EXP – 2

MLDL Experiment – Multiple, Ridge and Lasso Regression

Name: Janhavi Mandhan
Class:D15C
Roll No:33
Batch B

Aim:

To implement Multiple Linear Regression, Ridge Regression and Lasso Regression on a real-world dataset and compare their performance.

---

1. Dataset Source

Medical Insurance Dataset

Source: Kaggle

https://www.kaggle.com/datasets/mirichoi0218/insurance

---

2. Dataset Description

The **Medical Insurance Dataset** is a real-world dataset used to analyze and predict medical insurance costs based on individual demographic and health-related factors. The dataset contains **1,338 records**, where each record represents a person along with their insurance

details. The dataset is provided in **CSV format** and does not contain missing values, making it suitable for regression analysis.

Attributes of the Dataset:

- **age:** Age of the individual (numerical)
- **sex:** Gender of the individual (male/female), later encoded numerically
- **bmi:** Body Mass Index, which indicates body fat based on height and weight
- **children:** Number of children or dependents covered by insurance
- **smoker:** Smoking status (yes/no), encoded as binary values
- **region:** Residential area in the US (northeast, northwest, southeast, southwest), encoded numerically
- **charges:** Medical insurance cost billed to the individual (continuous numerical value)

Target Variable:

- **charges** – Used as the dependent variable to predict medical insurance costs.

Why this Dataset is Suitable:

- Contains **multiple independent variables**, making it ideal for **Multiple Linear Regression**.
- Presence of correlated features allows effective demonstration of **Ridge and Lasso Regression**.
- Real-world nature helps in understanding **regularization techniques** to reduce overfitting.

Overall, this dataset is well-suited for comparing Multiple Linear Regression, Ridge Regression, and Lasso Regression models.

---

3. Mathematical Formulation

- ## Multiple Linear Regression:

Multiple Linear Regression is a supervised learning algorithm used to predict a **continuous dependent variable** using **multiple independent variables**. It assumes a linear relationship between the input features and the output variable.

Model Equation

y=β0+β1x1+β2x2+β3x3+⋯+ β nxny = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_ny=β0+β1x1+β2x2+β3x3+⋯+ β nxn

Where:

- yyy = predicted medical insurance charges
- x1,x2,...,xnx_1, x_2, \dots, x_nx1,x2,...,xn = independent variables (age, BMI, children, smoker, region, etc.)
- β0\beta_0β0 = intercept
- β1,β2,...,βn\beta_1, \beta_2, \dots, \beta_nβ1,β2,...,βn = regression coefficients

Objective Function

The coefficients are estimated by minimizing the **Residual Sum of Squares (RSS)**:

RSS=∑i=1n(yi−y^i)2RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2RSS=i=1∑n(yi−y^i)2

This is commonly solved using the **Ordinary Least Squares (OLS)** method.

- ## Ridge Regression:

Ridge Regression is an extension of Multiple Linear Regression that introduces **L2 regularization** to reduce model complexity and prevent overfitting, especially when multicollinearity exists among features.

Cost Function

J(β)=∑i=1n(yi−y^i)2+α∑j=1nβj2J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{n} \beta_j^2J(β)=i=1∑n(yi−y^i)2+αj=1∑nβj2

Where:

- α\alphaα = regularization parameter
- ∑βj2\sum \beta_j^2∑βj2 = L2 penalty term

Effect of Ridge Regression

- Shrinks coefficient values toward zero
- Does **not** make coefficients exactly zero
- Helps stabilize the model when predictors are highly correlated

- ## Lasso Regression:

Lasso Regression adds **L1 regularization** to the loss function, which helps in both **overfitting control** and **feature selection**.

Cost Function

J(β)=∑i=1n(yi−y^i)2+α∑j=1n∣βj∣J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{n} |\beta_j|J(β)=i=1∑n(yi−y^i)2+αj=1∑n∣βj∣

Where:

- α\alphaα = regularization strength
- ∑∣βj∣\sum |\beta_j|∑∣βj∣ = L1 penalty term

Effect of Lasso Regression

- Shrinks some coefficients **exactly to zero**
- Performs automatic **feature selection**
- Produces simpler and more interpretable models

Role of Regularization Parameter (α)

- **Small *α*:** Model behaves like Multiple Linear Regression
- **Large *α*:** Strong regularization, simpler model
- α controls the trade-off between **bias and variance**

---

4. Algorithm Limitations

Multiple Linear Regression:

- Sensitive to multicollinearity
- Assumes linear relationship
- Affected by outliers
- Overfits when many correlated features exist

Ridge Regression:

- Does not perform feature selection
- Requires tuning of α

Lasso Regression:

- Can remove important features if α is too large
- Less stable when features are highly correlated

---

5. Methodology / Workflow

Dataset → Preprocessing → Train/Test Split → Model Training → Prediction → Evaluation

---

6. Performance Analysis

- Multiple Linear Regression, Ridge Regression, and Lasso Regression were evaluated using Mean Squared Error and R^2 Score.
- Ridge Regression showed improved stability compared to Multiple Linear Regression.
- Lasso Regression helped reduce less important features.

---

7. Hyperparameter Tuning

The **α (alpha)** parameter was tuned for Ridge and Lasso Regression to control regularization strength. Proper tuning improves model generalization and prevents overfitting.

Conclusion:

Ridge and Lasso Regression improve model generalization over Multiple Linear Regression by reducing overfitting.