DỰ BÁO KHẢ NĂNG RÒ RỈ TRÊN MẠNG LƯỚI CẤP NƯỚC BẰNG MỘT SỐ KỸ THUẬT HỌC MÁY: NGHIÊN CỨU ĐIỂN HÌNH CHO HỆ THỐNG CẤP NƯỚC TRUNG AN - THÀNH PHỐ HỒ CHÍ MINH

Nguyễn Hoàng Tuấn[1], Trần Đăng An[2], Triệu Ánh Ngọc[2], Huỳnh Duy Linh[3]

Tóm tắt: *Dự báo khả năng rò rỉ trên mạng lưới cấp nước luôn là vấn đề khó và được quan tâm hàng đầu, đặc biệt là những thành phố có mạng lưới cấp nước lớn, phức tạp như thành phố Hồ Chí Minh. Nghiên cứu này được thực hiện dựa trên 126 mẫu thu thập được trên cơ sở dữ liệu không gian với 11 yếu tố ảnh hưởng đến khả năng rò rỉ: tuổi ống, đường kính, vật liệu, sức chịu tải nền đất, tải trọng giao thông, độ sâu lắp đặt, áp lực, lưu lượng, chênh lệch áp lực, số đấu nối và mật độ dân số. Các mô hình học máy được sử dụng: Random Forest Regression, Extreme Gradient Boosting Regression, Light Gradient Boosting Regression và Catboost Regression để đánh giá khả năng dự báo rò rỉ trên mạng lưới thông qua các thông số: sai số bình phương gốc (RMSE), hệ số xác định (R^2), tiêu chí thông tin Akaike (AIC) và tiêu chí thông tin Bayes (BIC) để lựa chọn ra mô hình phù hợp nhất. Kết quả mô phỏng cho thấy, mô hình CastBoost cho kết quả dự báo về khả năng rò rỉ trên mạng lưới tốt nhất. Các mô hình khác cũng có kết quả khá tốt. Tuy nhiên, mô hình SVR được đánh giá không phù hợp với bộ số liệu thu thập. Kết quả cũng chỉ ra rằng, các yếu tố khác cần được bổ sung để nâng cao hiệu quả dự báo của mô hình và có khả năng ứng dụng trong thực tế giảm thất thoát nước trên mạng lưới cấp nước.*

Từ khóa: Thất thoát nước, dự báo rò rỉ, học máy, Tp.Hồ Chí Minh.

1. ĐẶT VẤN ĐỀ

Thất thoát nước là một trong những thách thức lớn đối với các công ty quản lý cấp nước trên giới nói chung và Việt Nam nói riêng, đặc biệt là Tổng công ty cấp nước Sài Gòn (SAWACO) và các công ty cấp nước thành viên. Hiện nay tỷ lệ thất thoát nước trung bình của SAWACO là khoảng 18% điều này đã thúc đẩy công ty phải tiến hành giảm thiểu mức độ thất thoát nước trên mạng lưới cấp nước, đặc biệt là thất thoát nước do rò rỉ thông qua việc phát triển và ứng dụng nhiều kỹ thuật để xác định, định vị và khắc phục các vị trí rò rỉ và vỡ ống.

Các phương pháp truyền thống đang được sử dụng rộng rãi để điều tra, xác định vị trí, số lượng và quy mô rò rỉ trên mạng thường yêu cầu nguồn nhân lực và tài chính lớn. Hiện nay, nhiều công ty cấp nước đã được áp dụng việc giám sát mạng trực tuyến theo thời gian thực, tạo điều kiện phát hiện sớm và khoanh vùng rò rỉ; phương pháp này có ưu điểm là giúp đơn vị quản lý vận hành dễ dàng theo dõi và có phương án khắc phục hiệu quả. Tuy nhiên, phương pháp cũng đòi hỏi kinh phí đầu tư rất lớn và nguồn nhân lực có trình độ cao mới phát huy được hiệu quả. Bên cạnh đó, phương pháp này cần đòi hỏi đội ngũ vận hành có trình độ cao. Để rút ngắn thời gian và tăng hiệu quả trong việc giám sát, quản lý và xử lý rò rỉ trên mạng lưới cấp nước, kỹ thuật học máy đã được ứng dụng nhiều trong những năm qua tại các nước trên thế giới và đem lại những hiệu quả hết sức tích cực (Banjara, Sasmal, & Voggu, 2020; Hu, Han, Yu, Geng, & Fan, 2021). Hu và ctv (Hu et al., 2021) sử dụng mạng nơ-ron đa tầng để xác

[1] *Phòng Công nghệ Thông tin, Tổng Công ty Cấp nước Sài Gòn – TNHH MTV.*
[2] *Phân hiệu Trường Đại học Thủy lợi.*
[3] *Phòng kỹ thuật, Công ty Cổ phần Sonadezi Long Bình*

