Title: Exploring Nonlinear Models and Model Selection for Real-World Problems
MSA 8150 - Machine Learning for Analytics
Homework Assignment 2
Student Name: Harshal Kamble
Date: March, 13 2025

---

1. Introduction

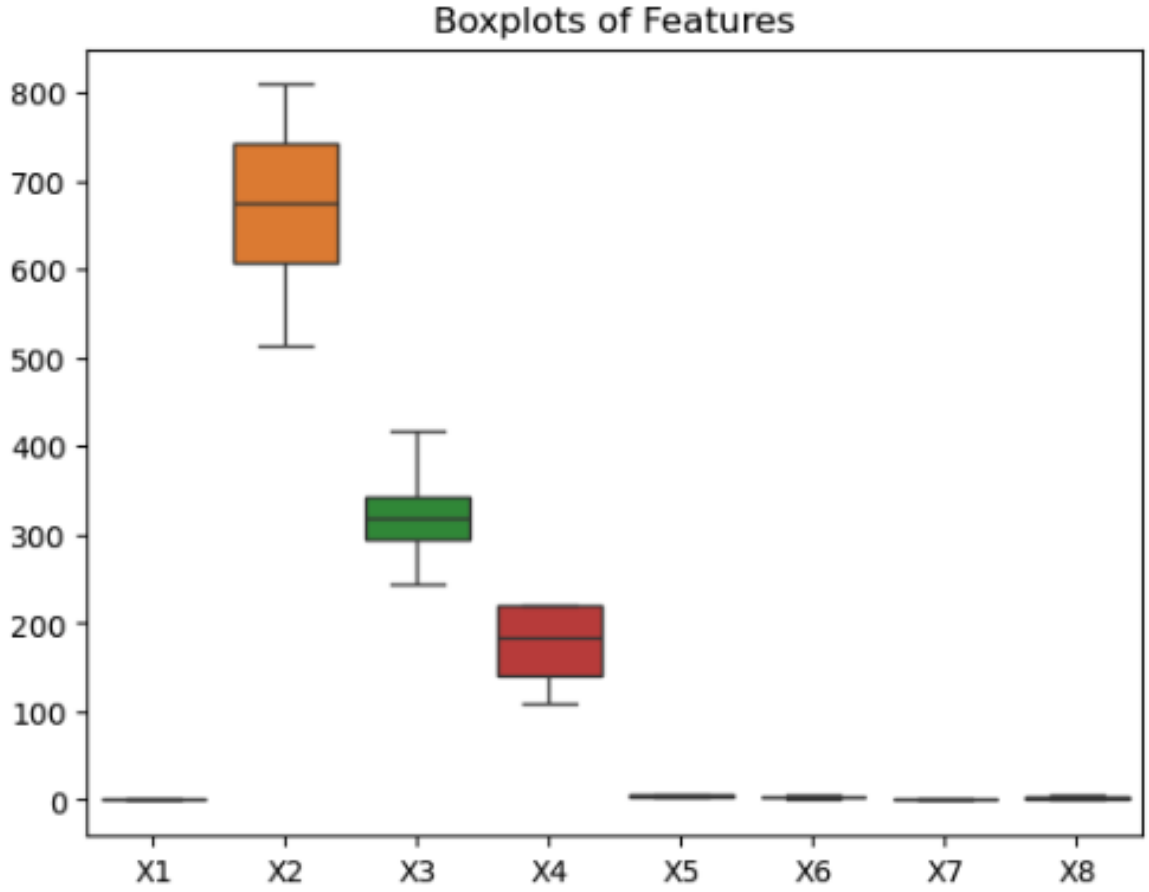
The goal of this assignment was to explore nonlinear machine learning models and implement techniques for model selection to predict the heating load (Y1) of buildings. The prediction was based on features such as relative compactness and surface area, and glazing area. The emphasis was on evaluating the performance of various models and selecting the most appropriate one by using both cross-validation and testing metrics.

---

2. Dataset Description

- **Source:** UCI Energy Efficiency Dataset
- **Instances:** 768
- **Features (Inputs):**
  - X1: Relative Compactness
  - X2: Surface Area
  - X3: Wall Area
  - X4: Roof Area
  - X5: Overall Height
  - X6: Orientation (Categorical)
  - X7: Glazing Area
  - X8: Glazing Area Distribution (Categorical)
- **Targets:**
  - Y1: Heating Load (target for this task)
  - Y2: Cooling Load (not used in this task)

Summary Statistics and EDA:



No missing values were present, and the features were a mix of numerical and categorical variables. Outliers were visually checked using boxplots.

