SEA600

INTRODUCTION TO MACHINE LEARNING

ASSIGNMENT #1 - REPORT

7 MARCH' 2024

BY

ARNAV NIGAM

CONTENTS

1. Introduction

It is widely understood that sustainability is crucial in today's world. Therefore, improving energy efficiency in the construction and design of buildings is more important than ever before. Buildings consume a significant portion of the world's energy and contribute significantly to carbon emissions. Therefore, improving building energy efficiency is not only an environmental need but also a social and economic responsibility.

This report is based on the analysis of the "energy_efficiency.csv" dataset, which includes a vast amount of data points. The dataset illustrates the complex relationship between building design parameters and energy consumption for heating and cooling. The dataset contains 768 samples and 8 architectural features, including relative compactness, surface and wall area, roof area, overall height, orientation, and glazing attributes. This dataset is ideal for exploring how different design elements can impact a building's thermal performance.

Our assignment entails a multifaceted approach to analyzing the dataset and deriving meaningful insights. We will begin by exploring baseline regression techniques, such as Linear Regression and Lasso Regression, to establish a foundational understanding of the dataset's characteristics and predictive capabilities. Subsequently, we will delve into advanced modeling approaches, including Decision Tree Regression and K-Nearest Neighbors Regression, to unlock deeper insights and achieve enhanced predictive accuracy. Furthermore, we will explore feature engineering methods to preprocess the dataset and optimize model performance. Finally, we will conduct hyperparameter tuning to fine-tune our models and evaluate their efficacy in predicting heating load demands. Through these comprehensive analyses, we aim to provide actionable recommendations for designing energy-efficient buildings and advancing sustainable architectural practices.

2. Objective

The purpose of this report is to identify patterns and generate insights that can be used to create smarter and more sustainable building designs. By predicting heating demands using specific building characteristics, we hope to provide actionable strategies to reduce energy consumption, which will contribute to environmental sustainability, cost savings, and improved occupant comfort, and well-being.

It is essential to recognize that the implications of our findings extend beyond design and construction. They address significant societal issues and underscore the importance of data and analytics in creating a more sustainable future. This paper aims to clarify the relationship between building design and energy efficiency and stimulate innovation and policy-making that promotes sustainable development.

The primary objective of analyzing the "energy_efficiency.csv" dataset is to identify and understand the factors that significantly impact building energy efficiency. By analyzing these elements, we aim to accurately forecast the energy requirements for heating buildings. Finally, our research aims to advise and assist in the construction of more energy-efficient buildings, reducing energy use while minimizing environmental impact.

This translates into a supervised regression machine learning task, to predict continuous values—specifically, the building's heating load (energy consumption). These energy needs (dependent variables) are predicted using dataset properties (independent variables) such as building size, orientation, window area, and insulation levels.

3. Implementation Constraints

Computational resources and time constraints may limit the complexity of models that may be trained and tested. More complicated models, such as deep learning, may demand significant processing power and time. The volume and quality of data available have a substantial impact on the model's performance. Incomplete or noisy data can result in inaccurate forecasts.

Increasing energy efficiency in buildings reduces energy consumption, which is critical for environmental sustainability and carbon footprint reductions. Prioritizing models that can identify the most effective energy-saving methods can help to concentrate efforts on regions with the highest environmental benefit. The costs of installing energy-efficient technologies and designs can influence adoption rates. Improved energy efficiency is frequently associated with higher indoor air quality and thermal comfort, which affects occupant health and productivity.

4. About Dataset

The dataset includes information about numerous building attributes, with a focus on elements that influence energy efficiency. The data captures a wide range of architectural and design features, emphasizing their impact on energy use for heating and cooling.

