SWE4012 Machine Learning Lab

|  |  |  |  |
| --- | --- | --- | --- |
| Register Number | 22MIS1150 | Subject Code | SWE4012 |
| Name of the Student | Divyanshu Singh | Subject Name | Machine Learning |
| Programme | Int. MTech | Slot | L3-L4 |
| Course | Software Engg. | Semester | Winter Sem 24-25 |
| Date | 10/02/25 | Exp No | 07 |

**Aim**: Compare the performance of **Linear Regression, Ridge Regression, and Lasso Regression** on different train-test splits and identify the best ratio based on error metrics on a Random Dataset

**Dataset Description:**

The dataset consists of **32 columns**, including an **ID column**, a **diagnosis label**, and multiple feature measurements related to tumor characteristics. The key attributes are:

* **id**: Unique identifier for each sample (to be dropped in analysis).
* **diagnosis**: The target variable indicating whether the tumor is **malignant (M)** or **benign (B)**.
* **Feature Groups**:
  + **Mean values** (e.g., radius\_mean, texture\_mean): Representing average measurements of tumor properties.
  + **Standard error values** (e.g., radius\_se, texture\_se): Measuring variability in the feature values.
  + **Worst values** (e.g., radius\_worst, texture\_worst): The most extreme feature values recorded.
* **Other columns**: Unnamed: 32 appears to be an empty or unnecessary column, which should be removed before analysis.

This dataset is widely used in **breast cancer diagnosis**, helping to classify tumors based on extracted cell characteristics.

**Experiment Procedure:**

1. Loading Dataset in the Google Collab Environment
2. Splitting the dataset into Training and testing as per the split mentioned above.
3. Loading the Models and Implementing them on each dataset split

Linear Regression

Ridge Regression

Lasso Regression

1. Converting regression output to binary values.
2. Metrices Calculation for each split based on following parameters:

MSE ( Mean Squared Error )

Accuracy

Precision

1. Plot the Results

**Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ratio** | **Model** | **Accuracy** | **Precision** | **MSE** |
| **60:40** | **Linear**  **Ridge**  **Lasso** | **0.960526**  **0.969298**  **0.960526** | **0.973333**  **0.974026**  **0.961039** | **0.064079**  **0.060919**  **0.069856** |
| **70:30** | **Linear**  **Ridge**  **Lasso** | **0.953216**  **0.964912**  **0.964912** | **0.966102**  **0.967213**  **0.983051** | **0.067284**  **0.059968**  **0.066144** |
| **80:20** | **Linear**  **Ridge**  **Lasso** | **0.956140**  **0.956140**  **0.947368** | **0.975000**  **0.975000**  **0.974359** | **0.064109**  **0.059398**  **0.068186** |
| **90:10** | **Linear**  **Ridge**  **Lasso** | **0.964912**  **0.964912**  **0.964912** | **0.941176**  **0.941176**  **1.000000** | **0.069947**  **0.058609**  **0.057210** |

**Based on the Dataset, following things can be inferred:**

**Accuracy is High Across All Models**

The accuracy values range between 94.73% and 97.03%, indicating that all three regression models (Linear, Ridge, and Lasso) are performing well in distinguishing between malignant and benign tumors.

Ridge Regression consistently provides slightly better accuracy across different train-test splits.

**Lower MSE values suggest better model performance.**

Ridge Regression generally has the lowest MSE, making it the most reliable in minimizing prediction errors.

**Effect of Training Split on Model Performance**

* As the training percentage increases, MSE slightly fluctuates, showing that the dataset is not overly sensitive to different train-test splits.
* Precision values are generally high, with **Lasso Regression achieving a perfect 1.00 precision at 90% training data**, meaning it classifies malignant tumors with **zero false positives** in this case.
* Ridge Regression also maintains high precision across different splits, showing its reliability.

**Graphs:**

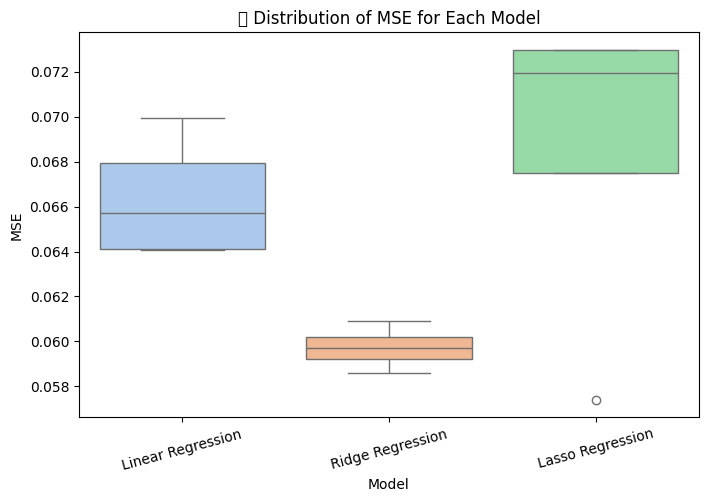
**MSE Graphs**

A graph of a graph of a number of different colored bars

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.



**Accuracy Graphs**

A graph of different colored bars

AI-generated content may be incorrect.

A graph of a graph showing the difference between a train and a train

AI-generated content may be incorrect.

A diagram of a distribution of accuracy

AI-generated content may be incorrect.

**Precision Graphs**

A graph showing a number of green bars

AI-generated content may be incorrect.

A graph with lines and numbers

AI-generated content may be incorrect.

A diagram of different colored boxes

AI-generated content may be incorrect.

**Conclusion:**

* Ridge Regression is the most consistent model, achieving the best balance between accuracy, and MSE.
* Lasso Regression shows competitive precision, especially at higher training percentages, but may have slightly higher errors in prediction.
* Linear Regression performs well but is slightly less stable, as seen in its fluctuating MSE scores.
* Overall, **Ridge Regression** is the recommended model for this dataset, as it consistently minimizes errors while maintaining high accuracy and precision.
* **90% Training Split is the most reliable**