Ứng dụng khuyến nghị xây dựng Công trình Xanh

CS313.P23 - Khai thác dữ liệu và ứng dụng

22520003 - Huỳnh Trọng Nghĩa
22520010 - Đinh Thiên Ân
22520019 - Nguyễn Ấn
22522021 - Nguyễn Hoàng Gia An
22520069 - Phạm Nguyên Anh
22520109 - Nguyễn Gia Bảo

Khoa Khoa học Máy tính
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

May 9, 2025

Phân công công việc

Thành viên	MSSV	Nhiệm vụ	Tiến độ hoàn thành
Huỳnh Trọng Nghĩa	22520003	Đọc và tổng hợp các paper Cài đặt mô hình Cài đặt ứng dụng Làm slide đồ án	100%
Đinh Thiên Ân	22520010	Đọc và tổng hợp các paper Tải và kiểm tra tính toàn vẹn của dữ liệu Làm sạch, tiền xử lý dữ liệu Xây dựng pipeline nạp dữ liệu Cài đặt ứng dụng	100%
Nguyễn Ấn	22520019	Đọc và tổng hợp các paper Tạo phát biểu bài toán và phạm vi dự án Làm sạch, tiền xử lý dữ liệu Viết báo cáo đồ án	100%

Phân công công việc

Thành viên	MSSV	Nhiệm vụ	Tiến độ hoàn thành
Nguyễn Hoàng Gia An	22520021	Khảo sát các bộ dữ liệu trên Kaggle/UCI Đọc và tổng hợp các paper Ghi chép các ràng buộc về dữ liệu Tìm hiểu kỹ thuật đặc trưng Làm slide đồ án	100%
Phạm Nguyên Anh	22520069	Đọc và tổng hợp các paper Thiết kế kiến trúc mô hình Cài đặt mô hình Cài đặt ứng dụng Kiểm tra và định dạng báo cáo, slide	100%
Nguyễn Gia Bảo	22520109	Đọc và tổng hợp các paper Thiết kế kiến trúc mô hình Huấn luyện và kiểm định mô hình Cài đặt tham số, đánh giá mô hình Kiểm tra và định dạng báo cáo, slide	100%

Mục lục

Bối cảnh

- Công trình xây dựng tiêu thụ lượng lớn năng lượng toàn cầu ($\approx 1/3$ tổng năng lượng toàn cầu).
- Là nguồn phát thải khí nhà kính (GHG) đáng kể, góp phần vào biến đổi khí hậu.
- Nhu cầu năng lượng cho HVAC (sưởi ấm, làm mát) chiếm phần lớn trong giai đoạn vận hành.

$\Rightarrow$ **Cần thiết phải tối ưu hóa thiết kế để giảm tác động môi trường.**

Giới thiệu vấn đề: **Heating & Cooling Load**

Heating Load (Y1)

Là lượng nhiệt cần bù đắp vào mùa lạnh để giữ ấm.
Cần bao nhiêu năng lượng để **sưởi ấm** tòa nhà?
Chịu ảnh hưởng bởi: Tổn thất nhiệt qua vỏ bao (tường, mái, kính), thông gió, xâm nhập khí lạnh.
Đơn vị đo: kW

Cooling Load (Y2)

Là lượng nhiệt cần hút ra vào mùa nóng (hoặc do nguồn nhiệt bên trong) để làm mát.
Cần bao nhiêu năng lượng để **làm mát** tòa nhà?
Chịu ảnh hưởng bởi: Thu nhiệt qua vỏ bao (bức xạ mặt trời qua kính, dẫn nhiệt qua tường/mái), nhiệt tỏa ra từ người, thiết bị, đèn, thông gió khí nóng.
Đơn vị đo: kW

Bộ dữ liệu: **Energy Efficiency Dataset**

- **Nguồn:** (Xifara & Tsanas, 2012) Mô phỏng bằng phần mềm Ecotect.
- **Quy mô:** 768 mẫu thiết kế nhà ở đa dạng.
- **Đầu vào:**
  - X1: Độ đặc tương đối (Relative Compactness)
  - X2: Tổng diện tích bề mặt (Surface Area)
  - X3: Diện tích tường (Wall Area)
  - X4: Diện tích mái (Roof Area)
  - X5: Chiều cao tổng thể (Overall Height)
  - X6: Hướng công trình (Orientation) (Quy ước 2/3/4/5 lần lượt là Nam/Bắc/Đông/Tây)
  - X7: Tỷ lệ diện tích kính (Glazing Area Ratio)
  - X8: Phân bố diện tích kính (Glazing Area Distribution)
- **Đầu ra:**
  - Tải nhiệt (Heating Load)
  - Tải lạnh (Cooling Load)

Khám phá dữ liệu (EDA)

Mục tiêu EDA:

- Kiểm tra chất lượng dữ liệu (thiếu, ngoại lệ).
- Hiểu phân bố của từng đặc trưng (Univariate Analysis).
- Khám phá mối quan hệ giữa các đặc trưng và giữa đặc trưng với biến mục tiêu (Multivariate Analysis).
- Đưa ra quyết định về tiền xử lý và lựa chọn đặc trưng.