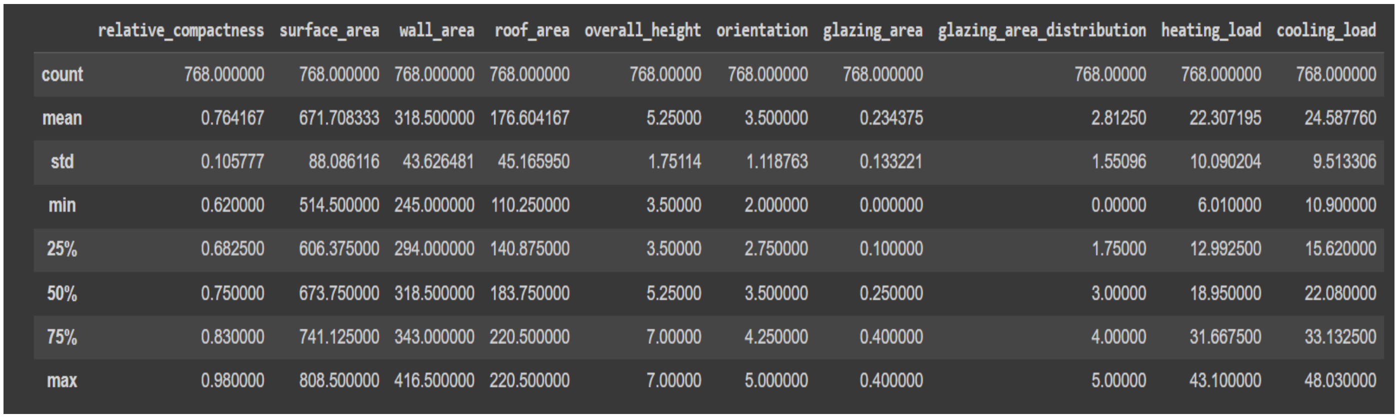
Below are the key details of the dataset:

- **Number of Samples**: The dataset has 768 observations. Each observation relates to a separate building or simulated scenario that reflects distinct combinations of the features examined.
- **Number of Features**: There are 8 features in all, each reflecting a different aspect of building design that is known to influence energy efficiency. These elements provide a detailed picture of the building's physical attributes as well as potential energy demands.
- **Features**: This includes relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and glazing area distribution.
- **There is only one target variable chosen which is the heating load**.

This dataset is a valuable resource for studying and predicting how different design choices affect a building's energy efficiency. The purpose of analyzing these variables is to discover trends and insights that can lead to more energy-efficient building designs, hence lowering energy consumption and contributing to environmental sustainability.

The study of such a collection not only achieves the primary goal of improving building energy efficiency, but it also fits with broader societal and environmental goals. It establishes a data-driven framework for making sound judgements about sustainable architecture and building practices.

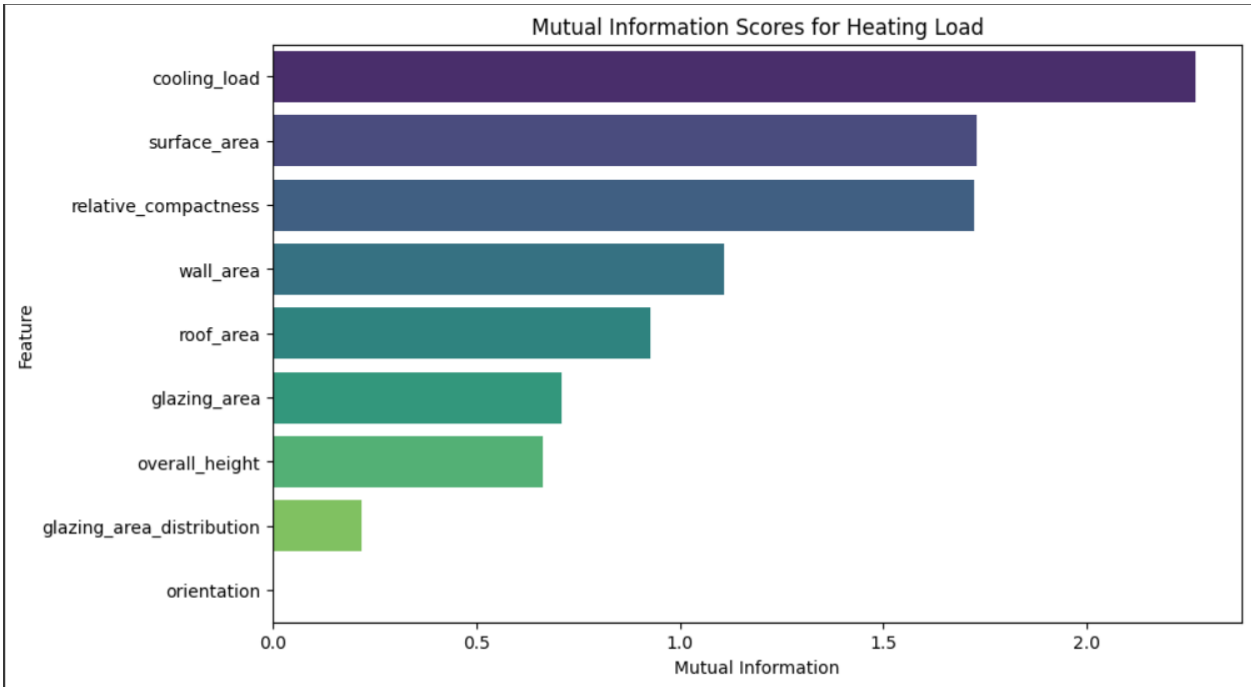
There were no missing values in the dataset, which had 768 samples and 8 characteristics. This indicates a well-curated dataset in which all data points were carefully collected or imputed, resulting in a comprehensive dataset ready for analysis. A first statistical assessment and visualization of the dataset revealed that all values are within predicted ranges, implying that any outliers are actual observations rather than errors or anomalies. Given the dataset's technical character, with each feature representing architectural and design variables, the occurrence of such data points could indicate unique design choices rather than errors in data collection or entry. Given these considerations, it was determined that the dataset was of excellent quality and ready for further analysis, eliminating the necessity for preliminary data cleaning. This scenario emphasizes the need for data gathering and curation techniques, which greatly speed up future phases of data analysis and model creation. Proceeding directly to the analysis step without first cleaning saves time while also lowering the possibility of adding bias or errors through excessive data manipulation. The surface area of the dataset has only 12 unique values, which means that there are only 12 different building shapes. All these 12 building shapes can have any of the 7 Wall Areas and 4 Roof Areas amongst other properties. Different combinations of these unique variables gave rise to the 768 examples. train_test_split from sklearn.model_selection plays a crucial role in preparing the dataset for machine learning tasks. By splitting the data into separate training and testing subsets, it enables robust model evaluation and validation. With this approach, the algorithm can learn patterns and relationships from the training data while retaining unseen data for evaluation, ensuring the model's generalization capability. Employing an 80-20 split ratio for training and testing, respectively, and utilizing a fixed random state for reproducibility, it ensures consistency in model evaluation.



