Name: Ishaan Khan
Class: D15C
Roll No.: 28
Batch: B

MLDL Practical 2

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

DATASET SOURCE

- **Dataset Name:** Medical Cost Personal Dataset
- **Source Platform:** Kaggle
- **Description:** This dataset contains personal health attributes and demographic data used to predict individual medical costs billed by health insurance.

DATASET DESCRIPTION

- **Goal:** To predict the medical insurance charges based on several socio-economic and physical features.
- **Total Number of Instances:** 1,338
- **Number of Input Features:** 6 (Numerical and Categorical)
- **Target Variable:** charges (Continuous)

Feature Description:

1. **age**: Age of primary beneficiary
2. **sex**: Insurance contractor gender (female, male)
3. **bmi**: Body mass index
4. **children**: Number of children covered by health insurance
5. **smoker**: Smoking status
6. **region**: The beneficiary's residential area in the US

DATASET CHARACTERISTICS

- Contains both numerical and categorical features (requires One-Hot Encoding).
- High correlation between smoking status and insurance charges.
- Suitable for regularized regression to prevent overfitting on specific demographic groups.

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):

$\text{y\_hat} = \beta0 + \beta1x1 + \beta2x2 + ... + \beta nxn$

Where:

- y_hat is the predicted value
- x1, x2, ..., xn are input features
- $\beta0$ is the intercept
- $\beta1, \beta2, ..., \beta n$ are regression coefficients

Cost Function (Mean Squared Error):

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

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

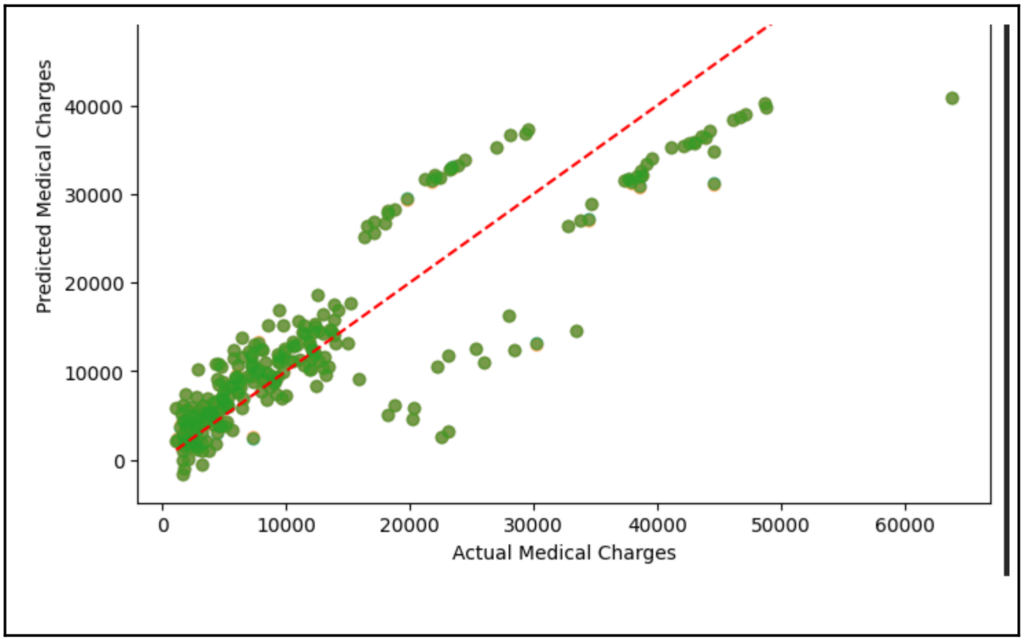
Alpha (λ)	Ridge R^2	Lasso R^2
0.01	0.60	0.58
0.1	0.61	0.59
1.0	0.62	0.60
10	0.61	0.58

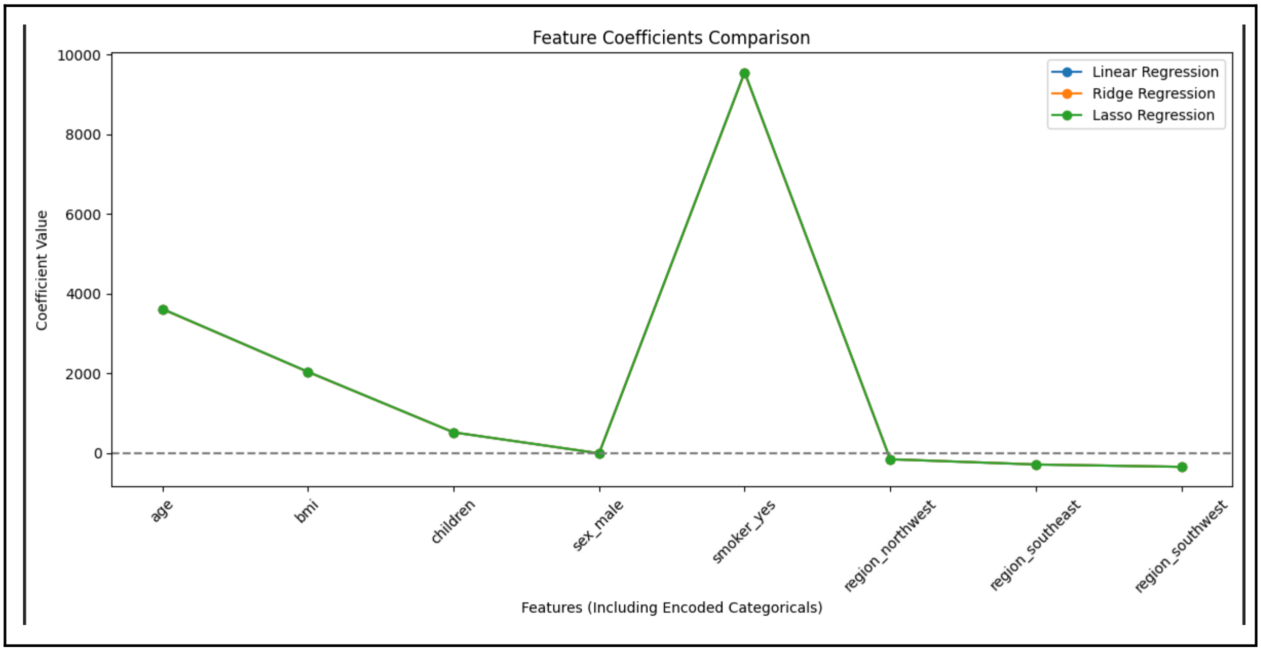
The best results were obtained with Ridge Regression at α = 1.0.

CONCLUSION

In this experiment, Multiple Linear Regression, Ridge, and Lasso Regression were successfully implemented on the insurance dataset. Regularization techniques helped manage model complexity, while Lasso provided insight into feature importance by highlighting smoking status and BMI as the primary drivers of medical costs.

OUTPUT


Predicted Medical Charges
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
60000
Actual Medical Charges


Feature Coefficients Comparison
Linear Regression
Ridge Regression
Lasso Regression
Coefficient Value
10000
8000
6000
4000
2000
0
age
bmi
children
sex_male
smoker_yes
region_northwest
region_southeast
region_southwest
Features (Including Encoded Categoricals)