định chính xác vị trí các điểm rò rỉ nước trên mạng lưới cấp nước. Ngoài ra, Candelieri và ctv (Candelieri, Soldi, Conti, & Archetti, 2014)đề xuất cách tiếp cận dựa trên mô phỏng thủy lực và học máy để cải thiện kiểm soát rò rỉ thông qua phân tích các thông tin của điểm rò rỉ trên mạng lưới cấp nước. Cantos và ctv (Cantos Wilmer, Juran, & Tinelli, 2020) đã kết hợp mô phỏng thủy lực và học máy để xác định điểm rò rỉ trên mạng lưới cấp nước. Kỹ thuật học máy cũng được sử dụng để hỗ trợ hiệu quả phương pháp đo âm thanh trong phát hiện rò rỉ trên hệ thống đường ống dẫn nước (Banjara et al., 2020).

Tại Việt Nam đặc biệt là ở khu vực Tp.HCM, trong những năm gần đây nghiên cứu về dự báo rò rỉ trên mạng lưới cấp nước đã được nhiều tác giả quan tâm. Võ Anh Tuấn, 2015 đã tiến hành nghiên cứu đặc điểm rò rỉ thất thoát nước trên hệ thống cấp nước SAWACO bằng phương pháp điều tra, quan trắc và phân tích đặc điểm rò rỉ nước trên hệ thống từ đó xác định nguyên nhân gây ra hiện tượng này. Phạm Thi Minh Lành và Nguyễn Quang Trưởng (Phạm Thị Minh Lành, 2022) đã sử dụng kết hợp điều tra, quan trắc, sử dụng mô hình thủy lực WaterGEMs và mô hình lý thuyết mờ (Fuzzy Logic) để xác định hệ số rò rỉ nước trên mạng lưới cấp nước. Trong nghiên cứu này, tác giả đã sử dụng dụng một số thuật toán học máy bao gồm mô hình hồi quy Logistic (Logistic Regression Model), mô hình cây quyết định (Decision Tree Model) và mô hình mạng Nơ-ron nhân tạo (Artificial Neural Network model) để xây dựng mô hình dự báo rủi ro do rò rỉ nước gây ra trên mạng lưới cấp nước Phường 17, Quận Gò Vấp, Tp. Hồ Chí Minh (Phạm Thi Minh Lanh, 2022).

Có thể thấy rằng kỹ thuật học máy đã được ứng dụng rộng rãi trong trong nghiên cứu xác định khả năng, số lượng và lưu lượng rò rỉ nước trên mạng lưới ở nhiều nước trên thế giới. Tại Việt Nam, một số nghiên cứu ban đầu về rò rỉ thất thoát nước theo hướng tiếp cận mới này đã đạt được một số kết quả nhất định. Tuy nhiên, ứng dụng kỹ thuật học máy trong nghiên cứu rò rỉ nước trên mạng lưới cấp nước đô thị ở nước ta vẫn còn là một trong lĩnh vực rất mới mẻ và chưa được ứng dụng rộng rãi. Do đó, việc nghiên cứu, đánh giá khả năng rò rỉ và các yếu tố ảnh hưởng đến rò rỉ trên mạng lưới cấp nước trên địa bàn Tp.HCM dựa trên kỹ thuật học máy là cần thiết và có ý nghĩa khoa học, ý nghĩa thực tiễn, góp phần nâng cao hiệu quả giảm thất thoát nước của ngành cấp nước Tp.HCM nói riêng và ngành cấp nước Việt Nam nói chung.

Mục tiêu của nghiên cứu này là đánh giá và lựa chọn các thuật học máy tiên tiến hiện nay bao gồm mô hình Random Forest (RFR), Mô hình Support Vector Machine (SVR), Mô hình Extreme Gradient Boosting (XGB), Mô hình Light Gradient Boosting (LGB), và Mô hình CatBoost (CBR) phục vụ dự báo số điểm rò rỉ nước trên mạng lưới cấp nước. Trên cơ sở đó sẽ đề xuất mô hình phù hợp với mô phỏng dự báo điểm rò rỉ phục vụ quản lý hiệu quả thất thoát nước trên mạng lưới cấp nước điển hình tại Tp. Hồ Chí Minh.

