Liu, H., et al., *Actual construction cost prediction using hypergraph deep learning techniques.* Advanced Engineering Informatics, 2025. **65**: p. 103187.
19. Luu, V. and S.Y. Kim, *Neural Network Model for Construction Cost Prediction of Apartment Projects in Vietnam.* Korean Journal of Construction Engineering and Management, 2009. **10**.
20. Phạm, H., *Nghiên cứu dự báo chi phí biện pháp thi công xây dựng theo lý thuyết độ tin cậy.* Tạp chí Xây dựng, 2025. **5**: p. 255-259.
21. Lê, H.Q.P., et al., *Ước lượng chi phí xây dựng nhà xưởng trong giai đoạn đấu thầu ứng dụng mạng Neural nhân tạo (ANN).* Tạp chí Xây dựng, 2022. **7/2022**.
22. Dang, C.N. and L. Le-Hoai, *Revisiting storey enclosure method for early estimation of structural building construction cost.* Engineering, Construction and Architectural Management, 2018. **25**(7): p. 877-895.
23. Tôn, H.V. and C.T. Đinh, *Ước lượng chi phí xây dựng công trình nhà ở cao tầng trên địa bàn thành phố Hồ Chí Minh* Tạp chí Xây dựng, 2019. **7/2019**: p. 252-255.
24. Đức, P.A.H., Nguyễn Ngọc Thuận *Nghiên cứu mô hình tối ưu hóa lợi nhuận của nhà thầu xây dựng trong triển khai thi công các dự án nhà cao tầng.* Tạp chí Xây dựng, 2019. **07**: p. 137-141.
25. Tự, B.X., Đ.T. Sỹ, and N.T. Việt, *Xác định các yếu tố gây vượt chi phí thi công các dự án nhà cao tầng xảy ra tại các thầu ở Việt Nam.* Tạp chí Vật liệu và Xây dựng, 2023. **13**(6): p. 73-79.
26. Phước, Đ.Q., et al., *Phát triển chương trình ứng dụng mô hình thông tin (BIM) trong việc tự động hóa lập dự toán công trình xây dựng.* Tạp chí Xây dựng, 2019. **06**: p. 24-28.
27. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.* The bulletin of mathematical biophysics, 1943. **5**(4): p. 115-133.
28. Vapnik, V.N., *The Nature of Statistical Learning Theory*. Information Science and Statistics. 2013, New York: Springer New York.

29. Kavzoglu, T. and I. Colkesen, *A kernel functions analysis for support vector machines for land cover classification.* International Journal of Applied Earth Observation and Geoinformation, 2009. **11**(5): p. 352-359.
30. Breiman, L., et al., *Classification and Regression Trees* ed. s. ed. 1984: Chapman and Hall/CRC. .
31. Loh, W.-Y., *Classification and Regression Trees.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011. **1**: p. 14-23.
32. Kuncheva, L.I., *Combining Pattern Classifiers: Methods and Algorithms*. 2004, New York City, United States: Wiley.
33. Breiman, L., *Bagging predictors.* Machine Learning, 1996. **24**(2): p. 123-140.
34. Wolpert, D.H., *Stacked generalization.* Neural Networks, 1992. **5**(2): p. 241-259.
35. Kohavi, R., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. 1995, Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada. p. 1137–1143.
36. Chou, J.-S., N.-T. Ngo, and W.K. Chong, *The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate.* Engineering Applications of Artificial Intelligence, 2017. **65**: p. 471-483.
37. Chou, J.-S., et al., *Shear strength prediction of reinforced concrete beams by baseline, ensemble, and hybrid machine learning models.* Soft Computing, 2020. **24**(5): p. 3393-3411.