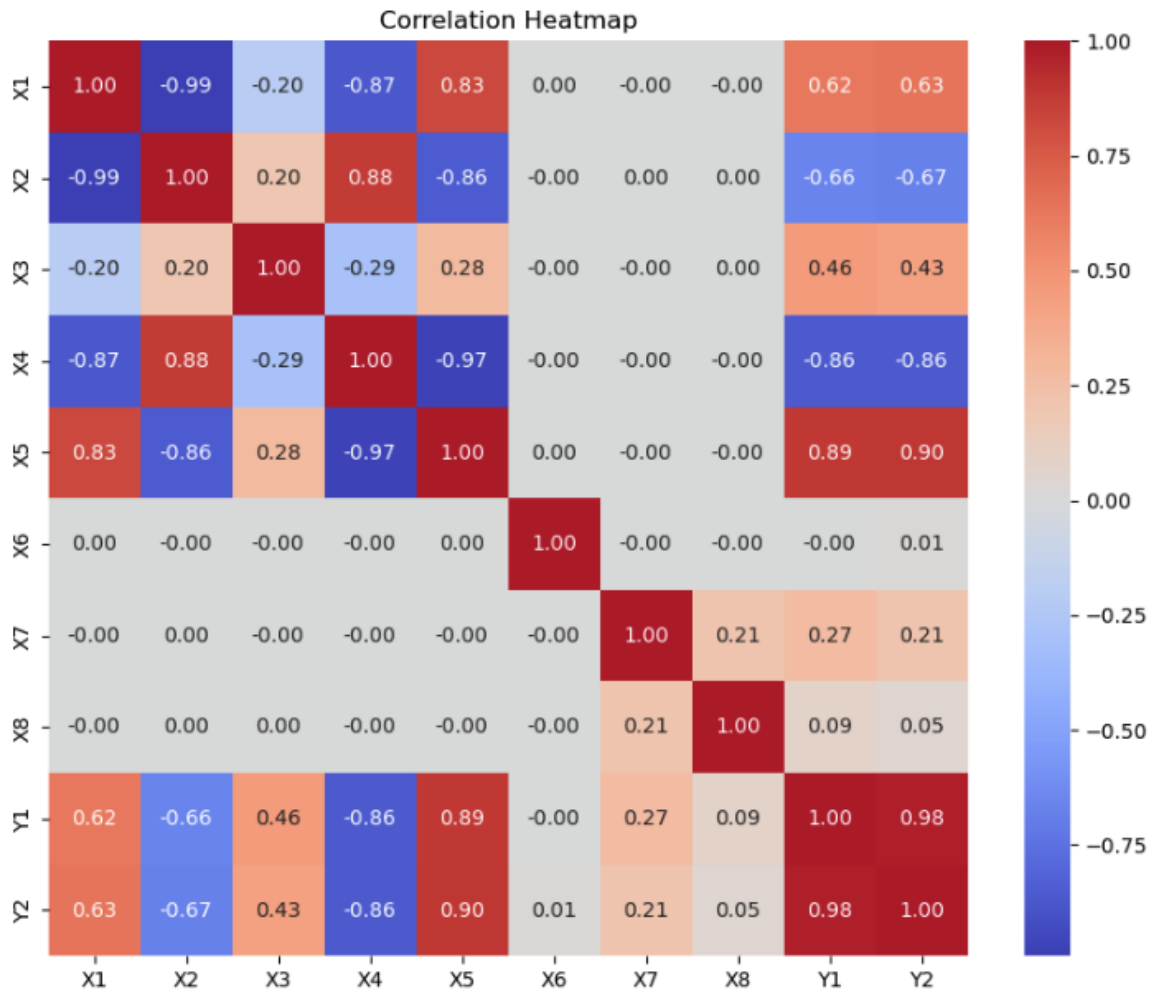
Correlation Analysis

The following heatmap visualizes the correlation between all numerical features and target variables.

- Relative Compactness (X1) shows a strong negative correlation with Heating Load (Y1).

- Surface Area (X2) and Wall Area (X3) have moderate correlations.

These insights informed the selection of features and models for further analysis.



---

3. Data Preprocessing

- Categorical variables (X6 and X8) were one-hot encoded.
- Numerical features were standardized. The dataset was split into training (80%) and testing (20%) sets.
- Preprocessing pipelines were implemented to ensure consistency across all models.