Kiểm tra ban đầu:

- Đổi tên cột cho dễ hiểu.
- Kiểu dữ liệu: Chủ yếu là số thực (float), một vài cột số nguyên (int) cho X6, X8.
- Không có giá trị thiếu (Missing Values) trong toàn bộ dữ liệu.

Khám phá dữ liệu (EDA)

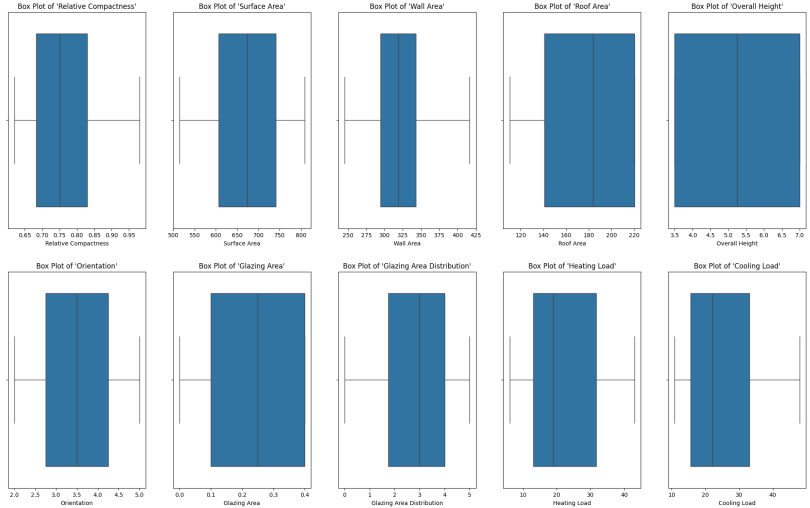
Thống kê mô tả:

- Giá trị trung bình (mean) và trung vị (median - 50 %) khá gần nhau ở nhiều cột $\rightarrow$ Ít bị lệch nhiều bởi ngoại lệ.
- Phạm vi giá trị (min/max) của các đặc trưng khác nhau $\rightarrow$ Cần Scaling.

Chuẩn bị dữ liệu:

- Loại bỏ các đặc trưng dư thừa, xử lý giá trị không hợp lệ nếu có.
- Phân tích mối quan hệ giữa biến đầu vào và đầu ra (correlation, scatterplot).
- Tạo biểu đồ boxplot để phát hiện ngoại lệ nếu cần.

Khám phá dữ liệu (EDA)

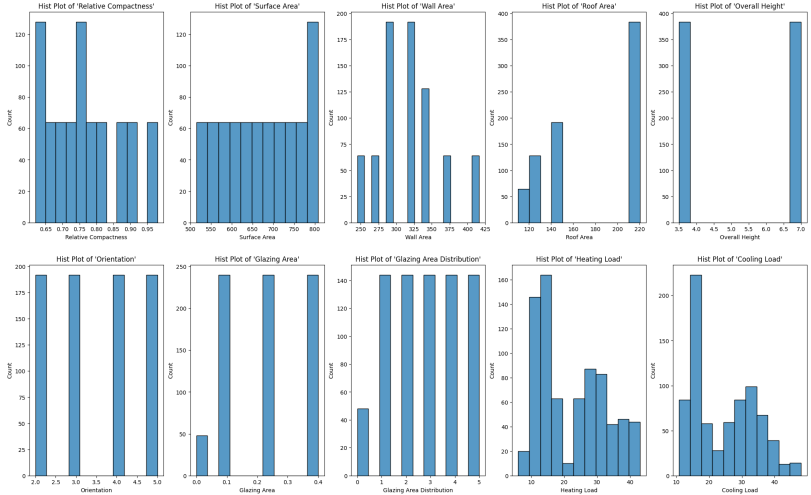


Boxplot

Khám phá dữ liệu (EDA)

Nhận xét: Biểu đồ hộp xác nhận phân bố tập trung hoặc rời rạc của nhiều đặc trưng đầu vào và sự đa dạng của tải năng lượng đầu ra. Quan trọng nhất, không phát hiện thấy điểm ngoại lệ (outliers) đáng kể nào trong bộ dữ liệu.
Kết luận: Dữ liệu có chất lượng tốt, không cần các bước xử lý ngoại lệ phức tạp.