	relative_compactness	surface_area	wall_area	roof_area	overall_height	orientation	glazing_area	glazing_area_distribution	heating_load	cooling_load
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375	2.81250	22.307195	24.587760
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221	1.55096	10.090204	9.513306
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000	0.00000	6.010000	10.900000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000	1.75000	12.992500	15.620000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000	3.00000	18.950000	22.080000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000	4.00000	31.667500	33.132500
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000	5.00000	43.100000	48.030000

5. Baseline methods

This step involved employing two fundamental regression techniques: Linear Regression and Lasso Regression. The rationale behind choosing these models as baselines lies in their widespread use in regression analysis, driven by their simplicity and interpretability. For the Lasso Regression model applied to predict the heating load, the mean root mean squared error (RMSE) across 5-fold cross-validation was found to be 2.1183. The model exhibited a memory usage of 299.68 MB and a training time of 0.0738 seconds. Similarly, the Linear Regression model showcased a mean RMSE of 1.7876, with a memory utilization of 299.68 MB and a training time of 0.0402 seconds. These documented results serve as a foundational reference point for subsequent stages of our project. They provide a clear understanding of the baseline performance and resource utilization associated with the chosen regression models.



The bar graph depicts the mutual information scores between the features and the target variable (heating load) in the dataset. Mutual information measures the dependency between two variables, indicating how much information about one variable can be obtained from another. Each bar represents the mutual information score of a feature for the heating load. Features with higher mutual information scores are considered to have stronger relationships with the target variable.

We also found overfitting in initial training examples for all the models, but later on all the models showed significant test score as the number of training examples increased.

6. Advanced Models

We explored various modelling techniques and evaluated their efficacy against our established baselines. Beyond Linear Regression and Lasso Regression, we introduced two other prominent regression algorithms: Decision Tree Regression and K-nearest neighbours Regression. The Linear Regression model yielded an MSE of 3.2827, with memory usage and training time recorded at 302.71 MB and 0.0037 seconds, respectively. Similarly, the Lasso Regression model exhibited an MSE of 4.7825, with comparable memory usage but slightly longer training time, at 0.0051 seconds.

Interestingly, the Decision Tree Regression model demonstrated a notably lower MSE of 2.1773, indicating superior predictive accuracy compared to the previous models. Despite its improved performance, this model incurred a marginally longer training time of 0.0069 seconds. Conversely, the K-Nearest Neighbors Regression model achieved an MSE of 6.0867 with a relatively shorter training time of 0.0039 seconds. The space used by all the models was almost equal. We chose these two regressors because Decision Tree and KNN are known for their simplicity and ease of interpretation compared to more complex models like neural networks or ensemble methods. Also, Decision Tree Regression can capture nonlinear relationships between features and the target variable. Similarly, KNN Regression is inherently non-parametric and flexible, making it adept at handling non-linear patterns in the data. Overall, the selection of Decision Tree Regression and KNN Regression was driven by a combination of their simplicity, non-linear modelling capabilities, robustness to outliers, observed performance, computational efficiency, and the desire to diversify modelling approaches. These models were deemed well-suited for our predictive task, offering a balanced trade-off between predictive accuracy, interpretability, and computational efficiency.

