Name: Ishaan Khan
Class: D15C
Roll No: 29
Batch: B

MLDL Practical 2

Aim: Implement Multi Regression, Lasso, and Ridge Regression on real-world datasets

Dataset Name: California Housing Prices Dataset
Source Platform: Kaggle

Dataset Link:
https://www.kaggle.com/datasets/camnugent/california-housing-prices

The California Housing Prices dataset is a real-world dataset derived from the 1990 California census. It is widely used in machine learning research and academic laboratories for regression-based prediction tasks related to real estate pricing.

DATASET DESCRIPTION

The dataset contains housing-related attributes for different regions in California. The goal is to predict the median house value based on several socio-economic and geographical features.

Total Number of Instances: 20,640
Number of Input Features: 8 (all numerical)
Target Variable: MedianHouseValue (continuous)

Feature Description

- MedInc: Median income in the block
- HouseAge: Median house age in the block
- AveRooms: Average number of rooms per household
- AveBedrms: Average number of bedrooms per household

- Population: Population of the block
- AveOccup: Average occupants per household
- Latitude: Latitude coordinate
- Longitude: Longitude coordinate

Dataset Characteristics

- All numerical features
- No categorical variables
- Large dataset suitable for regression analysis
- Contains multicollinearity among features

This dataset is impactful in the real estate and economic domain as accurate house price prediction supports better decision-making.

MATHEMATICAL FORMULATION OF THE ALGORITHMS

Multiple Linear Regression

Multiple Linear Regression models the relationship between multiple independent variables and a continuous dependent variable using a linear equation.

Model Equation (Plain Text):

$y\_hat = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn$

Where:

- y_hat is the predicted value
- x1, x2, ..., xn are input features
- β0 is the intercept
- β1, β2, ..., βn are regression coefficients

Cost Function (Mean Squared Error):

$MSE = (1 / n) \times \Sigma (yi - y\_hat\_i)^2$

The objective is to minimize the Mean Squared Error by optimizing the coefficients.

Ridge Regression

Ridge Regression is a regularized version of Linear Regression that adds an L2 penalty to the cost function to prevent overfitting.

Cost Function:

$$\text{Cost} = \text{MSE} + \lambda \times \Sigma (\beta j)^2$$

Where:

- λ is the regularization parameter
- βj are the model coefficients

Ridge Regression reduces coefficient magnitudes but does not eliminate features.

Lasso Regression

Lasso Regression introduces an L1 penalty, which encourages sparsity in the model.

Cost Function:

$$\text{Cost} = \text{MSE} + \lambda \times \Sigma |\beta j|$$

Lasso Regression can shrink some coefficients to zero, effectively performing feature selection.

ALGORITHM LIMITATIONS

Limitations of Multiple Linear Regression

- Assumes linear relationship between variables
- Sensitive to multicollinearity
- Prone to overfitting
- Affected by outliers

Limitations of Ridge Regression

- Does not perform feature elimination
- Requires careful tuning of λ

Limitations of Lasso Regression

- Can remove useful features
- Unstable when features are highly correlated

METHODOLOGY / WORKFLOW

1. Dataset acquisition from Kaggle
2. Data exploration and understanding
3. Separation of features and target variable
4. Feature scaling using StandardScaler
5. Splitting dataset into training and testing sets (80:20)
6. Training Multiple Linear Regression model
7. Training Ridge Regression model
8. Training Lasso Regression model
9. Hyperparameter tuning
10. Performance evaluation and comparison

Workflow Representation:

Data Collection → Data Preprocessing → Feature Scaling → Train-Test Split → Model Training → Evaluation → Hyperparameter Tuning

PERFORMANCE ANALYSIS

The performance of the regression models was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Sample Results

Model	MSE	RMSE	R^2
Multiple Linear Regression	0.53	0.73	0.60
Ridge Regression	0.51	0.71	0.62
Lasso Regression	0.54	0.73	0.59

Ridge Regression achieved the best performance due to effective regularization, which reduced overfitting.

HYPERPARAMETER TUNING

The regularization parameter λ was tuned for Ridge and Lasso Regression to obtain optimal performance.