Khám phá dữ liệu (EDA)



Histplot

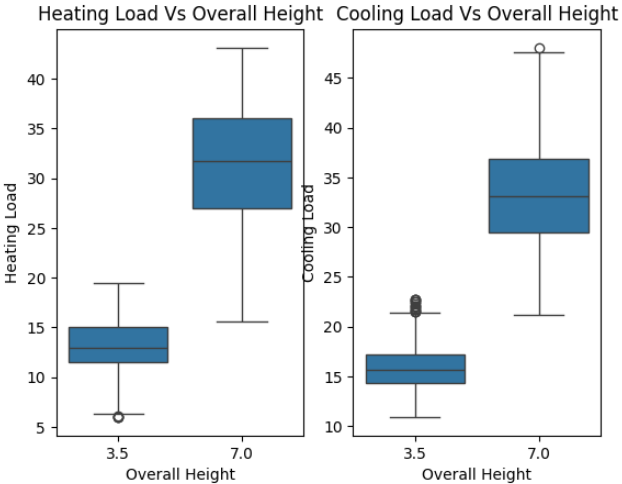
Khám phá dữ liệu (EDA)

Nhận xét: Histogram cho thấy nhiều đặc trưng đầu vào có các giá trị theo nhóm hoặc rời rạc (X1, X5, X6, X7, X8...). Tải nhiệt (Y1) và Tải lạnh (Y2) có phân bố tương tự nhau và hơi lệch phải, không tuân theo phân bố chuẩn.

Kết luận:

- Tính chất rời rạc/nhóm của nhiều đặc trưng đầu vào là kết quả của quá trình tạo dữ liệu mô phỏng.
- Sự lệch phải của Y1/Y2 có thể ảnh hưởng nhẹ đến một số mô hình hồi quy tuyến tính (vốn ưa chuộng phân bố chuẩn của phần dư), nhưng các mô hình dựa trên cây (DT, RF, XGBoost) thường ít bị ảnh hưởng bởi điều này hơn.
- Sự tương đồng phân bố của Y1 và Y2 tiếp tục củng cố việc có thể tập trung vào một biến mục tiêu.

Khám phá dữ liệu (EDA)



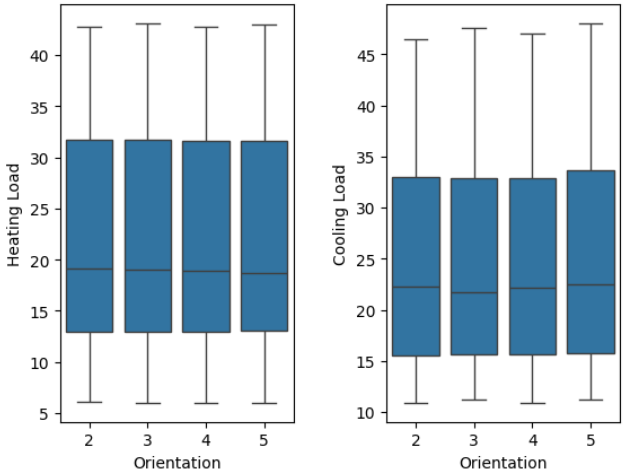
Phân bố Tải Nhiệt/Lạnh theo Chiều cao Tổng thể (X5)

Nhận xét:

- Chiều cao tòa nhà (X5) có ảnh hưởng rất lớn.
- Tòa nhà cao 7.0m yêu cầu Tải Nhiệt và Tải Lạnh trung bình cao hơn đáng kể so với tòa nhà cao 3.5m.

Kết luận: Chiều cao tổng thể (X5) là một trong những yếu tố hình học quan trọng nhất quyết định hiệu quả năng lượng của tòa nhà trong bộ dữ liệu này. Giảm chiều cao là một chiến lược hiệu quả để giảm nhu cầu năng lượng.

Khám phá dữ liệu (EDA)



Phân bố Tải Nhiệt/Lạnh theo Hướng Công trình (X6)

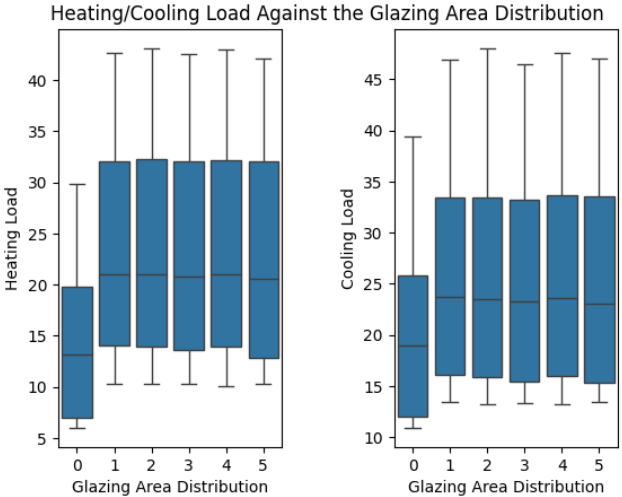
Khám phá dữ liệu (EDA)

Nhận xét:

- Hướng công trình (X6) hầu như không ảnh hưởng đến Tải Nhiệt và Tải Lạnh trong bộ dữ liệu này.
- Phân bố tải năng lượng là tương tự nhau cho cả 4 hướng.

Kết luận: Đặc trưng Hướng (X6) có thể được loại bỏ khỏi mô hình dự đoán mà không làm giảm đáng kể độ chính xác, giúp đơn giản hóa mô hình.