7. Feature engineering methods

We introduced manual feature engineering techniques, particularly feature scaling, to preprocess the dataset before training our regression models. Our analysis involved training Linear Regression, Lasso Regression, Decision Tree Regression, and K-nearest neighbours Regression models on the scaled training data and evaluating their performance on the scaled validation set.

We selected feature scaling as a preprocessing technique due to its ability to normalize the range of independent variables within the dataset, ensuring that features with larger scales do not dominate the learning process. This normalization facilitates the convergence of optimization algorithms and enhances the performance of machine learning models, particularly those sensitive to the scale of input features, such as gradient-based methods.

Upon analysis of the output, we observed significant differences in the performance metrics compared to the previous milestone. Specifically, for Linear Regression, the Mean Squared Error (MSE) on the validation set decreased to 3.7529. Conversely, Lasso Regression's MSE increased to 6.2612, indicating potential sensitivity to feature scaling. Decision Tree Regression exhibited a modest improvement, with an MSE of 0.8083, suggesting enhanced performance post-feature scaling. Notably, K-Nearest Neighbors Regression showcased a notable decrease in MSE to 1.7765, indicative of improved predictive accuracy facilitated by feature scaling.

Overall, the adoption of feature scaling as a preprocessing technique proved beneficial in our project by improving the convergence behaviour of regression models and refining their predictive accuracy. This underscores the importance of thoughtful feature engineering strategies in optimizing model performance and achieving reliable predictions in machine learning tasks.

8. Hyperparameter tuning

Following the application of feature scaling to our dataset, we proceeded to optimize our regression model using hyperparameter tuning. Leveraging the capabilities of Scikit-learn's GridSearchCV, we used GridSearchCV because it can automatically search from the best set of hyperparameters and can find the best combination according to our models and also uses cross-validation, we systematically explored various combinations of hyperparameters to identify the optimal configuration for our Decision Tree Regression model. The optimal hyperparameters for our Decision Tree Regression model were determined to be a max_depth of 7, a min_samples_leaf of 1, and a min_samples_split of 2. Notably, this configuration resulted in a significantly reduced MSE of approximately 0.27760 on the validation set, demonstrating substantial improvement over the advanced model.

The decision to adopt the Decision Tree Regression model was justified by its performance, as evidenced by the lowest MSE achieved during hyperparameter tuning. This outcome underscores the effectiveness of feature scaling in enhancing model performance and optimizing predictive accuracy. Through hyperparameter optimization, we were able to unlock the full potential of the Decision Tree Regression model and achieve superior predictive performance compared to other regression models considered in our analysis.