2. GIỚI THIỆU VỀ VÙNG NGHIÊN CỨU

Hệ thống cấp nước Trung An nằm ở phía Bắc – Tp.HCM, chiếm 14,7% diện tích nội thành, 8,7% tổng diện tích toàn Thành phố với cao độ địa hình biến đổi từ +15 m đến +1 m (các bờ sông Vàm Thuật, sông Sài Gòn) với nhiều loại hình địa chất khác nhau. Tổng dân số trong vùng khoảng 1.8 triệu người (Niên giám thống kê, 2019). Khu vực này có nhiều đối tượng sử dụng nước bao gồm sinh hoạt và ăn uống của dân cư trên địa bàn chiếm trên 70% bên cạnh đó nhu cầu nước cho sản xuất công nghiệp, tiểu thủ công nghiệp, thương mai- dịch vụ, và nông nghiệp. Đây là khu vực có tốc độ đô thị hóa và tỉ lệ tăng trưởng kinh tế nhanh điều này tạo ra sức ép rất lớn đối với mạng lưới đường ống truyền tải và phân phối hiện hữu.

Mạng lưới cấp nước Trung An bao gồm 72 km ống truyền dẫn; 2,018 km ống phân phối và

hơn 1,866 km ống dịch vụ. Tỷ lệ thất thoát nước bình quân năm 2014 là 41,8%, đến cuối năm 2020 là 18.05% và hướng tới năm 2025 là 16.5%. Thất thoát nước do rò rỉ vỡ ống trong khu vực nghiên cứu chủ yếu là do các yếu tố chính như đã đề cập ở **Bảng 1**. Trong đó, tuổi thọ đường ống, chênh lệch áp lực và đặc tính vật liệu làm ống được xem là những yếu tố chính ảnh hưởng tới khả năng thất thoát nước do rò rỉ và vỡ ống diễn ra trên mạng lưới cấp nước khu vực này.

3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Các mô hình học máy

3.1.1. Mô hình Random Forest (RFR)

RF là một kỹ thuật phân lớp và hồi quy (Friedman, 2001) bằng cách sử dụng nhiều cây phân lớp hoặc hồi quy trong một nhóm. Thuật toán này là một trong những thuật toán được xây dựng dựa trên mô hình cây quyết định. Mỗi cây đóng vai trò như một lá phiếu làm cơ sở ra quyết định cho thuật toán. Các phương pháp học nhóm kết hợp với các kết quả riêng lẻ của từng cây thường mang lại các kết quả tốt hơn. Random Forest là thuật toán được mở rộng dựa trên kỹ thuật đóng gói (bagging) hoặc tập hợp boostrap sử dụng các mẫu ngẫu nhiên (có lặp lại) của dữ liệu huấn luyện để tạo ra nhiều cây dữ liệu hồi quy không cần cắt tỉa và là tổng kết quả trung bình của chúng.

3.1.2. Mô hình Support Vector Machine (SVR)

SVM là một thuật toán học máy có giám sát (Balabin & Lomakina, 2011) được sử dụng rất phổ biến ngày nay trong các bài toán phân lớp hay hồi qui. Ý tưởng của SVM là tìm một mặt siêu phẳng để phân tách các điểm dữ liệu. Mặt siêu phẳng này sẽ chia không gian thành các miền khác nhau và mỗi miền sẽ chứa một loại dữ liệu.

3.1.3. Mô hình Extreme Gradient Boosting (XGB)

XGB là một giải thuật dựa trên Gradient Boosting (Friedman, 2001)dựa trên cây quyết định. Tuy nhiên, XGB là những cải tiến to lớn về mặt tối ưu thuật toán, về sự kết hợp hoàn hảo giữa sức mạnh phần mềm và phần cứng, giúp đạt được những kết quả vượt trội cả về thời gian học tập cũng như bộ nhớ sử dụng. Kể từ lần đầu ra mắt năm 2014, XGB nhanh chóng được đón nhận và là giải thuật được sử dụng chính, tạo ra nhiều kết quả vượt trội.

3.1.4. Mô hình Light Gradient Boosting (LGB)

LGB cũng là một thuật toán dựa trên Gradient Boosting (Tran et al., 2021). Đây là thuật toán có nhiều cải tiến: tốc độ huấn luyện và hiệu quả cao hơn, ít tốn bộ nhớ hơn, độ chính xác tốt hơn bất kỳ thuật toán Boosting nào khác.

3.1.5. Mô hình CatBoost (CBR)

Thuật toán CatBoost được xây dựng dựa trên cây quyết định được tăng cường gradient bao gồm tập dữ liệu đào tạo, với độ chính xác được xác định trên tập dữ liệu xác thực. Thuật tóa này được phát triển bởi các kỹ sư và nhóm nghiên cứu thuộc Công ty Yandex, Nga (Hancock & Khoshgoftaar, 2020). CB là sự kế thừa thuật toán MatrixNet được sử dụng rộng rãi trong xếp hạng các nhiệm vụ, dự báo và đưa ra các khuyến nghị. Thuật toán này đã trở thành một trong những thuật toán học máy phổ biến nhất và được áp dụng để xử lý các vấn đề khác nhau trên nhiều lĩnh vực khác nhau.