Khám phá dữ liệu (EDA)



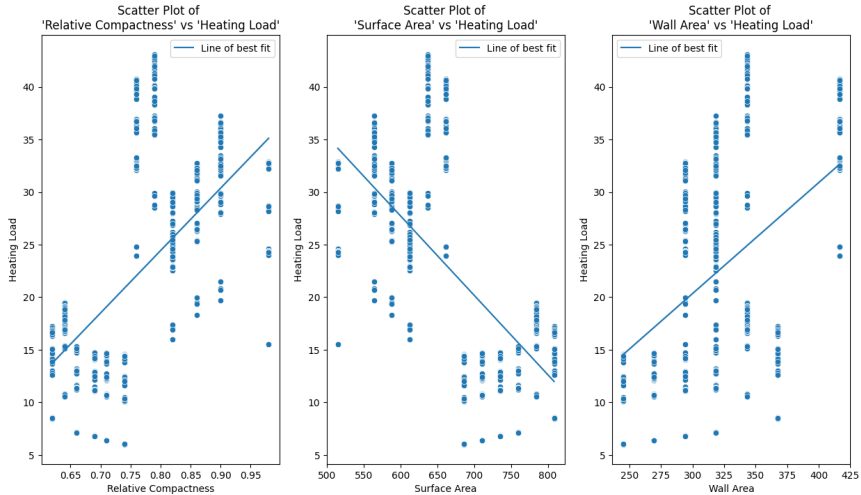
Phân bố Tải Nhiệt/Lạnh theo Phân bố Diện tích Kính (X8 - Giá trị gốc 0-5)

Nhận xét):

- Sự khác biệt lớn nhất về Tải Nhiệt/Lạnh là giữa nhóm không có kính (0) và các nhóm có kính (1-5).
- Sự khác biệt giữa các cách phân bố kính khác nhau (1-5) là không rõ rệt.

Kết luận: Việc có kính hay không có ảnh hưởng lớn hơn cách phân bố kính cụ thể. Binning X8 thành 0/1 là phù hợp.

Khám phá dữ liệu (EDA)



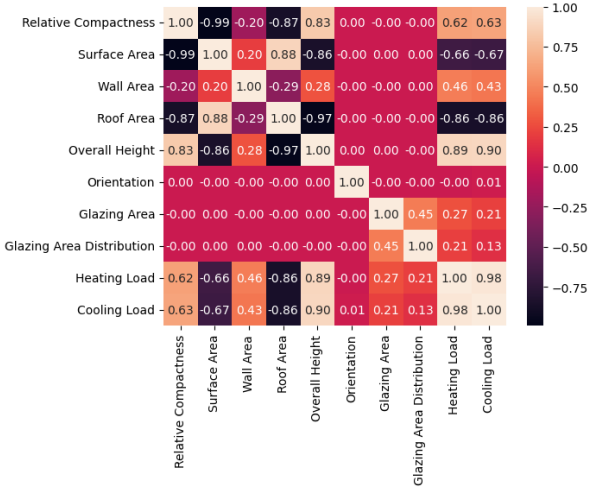
Mối quan hệ giữa các Đặc trưng Hình học và Tải Nhiệt (Y1)

Nhận xét:

- Tải nhiệt (Y1) giảm mạnh khi Độ đặc (X1) tăng và tăng khi Diện tích bề mặt (X2) tăng.
- Ảnh hưởng của Diện tích tường (X3) yếu hơn và phân tán hơn.
- Mối quan hệ không hoàn toàn tuyến tính.

Kết luận: Độ đặc (X1) và Diện tích bề mặt (X2) có mối liên hệ mạnh với Tải nhiệt, xác nhận kết quả từ ma trận tương quan. Sự phân tán lớn cho thấy mô hình tuyến tính đơn giản có thể không đủ để nắm bắt hết sự phức tạp; các mô hình phi tuyến (như cây quyết định) có thể phù hợp hơn.

Khám phá dữ liệu (EDA)



Ma trận tương quan

Khám phá dữ liệu (EDA)

Nhận xét:

- **Đa cộng tuyến:** X1 (Độ đặc) và X2 (Diện tích bề mặt) tương quan âm cực mạnh (-0.99).
- **Ảnh hưởng chính đến Tải Lạnh (Y2):** X5 (Chiều cao, +0.90), X4 (Diện tích mái, -0.86), X2 (Diện tích bề mặt, -0.67), X1 (Độ đặc, -0.63).
- **Ảnh hưởng yếu/không đáng kể (tuyến tính):** X6 (Hướng, $\approx$0), X8 (Phân bố kính), X7 (Diện tích kính).
- Y1 và Y2 tương quan rất cao (0.98).

Kết luận:

- Dữ liệu có cấu trúc tốt, phản ánh các nhóm thiết kế hình học khác nhau.
- Các đặc trưng hình học có mối liên hệ rõ ràng (tuyến tính và có thể phi tuyến) với tải năng lượng.
- X1, X5, X7 là các yếu tố quan trọng để dự đoán Y1/Y2.
- Sự tồn tại của đa cộng tuyến mạnh giữa X1 và X2 củng cố quyết định loại bỏ X2.
- Sự tương quan yếu của X6 với Y1/Y2 và các đặc trưng khác ủng hộ việc loại bỏ X6.
- Y1 và Y2 rất giống nhau, có thể tập trung dự đoán một trong hai.

