Introduction

This report presents the development of different predictive models for estimating heating load, an essential component of energy efficiency and management. Energy professionals can optimize heating system operations, minimize energy consumption, and assist sustainability initiatives by forecasting the daily heating energy requirements of buildings. The dataset covers a variety of building features and environmental parameters, including BuildingAge, BuildingHeight, Insulation, and environmental conditions like AverageTemperature and SunlightExposure, all of which impact total heating demand.

The intention is to construct an ideal prediction model for forecasting heating demand by utilizing a variety of regression models such as K-Nearest Neighbors (KNN), Polynomial Regression, Ordinary Least Squares (OLS), and Lasso Regression. These models are assessed using performance measures such as Root Mean Squared Error (RMSE), AIC, and BIC. Based on the model comparison, selecting the best-performing model and will later use it to estimate heating load values on the test dataset, which is critical for increasing building energy efficiency.

Variable	**Description**
HeatingLoad	Total daily heating energy required (in kWh)
BuildingAge	Age of the building (in years)
BuildingHeight	Height of the building (in meters)
Insulation	Insulation quality (1 = Good, 0 = Poor)
AverageTemperature	Average daily temperature (in °C)
SunlightExposure	Solar energy received per unit area (in W/m^2)
WindSpeed	Wind speed at the building's location (in m/s)
OccupancyRate	Proportion of the building that is occupied (percentage)

Table 1. Description of Variables

Exploratory Data Analysis

2.1 Data Splitting and Structure

The dataset was divided into two parts: a training set (70% of the data) and a validation set (30%), with a fixed random seed (random_state=1) to ensure consistency across numerous runs.

This approach ensures that the model is trained on a wide range of data points while also saving data for validation.

2.2 Data Summary on Relevant Variables

Building age: Ranges from 2.99 to 153.88 years, with a mean of around 22.73 years old.

Building height: Ranges from 3.07 to 106.36 meters, with an average of 20.79 meters.

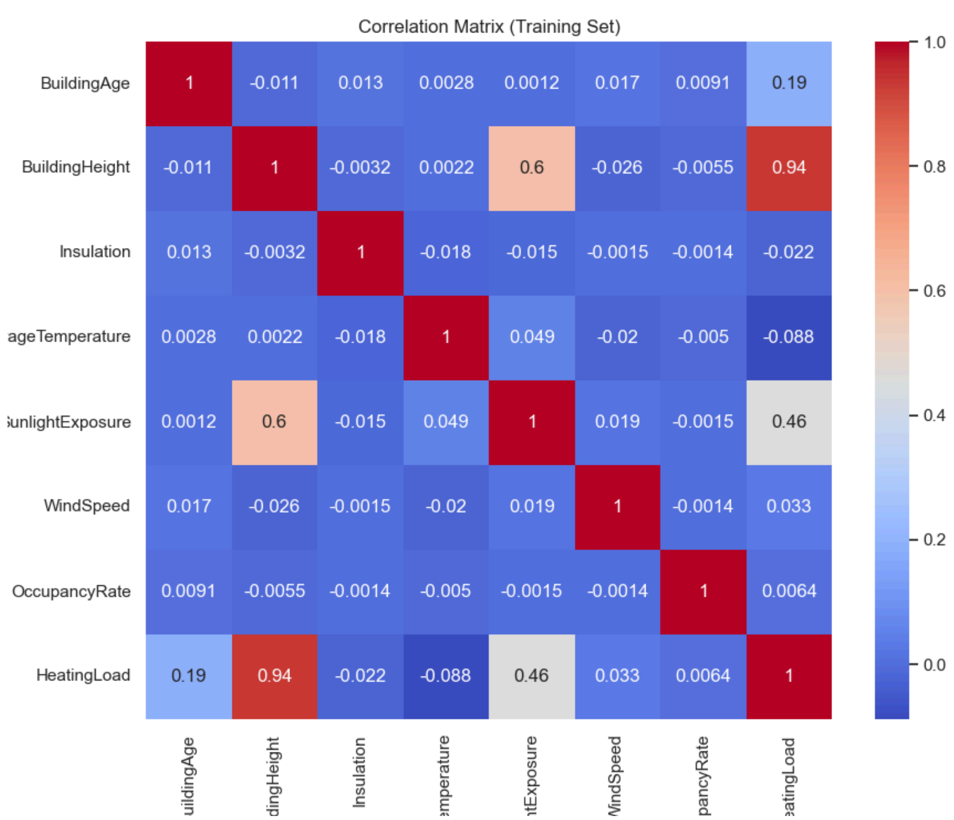
Insulation: Is a binary variable indicating good (1) or poor (0) insulation, with an average of 0.59, implying that more than half of the buildings have decent insulation.

