

Experiment: Comparative Analysis of Linear, Lasso, and Ridge Regression on Medical Insurance Costs

1. Dataset Source

For this experiment, we utilized the **Medical Cost Personal Datasets** (`insurance.csv`).

- **Source:** Machine Learning Repository / Kaggle (Medical Cost Personal Datasets).
- **File Name:** `insurance.csv` (Provided in the environment).

2. Dataset Description

The dataset contains information about medical insurance beneficiaries and the corresponding insurance charges. The goal is to predict the individual medical costs billed by health insurance.

- **Target Variable (y):** `charges` (Individual medical costs billed by health insurance).
- **Dataset Size:** 1,338 samples (instances) and 7 columns.
- **Features (X):**
 1. `age`: Age of the primary beneficiary.
 2. `sex`: Insurance contractor gender (female, male).
 3. `bmi`: Body mass index (kg/m²), providing an understanding of body weights that are relatively high or low relative to height.
 4. `children`: Number of children covered by health insurance / Number of dependents.
 5. `smoker`: Smoking status (yes, no).
 6. `region`: The beneficiary's residential area in the US (northeast, southeast, southwest, northwest).

3. Mathematical Formulation of the Algorithm

The experiment implements three regression algorithms, each optimizing a specific cost function.

A. Multiple Linear Regression (OLS)

The Ordinary Least Squares (OLS) method minimizes the sum of squared residuals between the observed targets and the predicted values.

Hypothesis:

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$$

Cost Function ($J(\theta)$):

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

B. Ridge Regression (L2 Regularization)

Ridge regression addresses overfitting by adding a penalty proportional to the square of the magnitude of coefficients. This shrinks the coefficients but keeps them non-zero.

Cost Function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2$$

C. Lasso Regression (L1 Regularization)

Lasso (Least Absolute Shrinkage and Selection Operator) adds a penalty proportional to the absolute value of the coefficients. This can shrink some coefficients to exactly zero, effectively performing feature selection.

Cost Function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n |\theta_j|$$

4. Algorithm Limitations

- **Linearity Assumption:** All three models assume a linear relationship between features (like Age, BMI) and the target (Charges). If the relationship is polynomial or highly non-linear, these models will underperform.
- **Sensitivity to Outliers:** Linear Regression is highly sensitive to outliers (e.g., extremely high medical charges for a few individuals). Ridge and Lasso are more robust but still affected.
- **Feature Scaling:** Lasso and Ridge require feature scaling (standardization) because the penalty terms depend on the magnitude of the coefficients. Without scaling, features with larger ranges (like "charges") would dominate the penalty.

5. Methodology / Workflow

The experiment follows a standard Machine Learning pipeline:

1. **Data Loading:** Ingest `insurance.csv` into a Pandas DataFrame.
 2. **Preprocessing:**
 - **Categorical Encoding:** Convert `sex`, `smoker`, and `region` into numerical format using One-Hot Encoding.
 - **Feature Scaling:** Apply `StandardScaler` to numerical features (`age`, `bmi`, `children`) to ensure they have a mean of 0 and variance of 1.
 3. **Data Splitting:** Split the dataset into **Training (80%)** and **Testing (20%)** sets.
 4. **Model Training:** Initialize Linear, Ridge, and Lasso regressors.
 5. **Hyperparameter Tuning:** Use `GridSearchCV` to find the optimal α (regularization strength) for Ridge and Lasso.
 6. **Evaluation:** Predict charges on the Test set and calculate RMSE and R2.
-

6. Performance Analysis

The models were evaluated on the test set.

Model	RMSE (Lower is better)	R2 Score (Higher is better)	Best Alpha (λ)
Linear Regression	5796.28	0.7836	N/A
Ridge Regression	5800.46	0.7833	1.0
Lasso Regression	5853.72	0.7793	100.0

Export to Sheets

Interpretation:

- All three models performed similarly, explaining approximately **78%** of the variance in medical charges.

- **Feature Importance (Lasso Analysis):** The Lasso model (with $\alpha=100$) zeroed out the coefficients for `sex`, and `region`. This indicates that **Smoker status, Age, and BMI** are the most significant drivers of medical costs, while gender and region have negligible impact in this linear context.
 - The high coefficient for `smoker_yes` (~23,000) confirms that being a smoker adds significantly to the predicted insurance cost.
-

7. Hyperparameter Tuning

We performed hyperparameter tuning for Ridge and Lasso using **Grid Search** with 5-fold Cross-Validation.

- **Ridge Regression:** We tested alphas [0.001,0.01,0.1,1,10,100]. The best performance was found at $\alpha=1$.
- **Lasso Regression:** We tested alphas [0.001,0.01,0.1,1,10,100]. The best performance was found at $\alpha=100$.
- **Impact:** Tuning allowed Lasso to aggressively penalize less important features (Region, Sex), simplifying the model without significantly sacrificing accuracy (RMSE increased only slightly compared to OLS).