---

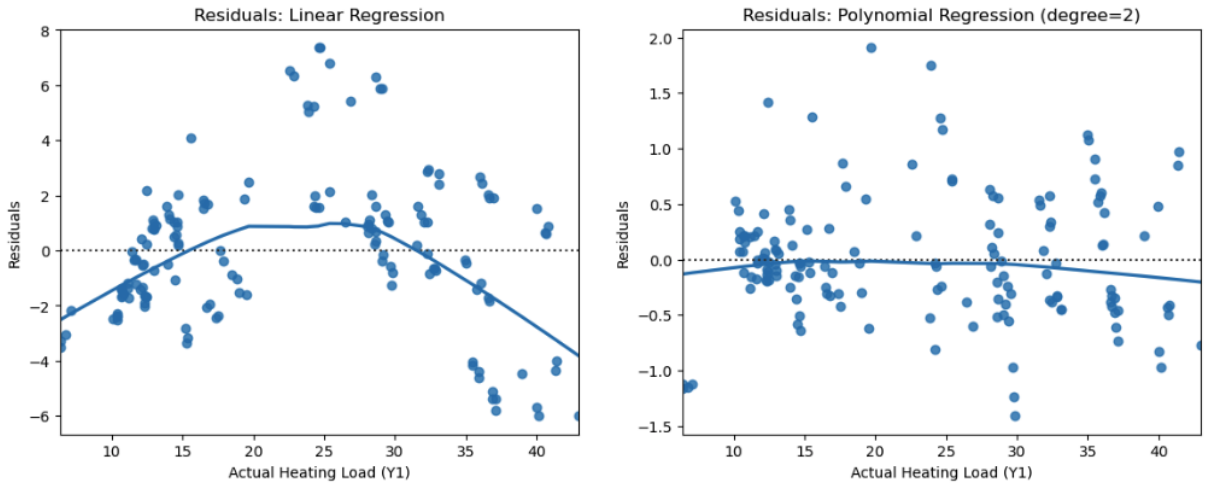
4. Models Implemented

4.1 Linear Regression (Baseline Model)

- A linear model was implemented as a baseline.
- **Performance on Test Set:**
  - MSE: **8.25**
  - RMSE: **2.87**
  - R^2: **0.92**

4.2 Polynomial Regression (Nonlinear Model)

- Degree = 2 was selected after cross-validation to avoid overfitting.
- Cross-validation MSE: **0.30**
- **Performance on Test Set:**
  - MSE: **0.33**
  - RMSE: **0.57**
  - R^2: **1.00**



4.3 Ridge Regression (Regularization Model)

- Hyperparameter (alpha) was tuned with cross-validation.
- Best Alpha: **0.1**
- **Performance on Test Set:**
  - MSE: **8.27**
  - RMSE: **2.88**
  - R^2: **0.92**

4.4 Decision Tree Regressor (Bonus Nonlinear Model)

- Hyperparameters tuned with cross-validation:
  - max_depth = 7
  - min_samples_leaf = 5
- **Performance on Test Set:**
  - MSE: **0.36**
  - RMSE: **0.60**
  - R^2: **1.00**

---

5. Model Comparison & Analysis

Model	MSE	RMSE	R^2
Linear Regression	8.25	2.87	0.92
Polynomial Regression (2)	0.33	0.57	1.00
Ridge Regression	8.27	2.88	0.92
Decision Tree	0.36	0.60	1.00

Discussion:

- Polynomial Regression (degree=2) outperforms all other models, demonstrating the best performance in both cross-validation and testing.
- The Decision Tree Regressor also performs well, but it requires careful pruning to prevent overfitting. Residual plots indicate that the Polynomial Regression effectively eliminates the bias found in the linear baseline model.
- On the other hand, Ridge Regression did not provide any improvement over the baseline, largely due to the absence of significant multicollinearity.

---

6. Challenges and Solutions

- ***Challenge***: The risks of overfitting with high-degree Polynomial Regression and unpruned Decision Trees.
- ***Solution***: Implemented cross-validation for hyperparameter tuning and conducted residual diagnostics to ensure model generalization.

---

7. Conclusion

- ***Best Model:*** Polynomial Regression (degree = 2) was chosen due to its consistent cross-validation scores and strong generalization performance.
- Cross-validation effectively helped prevent overfitting while maintaining robust model performance.
- Residual plots and the comparison of predicted versus actual values confirmed a strong model fit, with no issues related to bias or variance.