Average temperature: Ranges from 1.68 to 34.34°C, with a mean of 18.04°C, reflecting a wide range of climatic conditions.

Heating load: Ranges from 173.68 to 793.92 kWh, with an average of around 260 kWh.

Sunlight Exposure: Ranges from 1.15 to 1250.17 with an average of 270.91.

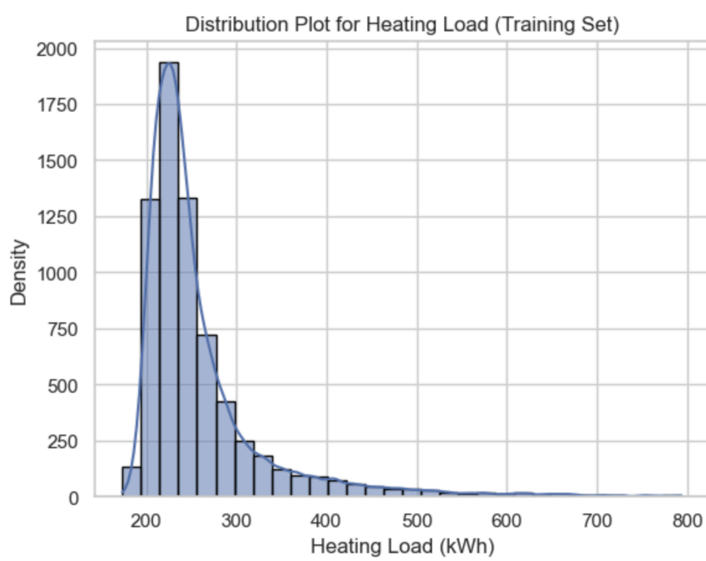
2.3 Correlation matrix



Graph 1. Correlation Matrix

The correlation matrix provides key insights on the interactions between variables in the dataset, discovering BuildingHeight as the most important predictor of HeatingLoad, with a strong positive correlation (0.94). SunlightExposure also has a moderately positive association (0.46), which supports its inclusion in the models. Insulation (-0.022) and OccupancyRate (0.0064) show low correlations, indicating that they have a little role in determining HeatingLoad. The modest correlation (0.60) between BuildingHeight and SunlightExposure indicates some possible collinearity, but not enough to cause considerable worry. This analysis directs the selection of relevant predictors while identifying factors that can be omitted due to their low influence for the later on model.

2.4 Distribution of Heating Load



Graph 2. Distribution Plot for Heating Load

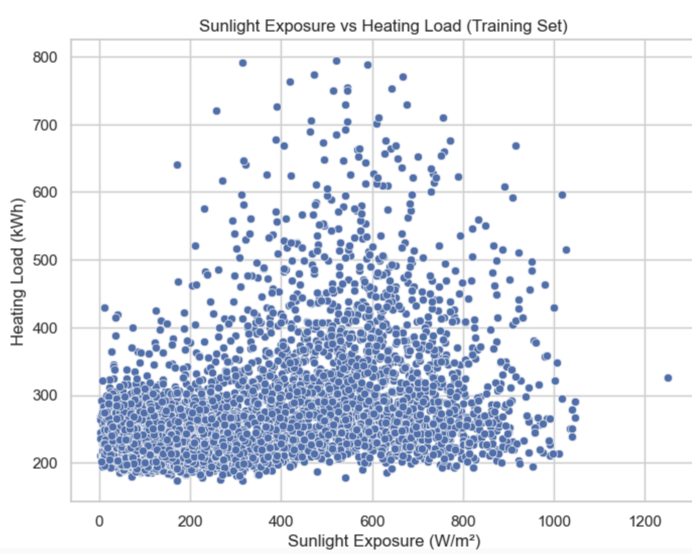
HeatingLoad's distribution plot demonstrates that the data is right-skewed, with most buildings having heating loads of 200 to 300 kWh. A few exceptions apply to heating loads greater than 500 kWh. This distribution indicates that the majority of buildings have modest energy requirements, but others have much greater demands, which might be attributed to variables such as bigger building sizes or more harsh environmental conditions.