3.2. Số liệu đầu vào mô hình

Dựa trên các nghiên cứu về các yếu tố ảnh hưởng đến rò rỉ mạng lưới phân phối được nghiên cứu bởi (Hu et al., 2021; Wéber, Huzsvár, & Hős, 2021; Xue et al., 2020), trong nghiên cứu này đã phân tích và lựa chọn 11 yếu tố ảnh hưởng đến khả năng rò rỉ mạng lưới cấp nước Trung An – Tp. HCM. Tổng cộng 126 mẫu dữ liệu không gian được thu thập từ nhiều nguồn khác nhau như trình bày ở **Bảng 1**, thể hiện những nguyên nhân chính gây ra rò rỉ trên hệ thống mạng lưới cấp nước hiện trạng của khu vực cấp nước Trung An.

Bảng 1. Các yếu tố ảnh hưởng đến khả năng rò rỉ của nghiên cứu

STT	Diễn giải	Ký hiệu	Đơn vị	Nguồn
1.	Đường kính ống	DIA	mm	TAWACO
2.	Module đàn hồi (Vật liệu ống)	ELA	Gpa	TAWACO
3.	Tuổi ống: tính từ năm thi công đến thời điểm hiện tại	AGE	năm	TAWACO
4.	Lưu lượng nước đi qua ống trong thời gian một giờ	QAN	m^3/h	TAWACO
5.	Áp lực trung bình của ống	PRS	mH_20	TAWACO
6.	Chênh lệch áp lực nước là hiệu số giữa áp lực cao nhất và thấp nhất trong lòng ống	DPRS	mH_20	TAWACO
7.	Sức chịu tải của nền đất	GRD	kN/m^2	TAWACO
8.	Ảnh hưởng của giao thông (tim đường, cấp đường, mật độ giao thông)	TIP		TAWACO
9.	Mật độ dân số	POP	Ng/km^2	TAWACO
10.	Độ sâu lắp đặt của đường ống	DPP	m	TAWACO
11.	Số đấu nối	CNT		TAWACO

Ghi chú: TAWACO – Công ty Cổ phần cấp nước Trung An

3.3. Phương pháp nghiên cứu

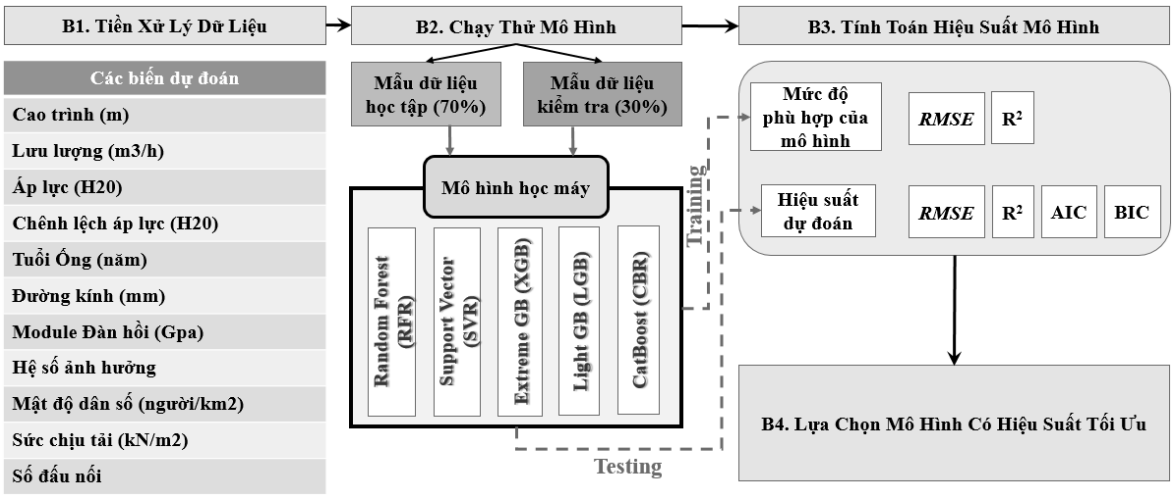
Trong nghiên cứu này mô hình dự báo các điểm rò rỉ nước trên mạng lưới cấp nước được thiết lập thông qua 04 bước cơ bản như **Hình 1**, chi tiết được diễn giải cụ thể dưới đây.

3.3.1. Chuẩn bị và xử lý dữ liệu

Tập dữ liệu thu thập được ở các nguồn thứ cấp và sơ cấp có nhiều dạng dữ liệu, cần phải được số hóa và chuẩn hóa để có thể chạy được các mô hình học máy. Ví dụ: cấp đường giao thông (A, B, C…) hoặc địa chất nền ống (đất sét mềm, cát mịn lỏng - khô, cát vừa nhỏ - gọn - khô …); sữa chữa những sai số về số học, lỗi trong ghi nhận dữ liệu. Các dữ liệu này được xử lý để đưa về dữ liệu chuẩn hóa mang giá trị liên tục để nhập vào mô hình học máy như Bảng 1.