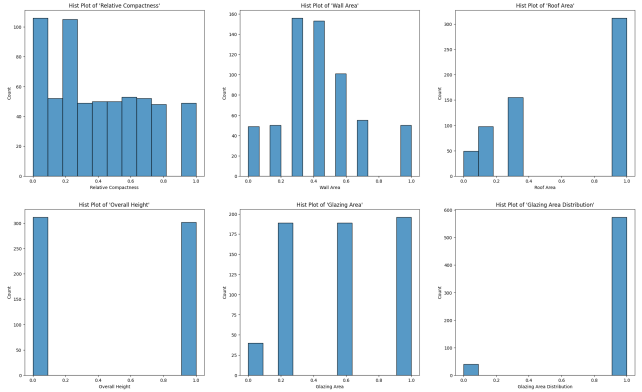
Kỹ thuật Đặc trưng (Feature Engineering)

Dựa trên Phân tích Dữ liệu (EDA):

- **Mục tiêu:** Tập trung dự đoán **Y2 (Tải lạnh)**.
- **Loại bỏ Đặc trưng:** X2 (Surface Area - đa cộng tuyến), X6 (Orientation - ít ảnh hưởng).
- **Biến đổi X8:** Đơn giản hóa thành **0 (Không kính)** và **1 (Có kính)**.
- **Scaling:** Áp dụng **MinMaxScaler** cho 6 đặc trưng còn lại (X1, X3, X4, X5, X7, X8).
- **Chia dữ liệu:** Train/Test (80/20).

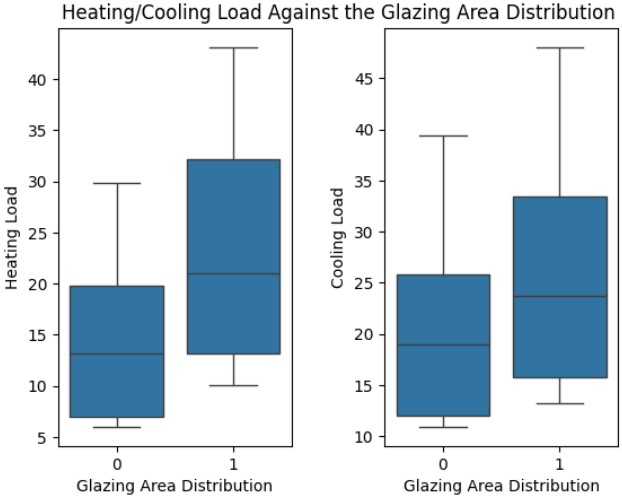
$\Rightarrow$ Dữ liệu sẵn sàng cho mô hình hóa với 6 đặc trưng đã chuẩn hóa.

Kỹ thuật Đặc trưng (Feature Engineering)



Histplot các cột X sau khi tiền xử lý

Kỹ thuật Đặc trưng (Feature Engineering)



Phân bố Tải Nhiệt/Lạnh theo Phân bố Diện tích Kính (Sau binning 0-1)

Xây dựng mô hình dự đoán

- Tinh chỉnh tham số bằng **GridSearchCV**.
- Thử nghiệm nhiều thuật toán hồi quy:
  - Linear Regression
  - SVM
  - K-Nearest Neighbor
  - Decision Tree
  - Random Forest
  - XGBoost

Xây dựng mô hình dự đoán: **Học chủ động**

- **Định nghĩa**:
  - Chiến lược nâng cao hiệu suất mô hình với dữ liệu gắn nhãn hạn chế.
  - Mô hình chọn mẫu **thông tin nhất** hoặc **không chắc chắn nhất** để gắn nhãn từ chuyên gia.
- **Lợi ích**:
  - Giảm **chi phí** và **thời gian** gắn nhãn.
  - Tập trung vào dữ liệu giá trị cao, cải thiện độ chính xác.
- **Nguyên lý hoạt động**:
  - Xác định mẫu khó dự đoán (dựa trên độ không chắc chắn).
  - Sử dụng các chiến lược truy vấn như:
    - *Uncertainty sampling*: Chọn mẫu có độ không chắc chắn cao nhất.
    - *Committee inquiry*: Dựa trên sự bất đồng giữa các mô hình trong một ủy ban.
    - *Information saturation sampling*: Chọn mẫu tối đa hóa thông tin thu được.

Xây dựng mô hình dự đoán: **Học chủ động**

- **Mô phỏng Học Chủ Động**:
  - Sử dụng nhãn **ground-truth** từ tập kiểm tra thay cho chuyên gia.
  - Chọn mẫu dựa trên **độ không chắc chắn**, thêm vào tập huấn luyện qua từng vòng lặp.
- **Chiến lược truy vấn**:
  - Sử dụng **Error-Based Sampling** (thuộc Uncertainty Sampling):
    - Mô hình xác định mẫu có **lỗi dự đoán cao nhất** (dựa trên bình phương phần dư - squared residuals).
    - Các mẫu này được thêm vào tập huấn luyện, và mô hình được huấn luyện lại.
    - Áp dụng cho **tất cả các baseline models**.
- **Mục tiêu**:
  - Khám phá lợi ích AL trong cải thiện hiệu suất.