2.5 Scatter Plots and Feature Analysis

Graph 3. Scatter Plot of Building Height vs Heating Load

The scatter plot shows an obvious positive link between the building's height and its heating load. Taller structures have greater heating loads, most likely due to their increased volume and surface area, which requires more energy to maintain a reasonable internal temperature.

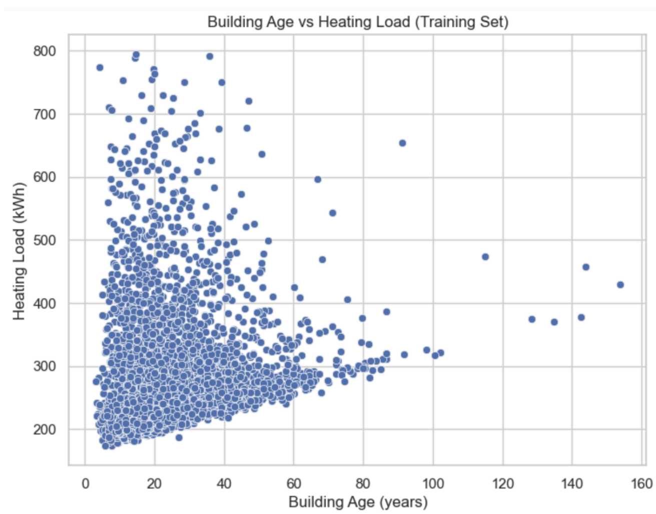
SunlightExposure vs HeatingLoad



Graph 4.Scatter Plot of Sunlight Exposure vs Heating Load

There is no obvious linear link between SunlightExposure and HeatingLoad. However, more sunshine exposure appears to lower heating load to some extent, as buildings with more sunlight require less heating. The scatter plot depicts the distribution of heating loads under varying amounts of solar exposure.

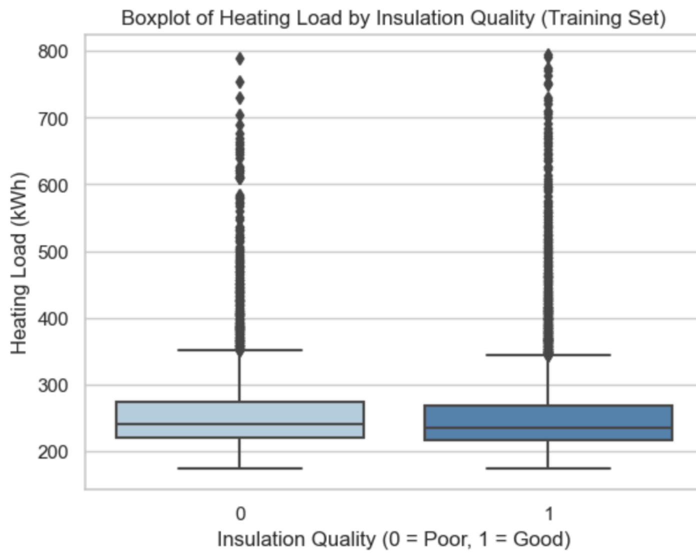
BuildingAge vs HeatingLoad



Graph 5. Scatter Plot of Building Age vs Heating Load

Older buildings typically have a greater variety of heating demands. Newer buildings typically have lower heating demands, which may imply that they were constructed with more energy-efficient features. This connection is relatively dispersed, with no clear linear trend, although newer buildings cluster around reduced heating loads.

