Regression Analysis on Energy Efficiency Dataset

1. Objective

The objective of this project is to apply regression models to predict the energy efficiency of residential buildings. Specifically, we aim to model two target variables:

- Y1: Heating Load
- Y2: Cooling Load

We evaluate the performance of two types of regression models: Linear Regression and Ridge Regression.
The dataset is split into training, validation, and testing subsets (70/10/20) to ensure proper evaluation.

2. Data

- Source: ENB2012 dataset (Energy Efficiency Dataset).
- Features (X1–X8):
    - Relative Compactness
    - Surface Area
    - Wall Area
    - Roof Area
    - Overall Height
    - Orientation
    - Glazing Area
    - Glazing Area Distribution
- Targets:

    - Y1 (Heating Load)
    - Y2 (Cooling Load)
- Split:
    - 70% Training
    - 10% Validation
    - 20% Test

3. Method

1. Preprocessing:
    - Loaded the dataset from Excel.
    - Features X1−X8 used to predict Y1 and Y2.
2. Data Splitting:
    - Applied a 70/10/20 split using `train_test_split`.
    - Training: 70% of data
    - Validation: 10% of data
    - Testing: 20% of data
3. Models Trained:
    - Linear Regression (no regularization)
    - Ridge Regression (regularized, with alpha values like 0.0, 0.1, 1, 10, 100)
4. Evaluation Metrics:
    - RMSE (Root Mean Squared Error) – penalizes larger errors.
    - MAE (Mean Absolute Error) – measures average prediction error.
    - R^2 (Coefficient of Determination) – shows how well the model explains variance in the data.

5. Outputs:
    - Results stored in `metrics.csv` and `metrics.json`.
    - Console summary printed for quick comparison.

4. Results

Target	Model	Split	RMSE	MAE	R^2
Y1	Linear	Validation	3.361	2.440	0.893
Y1	Linear	Test	2.765	2.014	0.922
Y1	Ridge	Validation	3.897	2.626	0.857
Y1	Ridge	Test	2.765	2.014	0.922
Y2	Linear	Validation	3.618	2.529	0.868
Y2	Linear	Test	2.984	2.134	0.895
Y2	Ridge	Validation	4.268	2.739	0.816
Y2	Ridge	Test	2.984	2.134	0.895

Linear Regression performed slightly better overall than Ridge Regression for both targets.

Validation scores were consistent with test scores, suggesting the models generalize well.

Best R^2: ~0.92 on Heating Load (Y1, Linear Regression, Test).

Ridge Regression did not improve over Linear Regression in this dataset, possibly because the dataset is not highly noisy and multicollinearity is limited.

5. Conclusion

- The dataset was successfully split into training, validation, and test sets (70/10/20).
- Both Linear and Ridge Regression models were trained and evaluated.
- Metrics (RMSE, MAE, R^2) were computed to assess performance.
- Linear Regression showed slightly better performance than Ridge Regression.