Phân tích Độ quan trọng Đặc trưng (Feature Importance)

Mục tiêu: Xác định các đặc trưng đầu vào có ảnh hưởng lớn nhất đến dự đoán Tải lạnh của mô hình.

Nhận xét chính:

- **Độ đặc tương đối (X1):** Là đặc trưng có **ảnh hưởng lớn nhất** trong mô hình RF.
- **Chiều cao tổng thể (X5):** Là đặc trưng **quan trọng nhất** trong mô hình XGB, DT và cũng quan trọng trong RF.
- **Các đặc trưng khác:**
  - **Diện tích tường (X3)**, **Diện tích mái (X4)** có ảnh hưởng vừa phải.
  - **Tỷ lệ kính (X7)**, **Phân bố kính (X8 - sau binning)** có ảnh hưởng thấp hơn.

Phân tích Độ quan trọng Đặc trưng (Feature Importance)

Kết luận từ Feature Importance

- **X1 (Độ đặc)** và **X5 (Chiều cao)** là hai yếu tố hình học then chốt ảnh hưởng đến hiệu quả năng lượng và tác động môi trường vận hành.
- Việc tập trung tối ưu hóa các đặc trưng này trong giai đoạn thiết kế sớm có thể mang lại hiệu quả giảm thiểu năng lượng đáng kể.

Kết quả dự đoán: **Baseline Model**

Đánh giá (Metrics): Sử dụng **MSE, MAE, RMSE, R^2**

	MAE	MSE	RMSE	R^2
XGBoost	1.152116	2.952587	1.718309	0.968134
Decision Tree	1.165358	3.055920	1.748119	0.967019
Random Forest	1.169375	3.058888	1.748968	0.966987
K-Nearest Neighbor	1.214726	3.342413	1.828227	0.963927
SVM	1.637182	7.093737	2.663407	0.923441
Linear Regression	2.187545	9.650917	3.106592	0.895843

Kết quả Baseline Models

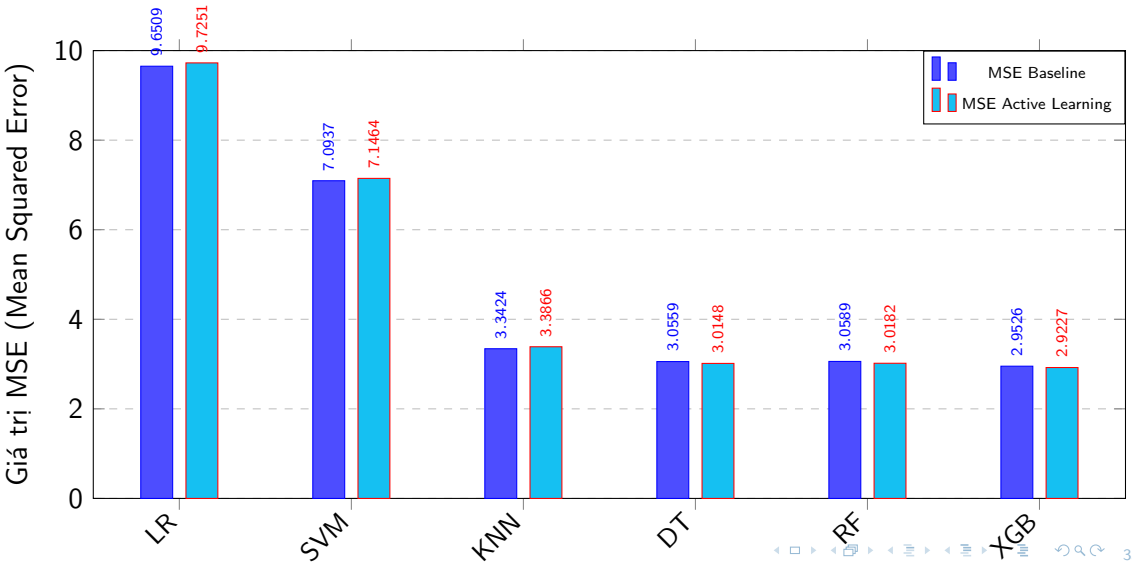
Kết quả dự đoán: **Active-Learning Model**

Đánh giá (Metrics): Sử dụng **MSE, MAE, RMSE, R^2**

	MAE	MSE	RMSE	R^2
XGBoost	1.148804	2.922741	1.709603	0.968456
Decision Tree	1.159368	3.014823	1.736325	0.967463
Random Forest	1.163210	3.018240	1.737308	0.967426
K-Nearest Neighbor	1.207157	3.386573	1.840264	0.963451
SVM	1.639293	7.146387	2.673273	0.922873
Linear Regression	2.192641	9.725087	3.118507	0.895042