3.3.2. Chạy thử mô hình học máy

Sau khi có bộ dữ liệu được chuẩn hóa là một ma trận có 126 cột (DMA) và 11 biến độc lập (các yếu tố ảnh hưởng tới rò rỉ), tập dữ liệu được chia thành 2 tập dữ liệu huấn luyện và tập dữ liệu kiểm tra với tỷ lệ 70/30 một cách ngẫu nhiên. Dữ liệu dự đoán đầu ra là những giá trị liên tục, do đó với những mô hình học máy sẽ được chạy ở kỹ thuật hồi quy. Mô hình học máy được lựa chọn để thực hiện là các mô hình hồi quy RFR, SVR, XGB, LGB và CBR.



Hình 1. Phương pháp nghiên cứu

3.3.3. Tính toán hiệu suất mô hình

Việc tính toán hiệu suất của mô hình được căn cứ dựa trên các tiêu chí (Tran et al., 2021): Root mean squared error (RMSE) – sai số bình phương gốc là độ lệch chuẩn của lỗi dự đoán, cho biết mức độ tập trung dữ liệu xung quanh dòng phù hợp nhất. RMSE được sử dụng trong các mô hình học máy dự báo để xác minh kết quả. RMSE càng bé, mức độ chính xác càng tốt. Giá trị của RMSE được tính theo công thức:

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_i - \hat{y})^2} \qquad (1)$$

Coefficient of Determination (R^2) – hệ số xác định: là thước đo cho sự phù hợp của mô hình, cho biết tỷ lệ dự đoán của biến phụ thuộc đối với biến độc lập. R^2 càng lớn, mức độ phù hợp càng tốt. Giá trị của R^2 được tính theo công thức:

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

y_i : giá trị thực tế; $\hat{y}$: giá trị dự đoán; $\bar{y}$: giá trị trung bình

Akaike information criterion (AIC) (Akaike, 1974) – Tiêu chí thông tin Akaike: ước tính lượng thông tin tương đối bị mất bởi một mô hình nhất định, mô hình mất càng ít thông tin thì chất lượng của mô hình đó càng cao. Giá trị của AIC được tính theo công thức:

$$AIC = 2k - 2\log(L) \qquad (3)$$

Bayesian information criterion (BIC) (Stone, 1979) – Tiêu chí thông tin Bayes: là một tiêu chí để lựa chọn mô hình trong số các mô hình hữu hạn; mô hình có BIC thấp nhất được lựa chọn. Giá trị của BIC được tính theo công thức:

$$BIC = 2k\log(n) - 2\ln(L) \qquad (4)$$

k: *số biến độc lập;* ***n:*** *số lượng mẫu;* ***L:*** *Likelihood*

3.4. Lựa chọn mô hình tối ưu

Siêu tham số (Hyperparameter) được hiểu như là: Mọi mô hình học máy có thể được định nghĩa là một mô hình toán học với một số tham số. Giá trị của các tham số này ảnh hưởng đến việc huấn luyện và do đó độ chính xác của mô hình. Hiệu chỉnh siêu tham số là quá trình chọn một tập hợp các siêu tham số tối ưu cho một thuật toán học máy. Quá trình trình này được thực hiện một cách tự động, nhằm giúp đạt được độ chính xác tối đa có thể của dự đoán. Có nhiều cách để đạt được kết quả tối ưu của điều chỉnh siêu tham số trong học máy như: Grid Search, Random Search hay Bayesian Optimization. Trong nghiên cứu này, chúng tôi sử dụng bộ dữ liệu là một ma trận có 126 hàng tương ứng với số DAM và 11 cột tương ứng với số biến độc lập ảnh hưởng đến mô hình dự báo rò rỉ. Với dữ liệu này phương pháp Grid Search được chọn để điều chỉnh siêu tham số nhằm tìm ra mô hình tối ưu.