2.6 Insulation Quality and Heating Load



Graph 6. Box Plot of HeatingLoad by Insulation Quality

A boxplot comparing heating load by insulation quality (good = 1, poor = 0) demonstrates that buildings with good insulation have lower heating loads. Buildings with good insulation have a lower median heating demand and a smaller interquartile range (IQR). Buildings with inadequate insulation exhibit a wider range of heating loads, with some outliers reaching very high levels.

Modeling and Training

3.1 Model Selection Process and Justification

To forecast HeatingLoad, I investigated different machine learning algorithms, including K-Nearest Neighbors, Polynomial Regression, Ordinary Least Squares Regression, Lasso Regression, and Ridge Regression. Each model was thoroughly examined using the validation performance measures, with an emphasis on RMSE (Root Mean Square Error), AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion) when appropriate. The following is a full discussion of the model selection procedure, which included both data analysis and trial-and-error.

The major aim was to create a model that reduced prediction error (RMSE) while being interpretable, and predictors were chosen based on correlations and feature relevance.

3.2 K-Nearest Neighbors (KNN) Regression

	Model	RMSE	AIC	BIC	Selected Predictors
0	KNN (CV: BuildingHeight, k=32)	18.80	N/A	N/A	BuildingHeight
1	KNN (CV: BuildingHeight & SunlightExposure, k=26)	18.19	N/A	N/A	BuildingHeight, SunlightExposure
2	KNN (CV: BuildingHeight, SunlightExposure, BuildingAge, k=8)	11.50	N/A	N/A	BuildingHeight, SunlightExposure, BuildingAge
3	KNN (CV: All Variables, k=7)	15.66	N/A	N/A	All Predictors

The K-Nearest Neighbors (KNN) regression method was first chosen for its simplicity and non-parametric character. Starting with BuildingHeight, the variable most connected with HeatingLoad (0.94), I examined several k values using cross-validation, and found that k=32 resulted in an RMSE of 18.80. Adding SunlightExposure, another weakly correlated variable, lowered the RMSE to 18.19 with k=26, demonstrating that integrating environmental factors might improve results. BuildingAge was included to further improve the model, resulting in a considerable RMSE drop to 11.50 with k=8, as this variable represents the building's age impacts on heating efficiency. However, when all predictors were employed, the RMSE surprisingly increased to 15.66 with k=7, indicating overfitting or noise caused by less important factors such as Insulation and WindSpeed. As a result, the final KNN model was optimized for the three most important variables: BuildingHeight, SunlightExposure, and BuildingAge.

3.3 GridSearchCV for KNN

	Model	RMSE	AIC	BIC	Selected Predictors
4	KNN (GridSearch: BuildingHeight, k=32)	18.80	N/A	N/A	BuildingHeight
5	KNN (GridSearch: BuildingHeight & SunlightExposure, k=5)	21.37	N/A	N/A	BuildingHeight, SunlightExposure
6	KNN (GridSearch: BuildingHeight, SunlightExposure, BuildingAge, k=3)	16.29	N/A	N/A	BuildingHeight, SunlightExposure, BuildingAge
7	KNN (GridSearch: BuildingAge & Insulation, k=50)	72.80	N/A	N/A	BuildingAge, Insulation
8	KNN (GridSearch: All Variables, k=3)	14.74	N/A	N/A	All Predictors

GridSearchCV was used with KNN to systematically search for the best k values across several sets of predictors. Starting with BuildingHeight, GridSearchCV verified k=32 with an RMSE of 18.80, which is consistent with the previous cross-validation findings. Adding SunlightExposure raised the RMSE to 21.37 with k=5, demonstrating that the extra variable did not enhance performance in this situation. However, by combining BuildingAge, BuildingHeight, and SunlightExposure, the RMSE reduced dramatically to 16.29 with k=3, indicating that this combination successfully reflected the variance in HeatingLoad. The results indicate that Insulation did not provide a significant contribution to the model when tested in conjunction with BuildingAge, since the RMSE increased significantly to 72.80. Incorporating numerous relevant characteristics might enhance prediction accuracy, but doing so may make the model more complicated and prone to overfitting. Ultimately, utilizing all predictors produced a reduced RMSE of 14.74 with k=3.