Kết quả Active-Learning Models

So sánh hiệu năng: **Chỉ số MSE**

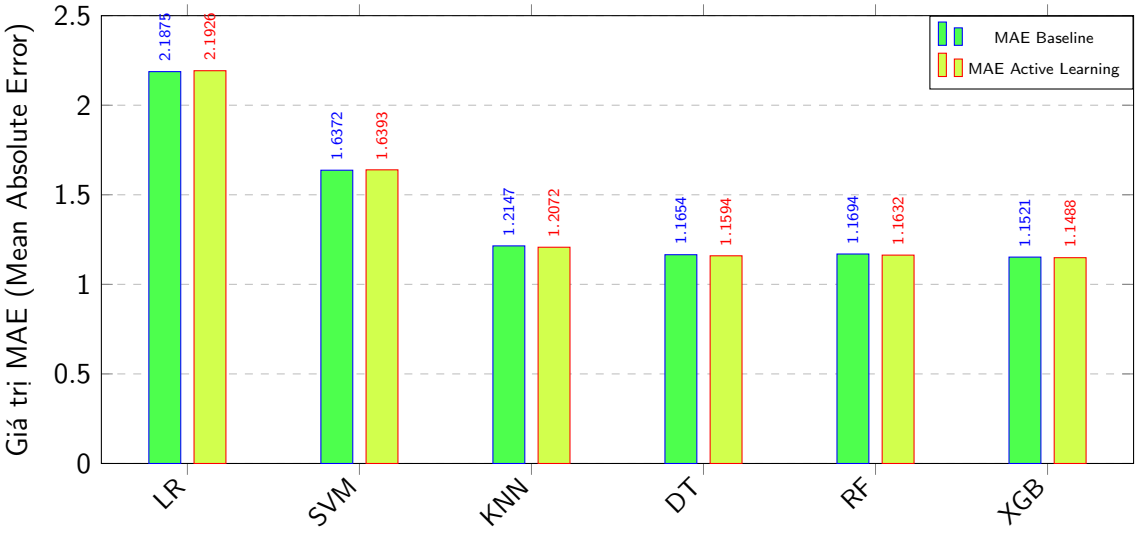


So sánh hiệu năng: **Nhận xét chỉ số MSE**

Nhận xét:

- Mô hình XGBoost và Decision Tree cho thấy MSE giảm nhẹ sau khi áp dụng Học Chủ động, thể hiện lỗi dự đoán trung bình bình phương nhỏ hơn.
- Các mô hình khác có sự thay đổi MSE không đáng kể.
- MSE thấp hơn thể hiện mô hình dự đoán chính xác hơn.

So sánh hiệu năng: **Chỉ số MAE**

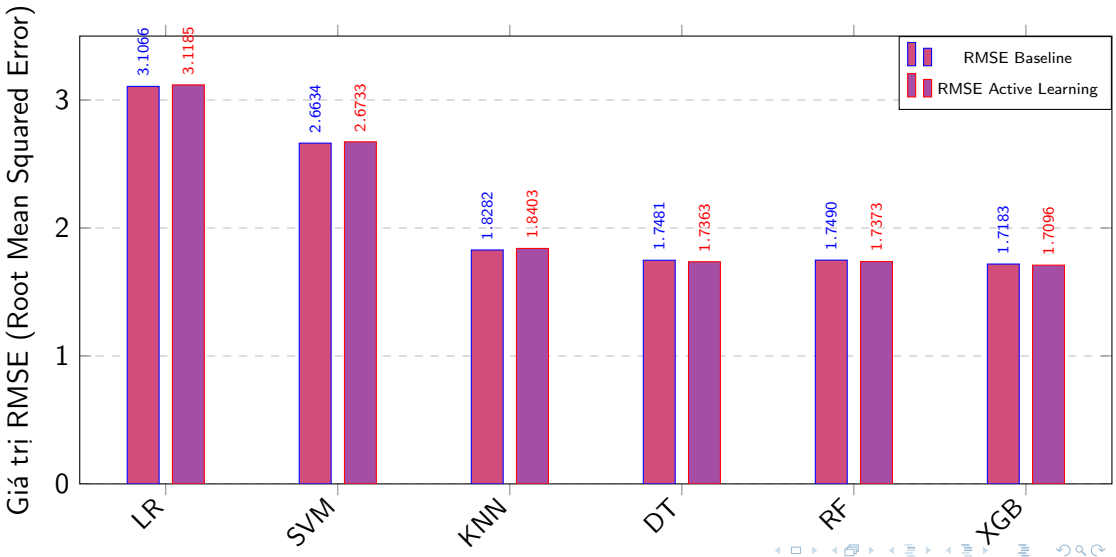


So sánh hiệu năng: **Nhận xét chỉ số MAE**

Nhận xét:

- Mô hình XGBoost, Decision Tree và Random Forest có MAE giảm nhẹ sau Học Chủ động, cho thấy sai số tuyệt đối trung bình nhỏ hơn.
- K-Nearest Neighbors cũng có cải thiện MAE.
- MAE thấp hơn là tốt hơn.