4. KẾT QUẢ VÀ THẢO LUẬN

4.1. Lựa chọn mô hình dự đoán điểm rò rỉ

Bảng 2. Hiệu suất các mô hình khi ở chế độ mặc định

	RFR	XGB	LGB	CBR	SVR
RMSE	190	195	192	120	229
R^2	0.46	0.42	0.446	0.81	0.21

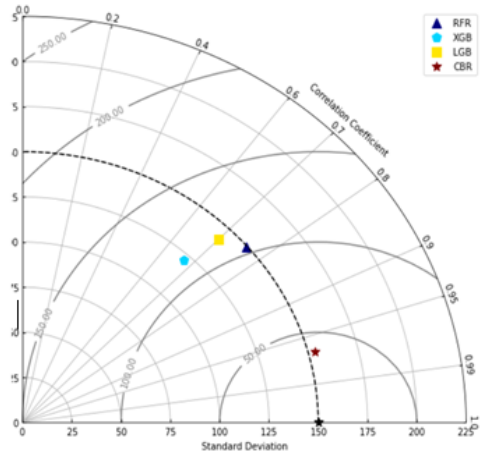
Căn cứ trên kết quả thống kê hiệu suất của mô hình theo Bảng 2, dễ dàng nhận thấy rằng mô hình CBR có độ chính xác nhất với R^2 =0.81 và RMSE = 120 tiếp theo các mô hình RFR, XGB, LGB. Kết quả cũng cho thấy rằngmô hình SVR có độ lệch chuẩn RMSE rất cao (229) và hệ số R^2 rất thấp chỉ khoảng 0.21. Điều này cho thấy, phương pháp dự báo bằng mô hình SVR – hồi quy dựa theo vector hỗ trợ với bộ số liệu đầu vào để mô hình hóa sự tương quan ảnh hưởng của các yếu tố gây ảnh hưởng đến rò rỉ mạng lưới cấp nước không đạt được hiệu quả cao trong trường hợp cụ thể mạng lưới cấp nước Trung An. Do đó, mô hình SVR sẽ không được sử dụng cho các bước phân tích dự báo tiếp theo.

Các bước thiết lập – cấu hình, huấn luyện (training) và kiểm nghiệm (testing) của bốn mô hình học máy được thực hiện trong công cụ Jupyter (Python) phiên bản 6.3.0 với 126 mẫu được chia ngẫu nhiên thành tập dữ liệu huấn luyện(70%) và tập dữ liệu kiểm nghiệm (30%) để đánh giá độ chính xác và phù hợp của mô hình bằng cách sử dụng gói Scikit-learning. Siêu tham số của bốn mô hình học máy (RFR, XGBR, CBR và LGBR) đã được điều chỉnh bằng cách sử dụng chức năng Grid Search với Cross Validation = 5 trong mô đun Scikit-learning.

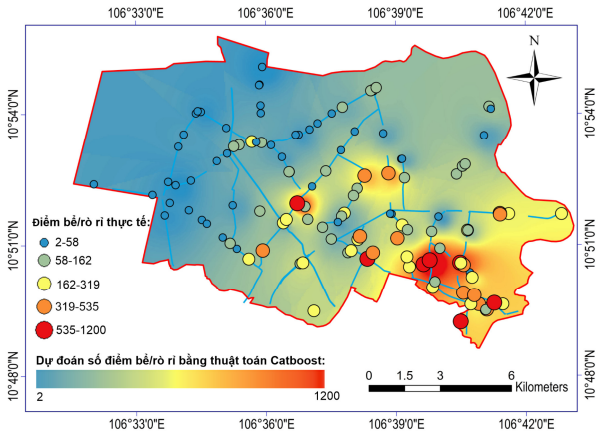
Bảng 3. Kết quả đánh giá các mô hình sau khi hiệu chỉnh siêu tham số

	Bước huấn luyện			Bước kiểm nghiệm			
	RMSE	**R^2**		**RMSE**	**R^2**	**AIC**	**BIC**
XGB	112	0.38		125	0.44	1237.94	1269.13
LGB	107	0.32		121	0.47	1231.44	1262.40
CBR	0	1.0		99	0.83	1179.87	1211.07
RFR	0	1.0		109	0.56	1204.25	1235.45

Từ Bảng 3 thấy mô hình CBR (với hiệu chỉnh siêu tham số: learning_rate = 0.01, depth = 3, n_estimators = 100) có kết quả dự đoán cao nhất với bộ kiểm nghiệm (RMSE = 99, R^2 = 0.82), tốt hơn đáng kể so với các mô hình XGB (RMSE = 125, R^2 = 0.44), LGB (RMSE = 121 , R^2 = 0.47), RFR (RMSE = 109, R^2 = 0.57). Bên cạnh đó, các giá trị AIC và BIC chỉ ra sự khác biệt đáng kể về mặt thống kê giữa các mô hình (theo Bảng 3). Việc đánh giá khả năng dự báo của các mô hình học máy bằng biểu đồ Taylor (Taylor, 2001) (theo Hình 2) cũng thể hiện rõ các kết quả này. Giá trị dự báo từ mô hình CBR có mối tương quan cao hơn và sai số bình phương gốc thấp hơn so với các mô hình XGB, LGB và RFR.