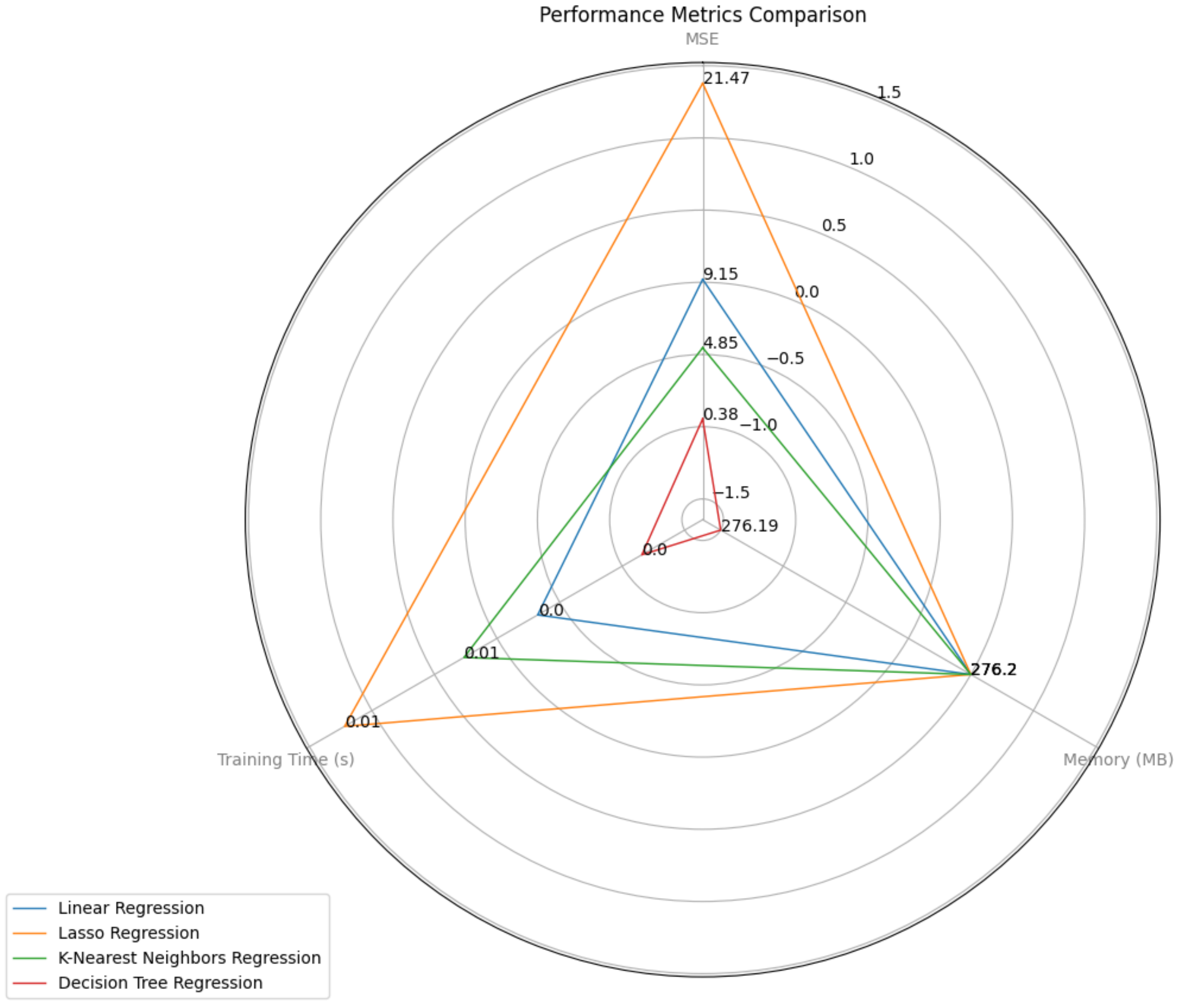
For the prediction of the Heating Load, the optimal decision tree has a maximum depth of 7. This suggests that a deeper tree, capable of capturing more intricate patterns within the data, yields the best predictive performance.

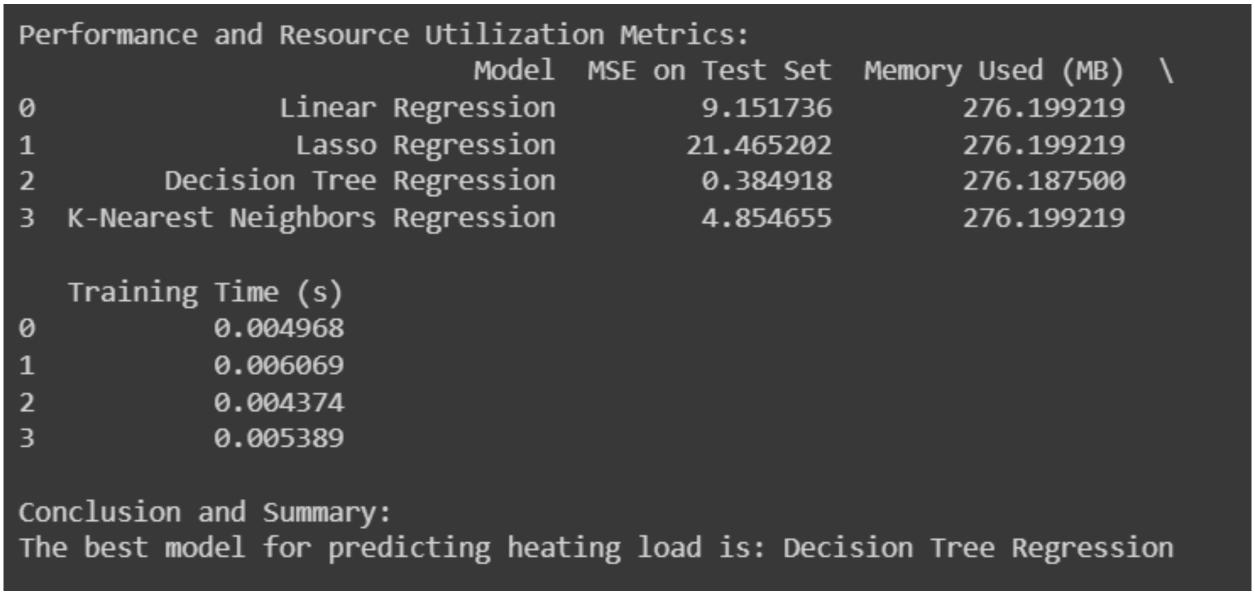
- The parameter 'min_samples_leaf' specifies the minimum number of samples required to be at a leaf node. With a value of 1 for the Heating Load, finer splits are favoured, allowing the model to capture more detailed nuances in the dataset.

