

Executive Summary

This project examines energy efficiency in machine learning using datasets from the UCI Machine Learning Repository to predict sentiment mapping. Using three regression models: Random Forest Regression (RF), K-Nearest Neighbors (KNN), and Support Vector Regression (SVR), we compared their performance and energy efficiency. Our findings show that while more complex models often deliver higher accuracy, they also consume significantly more energy. In contrast, simpler and well-optimized models can achieve comparable performance with lower energy costs. This highlights the importance of considering energy efficiency when developing and deploying machine learning systems.



Background

Introduction

As energy efficiency becomes a growing priority, reducing the heating and cooling demands of residential buildings is an important engineering challenge. Heating load (HL) and cooling load (CL) make up a large portion of a building's energy consumption, but traditional methods for calculating them, such as detailed simulation tools—can be time-consuming and complex. This creates a need for faster, more accessible prediction methods. The client's goal is to use key building design variables, such as surface area and glazing, to accurately estimate energy loads using statistical and machine learning techniques. This approach supports broader efforts to improve sustainable building design while saving time and resources in engineering practice.

Goal

The goal of this project is to develop and evaluate machine learning models that can accurately predict heating load (HL) and cooling load (CL) of residential buildings using key design variables. The project aims to identify input features have the strongest influence on energy demand and to compare the performance of different modeling approaches in terms of prediction accuracy and efficiency.

Experimental Setup

Data Source Identification

The data set includes 768 unique building shapes and conditions as designed by the study creators in Ecotest.

Each entry includes 8 predictors, covering glazing area, the glazing area distribution, and the orientation.

The dependent variable is Cooling Load, enabling analysis of how factors relate to energy efficiency.

Data Pre-Processing

Multicollinearity: having multiple predictors correlated, reduce to see individual factor effects.

Encoding: any non-numeric value was previously encoded by the study creators.

Scaling: each factor was scaled to ensure balanced affect across the various ranges.

Experimental Test Plan

The experimental test plan focuses on comparing the predictive performance of three machine learning algorithms—Random Forest, K-Nearest Neighbors (KNN), and Support Vector Regression (SVR)—using the UCI Machine Learning Repository Energy Efficiency dataset.

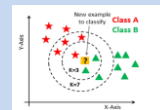
Random Forest

Random Forest creates multiple decision trees using random subsets of data and features, then averages their outputs. This reduces overfitting and improves prediction accuracy.



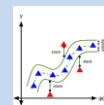
KNN

KNN predicts values based on the average of the k nearest data points, using distance measures like Euclidean distance. It relies on data similarity and has no explicit training phase.



SVR

SVR fits a function within a defined error margin (ϵ), using only key data points (support vectors). Kernel functions allow it to model nonlinear relationships.



Experimental Results

Final Model Performance

Table 1: Final Model Performance Comparison

Model	MAE	RMSE	R^2	Adj_ R^2
Random Forest (RF)	1.0604	1.7129	0.9683	0.9666
KNN	1.7048	2.6033	0.9269	0.9228
SVR	1.1369	1.7865	0.9656	0.9637

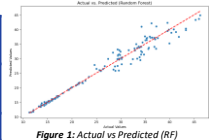


Figure 1: Actual vs Predicted (RF)

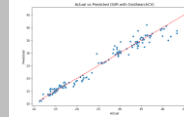


Figure 4: Actual vs Predicted (SVR)

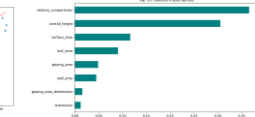


Figure 3: Feature Importance (RF)

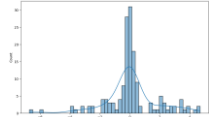


Figure 2: Residuals Distribution (RF)

Conclusions

Highlights

This study investigated the relationship between residential building design variables and energy loads, specifically heating load (HL) and cooling load (CL), using statistical and machine learning techniques. Our analysis identified the variables most strongly associated with energy demand, providing insight into key factors that influence building efficiency. The predictive models developed were evaluated for accuracy and efficiency, with performance metrics indicating that **Random Forest classification, with an R^2 of 0.9683 and an RMSE of 1.7129, achieved the best balance** between prediction accuracy and computational cost. These findings suggest that machine learning can be a practical and reliable tool for estimating building energy loads, supporting more efficient and sustainable building design.

Future Work

- Future work could incorporate additional datasets featuring diverse climate zones, occupant behavior patterns, and building materials to enhance model generalization
- Further improvements could include engineering new features like building orientation or window-to-wall ratios, alongside testing advanced deep learning architectures for higher prediction accuracy

References

- Liu, S., and Bobrow, J. E., "An Analysis of a Pneumatic Servo System and Its Application to a Computer-Controlled Robot," *ASME Journal of Dynamic Systems, Measurement, and Control*, 1988, Vol 110 pp 228-235.
- Al-arzouqi et al., "Generative AI in Creative Industries," *Studies in Computational Intelligence*, 2025, Vol 1208.
- IT Paragon. (n.d.). *Support vector machine regression*. Medium. <https://medium.com/it-paragon/support-vector-machine-regression-cf65348b6345>
- Plain English AI. (2021, March 3). *Introduction to k-nearest neighbors (KNN) algorithm*. Medium. <https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8>
- Günay, D. (n.d.). *Random forest*. Medium. <https://medium.com/@denizgunay/random-forest-af5bde5d7e1e>