Hình 2. Đánh giá các mô hình học máy dựa vào đồ thị Taylor



Hình 3. Kết quả dự đoán số điểm rò rỉ và kết quả thống kê số điểm rò rỉ trên thực tế trong khu vực nghiên cứu

4.2. Kết quả dự đoán điểm rò rỉ

Dựa vào kết quả dự đoán số lượng các điểm rò rỉ trong 126 DMAs của mạng lưới cấp nước Trung An từ mô hình Catboost ở phần 4.1, bản đồ phân bố theo không gian các điểm rò rỉ dự đoán trên mạng lưới này được thiết lập bằng cách sử dụng phương pháp nội suy nghịch đảo khoảng cách viết tắt là IDW (Inverse Distance Weight). Kết quả nội suy sự phân bố theo không gian các điểm rò rỉ được chia theo 5 lớp bao gồm lớp 2-58; 58-162; 162-319; 319-535 và lớp 535-1200 điểm rò rỉ. Ngoài ra, số liệu thống kê các điểm rò rỉ từ thực tế trong các DMAs của khu vực nghiên cứu được chồng xếp với dữ liệu bản đồ nội suy phân bố không gian các điểm rò rỉ để kiểm tra mức độ phù hợp giữa kết quả mô hình và số liệu thực đo như Hình 3. Có thể thấy rằng kết quả dự báo và số liệu thống kê các điểm rò rỉ trong các DMAs của mạng lưới cấp nước Trung An là khá phù hợp kể cả về số lượng và vị trí phân bố của chúng trên mạng lưới nghiên cứu này. Dựa vào bản đồ này có thể thấy rằng các DMAs nằm ở phía Đông Nam và khu vực trung tâm của mạng lưới cấp nước Trung An có số lượng điểm rò rỉ rất lớn dao động từ 535 tới 1200 điểm trong khi đó khu vực Tây Bắc và phía Nam có số lượng điểm rò rỉ trong các DMAs là khá nhỏ dưới 58 điểm. Kết quả này là thông tin hữu ích hỗ trợ các đơn vị quản lý vận hành có thể phân vùng ưu tiên thứ tự các khu vực cần sửa chữa nâng cấp mạng lưới để giảm lượng nước rò rỉ thất thoát hiệu quả hơn. Ví dụ như cần được quan tâm tập trung nguồn lực để giảm thiểu rò rỉ thất thoát ở khu vực phía Đông Nam và khu vực trung tâm của mạng lưới cấp nước Trung An do các khu vực này có số lượng rò rỉ trong các DMAs lớn hơn 500 điểm. Ngược lại, nếu nguồn lực tài chính còn hạn chế thì chưa cần phải tập trung đầu tư nhiều nguồn lực để giảm số lượng điểm rò rỉ xuống mức thấp hơn nữa các phía Tây Bắc, phía Nam và phía Bắc của mạng lưới cấp nước khu vực này do số điểm rò rỉ ở mức khá thấp dưới 58 điểm.

5. KẾT LUẬN

Kết quả nghiên cứu cho thấy rằng mô hình CBR cho hiệu quả dự đoán số lượng điểm rò rỉ trên mạng lưới là tốt nhất với R^2 = 0.83 và RMSE =99, trong khi đó mô hình SVR cho kết quả dự báo rất kém chính xác với hệ số R^2 = 0.29 và RMSE =229. Các mô hình học máy còn lại như RFR, XGB, và LGB cho kết quả kém chính xác hơn mô hình CBR trong nghiên cứu này do đó khi áp dụng cần phải xem xét đến quy mô và đặc tính dữ liệu đầu vào cho các mô hình dự báo. Nghiên cứu này cũng đã xác định, phân tích, đánh giá và giải thích được sự ảnh hưởng của

các yếu tố đối tác động đến khả năng gây ra rò rỉ trên mạng lưới cấp nước Trung An với 11 yếu tố ảnh hưởng chính.

Mặc dù các kết quả dự báo số lượng các điểm rò rỉ nước trên mạng lưới cấp nước từ nghiên cứu này chưa đạt được mức độ chi tiết cao tuy nhiên từ các kết quả của nghiên cứu này có thể thấy rằng (1) mô hình học máy có tiềm năng rất lớn trong việc hỗ trợ xác định số lượng các điểm rò rỉ, phân vùng ưu tiên đầu tư nâng cấp sửa chữa và quản lý hiệu quả thất thoát nước; (2) kết quả dự báo này có thể là thông tin hữu ích hỗ trợ giám sát, quản lý, vận hành và nâng cao chất lượng dịch vụ cấp nước cho các công ty cấp nước hiện này đặc biệt góp phần giảm thiểu đáng kể các sai sót trong công tác quản lý, tiết kiệm nguồn nhân lực, tận dụng được nguồn dữ liệu khổng lồ thu thập được từ các nguồn mang lại hiệu quả về kinh tế và kỹ thuật rất lớn.