- 'min_samples_split' denotes the minimum number of samples required to split an internal node. A value of 2 indicates a preference for splitting even smaller subsets, allowing the tree to make more nuanced decisions and potentially capture finer details in the data, leading to improved predictive accuracy.

Then, the decision tree model is trained and evaluated for predicting the heating load, and the MSE, memory usage, and training time are printed to assess the model's performance. The MSE on the test set is 0.3849, indicating the average squared difference between the predicted and actual heating load values. The memory usage during training is 276.1875 MB, and the training time is 0.0044 seconds.

Finally, we are comparing all the models again but this time on the test set. From the below graph and the table, we can understand the comparison better. (Based on most efficient and best outputs recorded during evaluation)





```
Performance and Resource Utilization Metrics:
                            Model  MSE on Test Set  Memory Used (MB)  \
0               Linear Regression         9.151736        276.199219   
1                Lasso Regression        21.465202        276.199219   
2        Decision Tree Regression         0.384918        276.187500   
3  K-Nearest Neighbors Regression         4.854655        276.199219   

   Training Time (s)  
0           0.004968  
1           0.006069  
2           0.004374  
3           0.005389  

Conclusion and Summary:
The best model for predicting heating load is: Decision Tree Regression
```

8. Conclusion

The exploration of the "energy_efficiency.csv" dataset offers actionable pathways toward reducing energy consumption while promoting environmental sustainability, cost savings, and improved occupant comfort. By accurately predicting heating load demands based on building characteristics, stakeholders can implement targeted strategies to optimize energy usage. Incorporating energy-efficient design elements not only minimizes the carbon footprint associated with building operations but also results in cost savings for building owners and occupants. Additionally, energy-efficient buildings are often associated with enhanced occupant comfort and well-being, contributing to increased productivity and satisfaction among occupants. Overall, the findings of this study underscore the importance of data-driven approaches in achieving sustainability goals and optimizing building performance.

Feature engineering techniques, such as feature scaling, played a crucial role in enhancing model convergence and refining predictive accuracy. This preprocessing step significantly improved the performance of regression models. Additionally, hyperparameter tuning further optimized the Decision Tree Regression model. The comprehensive comparative analysis of all models on the test set underscored the superior predictive performance of the Decision Tree Regression model, particularly after hyperparameter optimization. This study emphasizes the significance of data-driven approaches in informing sustainable architectural practices and addressing societal and environmental challenges. By leveraging advanced machine learning techniques, the research provides actionable insights for designing more energy-efficient buildings, contributing to the ongoing discourse on environmental sustainability and energy conservation in the built environment.

References

1) *Python KNN: Mastering K Nearest Neighbor Regression with sklearn – Kanaries*. (2023, August 19). Docs.kanaries.net. https://docs.kanaries.net/topics/Python/python-knn

2) (2024). Senecapolytechnic.ca. https://learning-oreilly-com.libaccess.senecapolytechnic.ca/library/view/machine-learning-with/9780134845708/ch07.xhtml

3) *3 Methods to Tune Hyperparameters in Decision Trees - Inside Learning Machines*. (2023, January 26). https://insidelearningmachines.com/tune_hyperparameters_in_decision_trees/#Regression-2

4) *8 Feature Engineering Techniques for Machine Learning*. (n.d.). ProjectPro. https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423

5) Mujtaba, H. (2020, September 29). *An Introduction to Grid Search CV | What is Grid Search*. GreatLearning. https://www.mygreatlearning.com/blog/gridsearchcv/

6) *A guide to technical report writing*. (n.d.). https://www.theiet.org/media/5182/technical-report-writing.pdf