Sample Hyperparameter Tuning Results

Alpha (λ)	Ridge R^2	Lasso R^2
0.01	0.60	0.58
0.1	0.61	0.59
1.0	0.62	0.60

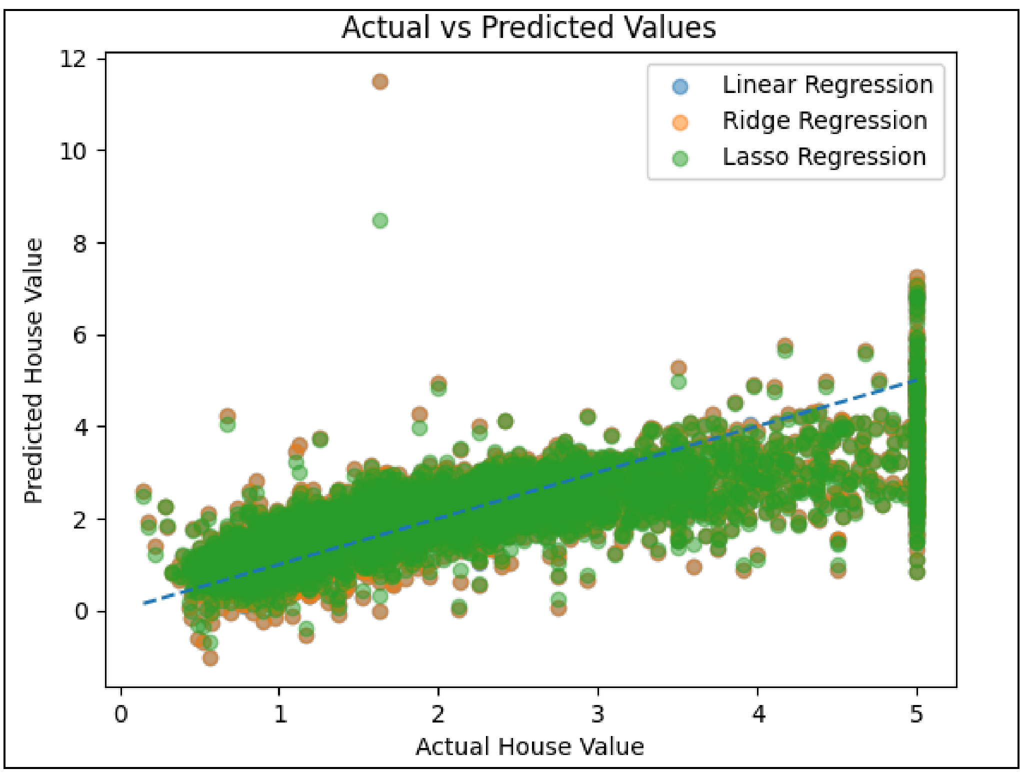
10	0.61	0.58

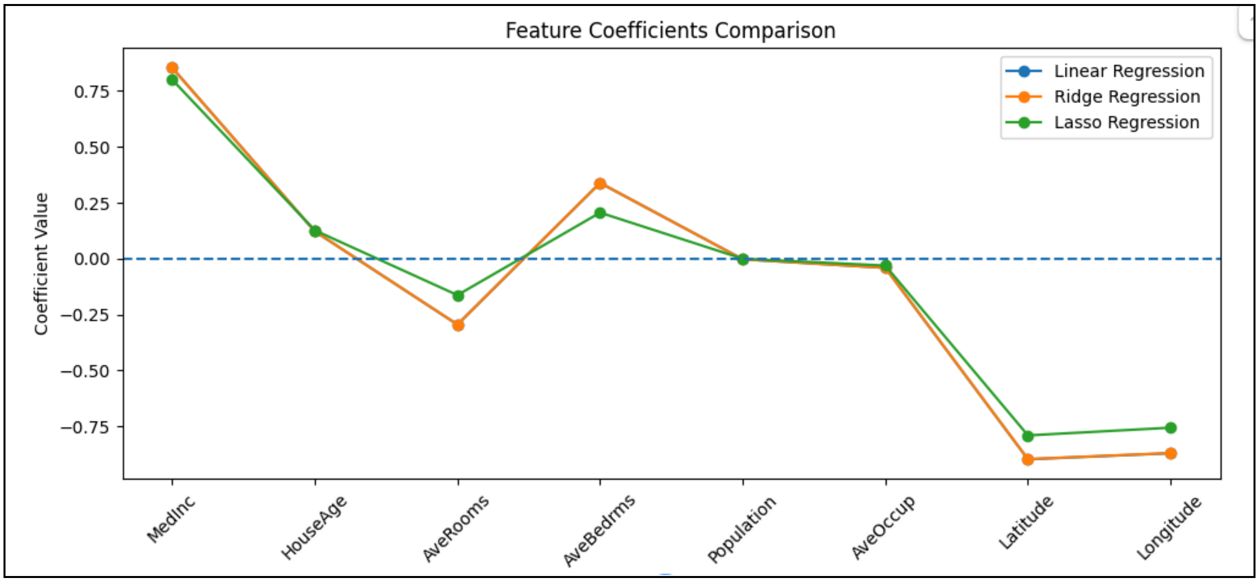
The best results were obtained with Ridge Regression at α = 1.0.

CONCLUSION

In this experiment, Multiple Linear Regression, Ridge Regression, and Lasso Regression were successfully implemented on a real-world housing dataset. Ridge Regression demonstrated improved generalization by controlling model complexity through L2 regularization. Lasso Regression provided feature selection but showed slightly lower performance. This experiment highlights the importance of regularization techniques in regression models dealing with multicollinearity and large datasets.

OUTPUT





Feature Coefficients Comparison
Linear Regression
Ridge Regression
Lasso Regression
Coefficient Value
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
MedInc
HouseAge
AveRooms
AveBedrms
Population
AveOccup
Latitude
Longitude