Điều cần lưu ý đó là nghiên cứu này chỉ dừng lại ở việc sử dụng các mô hình học máy và hiệu chỉnh các tham số đầu vào các mô hình học máy nhằm tìm ra mô hình phù hợp với dữ liệu đầu vào của mạng lưới cấp nước trong khu vực nghiên cứu. Để tăng độ chính xác và mức độ chi tiết của mô hình dự báo, các nghiên cứu tiếp theo cần xây dựng bộ dữ liệu đầu vào chi tiết tới các điểm đồng hồ sử dụng nước của từng hộ dân, xem xét và đánh giá các yếu tố ảnh hưởng chính tới khả năng rò rỉ nước dựa vào đặc điểm mạng lưới cấp nước thực tế nhằm loại bỏ các biến đầu có thể gây nhiễu cho các mô hình dự báo.

TÀI LIỆU THAM KHẢO

Phạm Thị Minh Lành, N. Q. T. (2022). *Mô hình ước lưu lượng lượng nước rò rỉ theo áp suất trên mạng lưới cấp nước.* Tạp chí Tài Nguyên Nước.

Akaike, H. (1974). *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 19(6), 716-723.

Balabin, R. M., & Lomakina, E. I. (2011). *Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data.* Analyst, 136(8), 1703-1712. doi:10.1039/C0AN00387E

Banjara, N. K., Sasmal, S., & Voggu, S. (2020). *Machine learning supported acoustic emission technique for leakage detection in pipelines.* International Journal of Pressure Vessels and Piping, 188, 104243.

Candelieri, A., Soldi, D., Conti, D., & Archetti, F. (2014). *Analytical Leakages Localization in Water Distribution Networks through Spectral Clustering and Support Vector MACHINES.* The Icewater Approach. Procedia Engineering, 89, 1080-1088.

Cantos Wilmer, P., Juran, I., & Tinelli, S. (2020). *Machine-Learning–Based Risk Assessment Method for Leak Detection and Geolocation in a Water Distribution System.* Journal of Infrastructure Systems, 26(1), 04019039. doi:10.1061/(ASCE)IS.1943-555X.0000517

Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine.* The Annals of Statistics, 29(5), 1189-1232.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). *CatBoost for big data: an interdisciplinary review.* Journal of Big Data, 7(1), 94. doi:10.1186/s40537-020-00369-8

Hu, X., Han, Y., Yu, B., Geng, Z., & Fan, J. (2021). *Novel leakage detection and water loss management of urban water supply network using multiscale neural networks.* Journal of Cleaner Production, 278, 123611.

Phạm Thi Minh Lanh, N. Q. T. (2022). *A comparison study of water pipe failure prediction models.* Journal of Water Resources.

Stone, M. (1979). *Comments on Model Selection Criteria of Akaike and Schwarz.* Journal of the Royal Statistical Society. Series B (Methodological), 41(2), 276-278.

Taylor, K. E. (2001). *Summarizing multiple aspects of model performance in a single diagram.* Journal of Geophysical Research: Atmospheres, 106(D7), 7183-7192.

Tran, D. A., Tsujimura, M., Ha, N. T., Nguyen, V. T., Binh, D. V., Dang, T. D., . . . Pham, T. D. (2021). *Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam.* Ecological Indicators, 127, 107790.

Wéber, R., Huzsvár, T., & Hős, C. (2021). *Vulnerability of water distribution networks with real-life pipe failure statistics.* Water Supply, ws2021447. doi:10.2166/ws.2021.447

Xue, P., Jiang, Y., Zhou, Z., Chen, X., Fang, X., & Liu, J. (2020). *Machine learning-based leakage fault detection for district heating networks.* Energy and Buildings, 223, 110161.

Abstract:

PREDICTION OF WATER LEAKAGES IN WATER DISTRIBUTION NETWORK USING MACHINE LEARNING TECHNIQUES: A CASE STUDY FOR TRUNG AN WATER SUPPLY SYSTEM - HO CHI MINH CITY

This study applied several novel machine learning algorithsms to predict the number of water leakage points in 126 DMA with 11 factors that affect the possibility of leakage: pipe age, diameter, materials, movement of the soil, traffic loads, depth of placement, pressure, flow, differential pressure, number of connections, and population density. The machine learning models are used as Random Forest Regression (RFR), Extreme Gradient Boosting Regression (XGB), Light Gradient Boosting Regression (LGB), and Catboost Regression (CBR) combined with the performance appraisals as well as reliability of the machine learning model by comparing the Root-Mean-Square Errors (RMSE), Coefficient of determination (R2), Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) to evaluate the effectiness of the models. The result revealed that the CBRt model showed the best prediction results of water leakage in DMAs. However, detailed dataset and presection of influenced factors should be performed to increase the accuracy of the model and to be more effective in reducing water loss.

Keywords: Non-revenue water, leak prediction, machine learning, HoChiMinh City.

Ngày nhận bài: 02/01/2022

Ngày chấp nhận đăng: 04/3/2022