So sánh hiệu năng: **Chỉ số RMSE**

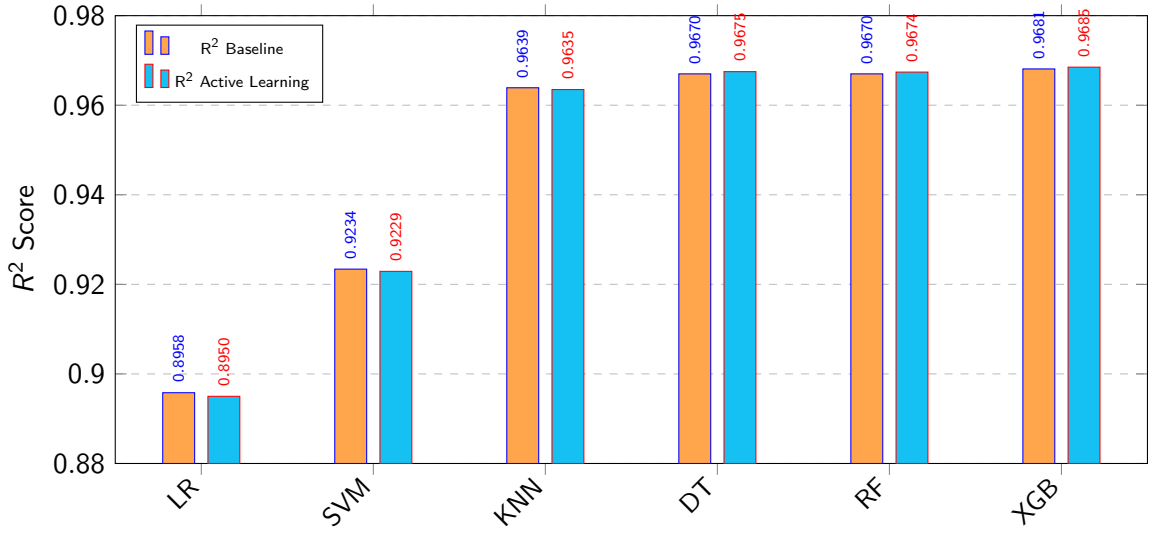


So sánh hiệu năng: **Nhận xét chỉ số RMSE**

Nhận xét:

- Tương tự MSE, XGBoost và Decision Tree cho thấy RMSE giảm nhẹ với Học Chủ động.
- RMSE có cùng đơn vị với biến mục tiêu, giúp diễn giải lỗi dễ hơn MSE. RMSE thấp hơn là tốt hơn.

So sánh hiệu năng: R^2 Score

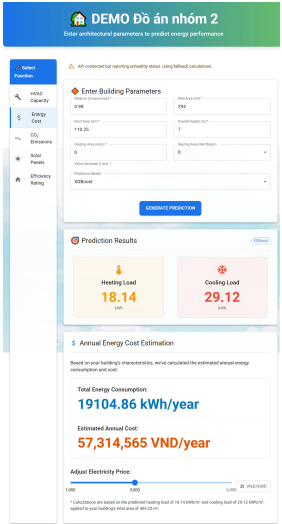


So sánh hiệu năng **Nhận xét chỉ số R^2**

Nhận xét:

- Mô hình XGBoost và Decision Tree cho thấy sự cải thiện nhẹ về chỉ số R^2 sau khi áp dụng Học Chủ động.
- Các mô hình Random Forest, K-Nearest Neighbors, SVM và Linear Regression có sự thay đổi không đáng kể hoặc giảm nhẹ về R^2.
- Nhìn chung, các mô hình dựa trên cây (XGBoost, Decision Tree, Random Forest) đạt R^2 cao nhất, cho thấy khả năng giải thích phương sai tốt của dữ liệu.

Xây dựng ứng dụng



Tính năng chính:

Dự đoán Năng lượng:

- Nhập thông số hình học.
- Dự đoán Tải nhiệt & Tải lạnh.

Ước tính Chi phí Vận hành.

Phân tích Môi trường & Giải pháp:

- Ước tính Phát thải CO_2.
- Đề xuất Pin Mặt trời.

$\Rightarrow$ **Mục tiêu:** Cung cấp giải pháp **nhanh chóng, trực quan** để tối ưu hóa thiết kế tòa nhà theo hướng hiệu quả năng lượng và bền vững môi trường.

Demo

Kết luận và Đề xuất

Kết luận:

- Mô hình XGBoost cho kết quả tốt nhất với $R^2 = 0.968$.
- Học chủ động (Active Learning) giúp tăng hiệu quả khi dữ liệu nhãn khan hiếm.

Hướng phát triển:

- Mở rộng mô hình cho các loại công trình khác nhau.
- Kết hợp thêm yếu tố chi phí và vật liệu để khuyến nghị tối ưu hơn.

References

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015).

Xgboost: extreme gradient boosting.

R package version 0.4-2, 1(4):1–4.

Mahmood, S., Sun, H., Ali Alhussan, A., Iqbal, A., and El-Kenawy, E.-S. M. (2024).

Active learning-based machine learning approach for enhancing environmental sustainability in green building energy consumption.

Scientific Reports, 14(1):19894.

[Mahmood et al., 2024] [Chen et al., 2015]

The End