3.3 Polynomial Regression

	Model	RMSE	AIC	BIC	Selected Predictors
9	Polynomial (Subset 1: BuildingHeight & SunlightExposure)	17.50	17186.64	17222.67	BuildingHeight, SunlightExposure
10	Polynomial (Subset 2: BuildingAge & Insulation)	72.31	25697.91	25733.95	BuildingAge, Insulation
11	Polynomial (Subset 3: BuildingHeight, SunlightExposure, BuildingAge)	8.27	12698.08	12758.15	BuildingHeight, SunlightExposure, BuildingAge
12	Polynomial (All Variables)	1.99	4201.89	4418.12	All Predictors

To get the best model, again, I examined multiple subsets of predictors in the Polynomial Regression analysis. An RMSE of 17.50 was obtained for Subset 1, which comprised BuildingHeight and SunlightExposure. The RMSE of 72.31 for Subset 2, which included BuildingAge and Insulation, was much higher, indicating that these predictors may not be sufficient to fully account for the variation in HeatingLoad. Subset 3, which included BuildingAge in the first subset, showed the benefit of adding more pertinent variables by lowering the RMSE to 8.27. An RMSE of 1.99 was obtained by using all predictors, which resulted in the best performance. Nevertheless the significant drop in RMSE for the entire model suggests that overfitting when the model is excessively adapted to the training set might occur. Due to the likelihood that the model's complexity would capture noise rather than significant associations, this could result in poor generalization to unseen data.

3.4 Ordinary Least Squares (OLS) Regression

	Model	RMSE	AIC	BIC	Selected Predictors
13	OLS (All Variables)	16.35	59690	59740	BuildingAge, BuildingHeight, Insulation, AverageTemperature, SunlightExposure, WindSpeed, Occupa...
14	OLS (Subset 1: BuildingHeight & SunlightExposure)	23.34	64560	64580	BuildingHeight, SunlightExposure
15	OLS (Subset 2: BuildingHeight, SunlightExposure & BuildingAge)	17.80	60999	61020	BuildingHeight, SunlightExposure, BuildingAge

The advantage of OLS regression is that it minimizes the sum of squared residuals while offering a clear, understandable model that aids in identifying linear correlations between variables. In Ordinary Least Squares (OLS) Regression, I first utilized every variable that was available, which produced a comparatively low RMSE of 16.35. This model captured a substantial variety in Heating Load by utilizing a wide range of building attributes and environmental conditions. The RMSE increased to 23.34 when we limited the model to only included BuildingHeight and SunlightExposure, demonstrating that the model's predictive potential is diminished when some variables are excluded. With an RMSE of 17.80, subset 2, which comprises BuildingHeight, SunlightExposure, and BuildingAge, increased the model's accuracy and showed the inclusion of BuildingAge as a predictor enhances the regression model's usefulness.

3.5 Lasso and Ridge Regression

The dataset's potential multicollinearity was explored using Lasso and Ridge regressions. By decreasing some coefficients to zero, Lasso regression eliminated variables that were deemed unnecessary through its feature selection capabilities. In particular, since they contribute little to the model, Insulation (which has a negative coefficient of -1.78) and OccupancyRate (which has a modest coefficient of 0.65) might be eliminated. For Lasso, the optimal alpha value determined by cross-validation was 0.001, which produced an RMSE of 16.45 that was similar to the OLS model as a whole. These regularization strategies did not much outperform the OLS model, despite the fact that they did enhance model generalization and somewhat reduce multicollinearity. They did, however, offer insightful information about which variables might be safely eliminated in order to simplify the model without compromising its predictive ability.

Final Model Selection

	Model	RMSE	AIC	BIC	Selected Predictors
0	KNN (CV: BuildingHeight, k=32)	18.80	N/A	N/A	BuildingHeight
1	KNN (CV: BuildingHeight & SunlightExposure, k=26)	18.19	N/A	N/A	BuildingHeight, SunlightExposure
2	KNN (CV: BuildingHeight, SunlightExposure, BuildingAge, k=8)	11.50	N/A	N/A	BuildingHeight, SunlightExposure, BuildingAge
3	KNN (CV: All Variables, k=7)	15.66	N/A	N/A	All Predictors
4	KNN (GridSearch: BuildingHeight, k=32)	18.80	N/A	N/A	BuildingHeight
5	KNN (GridSearch: BuildingHeight & SunlightExposure, k=5)	21.37	N/A	N/A	BuildingHeight, SunlightExposure
6	KNN (GridSearch: BuildingHeight, SunlightExposure, BuildingAge, k=3)	16.29	N/A	N/A	BuildingHeight, SunlightExposure, BuildingAge
7	KNN (GridSearch: BuildingAge & Insulation, k=50)	72.80	N/A	N/A	BuildingAge, Insulation
8	KNN (GridSearch: All Variables, k=3)	14.74	N/A	N/A	All Predictors
9	Polynomial (Subset 1: BuildingHeight & SunlightExposure)	17.50	17186.64	17222.67	BuildingHeight, SunlightExposure
10	Polynomial (Subset 2: BuildingAge & Insulation)	72.31	25697.91	25733.95	BuildingAge, Insulation
11	Polynomial (Subset 3: BuildingHeight, SunlightExposure, BuildingAge)	8.27	12698.08	12758.15	BuildingHeight, SunlightExposure, BuildingAge
12	Polynomial (All Variables)	1.99	4201.89	4418.12	All Predictors
13	OLS (All Variables)	16.35	59690	59740	BuildingAge, BuildingHeight, Insulation, AverageTemperature, SunlightExposure, WindSpeed, Occupa...
14	OLS (Subset 1: BuildingHeight & SunlightExposure)	23.34	64560	64580	BuildingHeight, SunlightExposure
15	OLS (Subset 2: BuildingHeight, SunlightExposure & BuildingAge)	17.80	60999	61020	BuildingHeight, SunlightExposure, BuildingAge
16	Lasso Regression (Best Alpha 0.001)	16.45	59780	59820	BuildingAge, BuildingHeight, AverageTemperature, SunlightExposure, WindSpeed

Table 2. Full Comparing Table with all Models

The optimal model I've chosen is the polynomial model with 'BuildingHeight', 'SunlightExposure', 'BuildingAge' for predictors since it strikes a compromise between simplicity and forecast accuracy. It performs more effectively than other models, such as the KNN variations, with an RMSE of 8.27, and is only somewhat less accurate than the Polynomial model employing all variables (RMSE of 1.99). However, with all predictors included in the Polynomial (All Variables) model, it raises complexity and boosts the possibility of overfitting to validation data, which weakens the model's ability to generalize effectively to unseen data. The Polynomial (Subset 3) model, on the other hand, simplifies the model while maintaining a low error rate as it only includes three important predictors: BuildingHeight, SunlightExposure, and BuildingAge. Though it has somewhat higher AIC/BIC values than certain OLS models, it nevertheless strikes a fair balance between model complexity and forecast accuracy. It is more accurate in predicting on validation dataset as its RMSE of 8.27 is significantly lower than all of the OLS models. It is a reliable and robust model for forecasting unseen data due to its simplicity and strong predictive performance.

Conclusion and Insights

The final prediction of HeatingLoad was generated using the 'HeatingLoad_test_without_HL.csv' dataset with the optimal model, which utilized BuildingHeight, SunlightExposure, and BuildingAge as predictors. A Polynomial Regression (degree 2) model was chosen to capture nonlinear relationships between the variables. The resulting predictions were saved in the 520397297_Assignment1_HL_prediction.csv file. The model's accuracy was assessed using Mean Squared Error (MSE), comparing the predicted HeatingLoad values to actual test set values. This process confirmed that the chosen model is suitable for making future predictions with similar data, demonstrating its effectiveness.

To sum up, the chosen model gives beneficial information for improving heating systems in buildings, hence lowering energy consumption and running expenses. Despite its strong performance, future enhancements include adding more building attributes and better modeling approaches to improve forecast accuracy. The model's usage in actual applications has the potential to considerably improve energy efficiency; however, future work may address issues such as the need for more detailed data.