EXPERIMENT 2

Aim of the Experiment

To implement Multiple Linear Regression, Lasso Regression, and Ridge Regression on a real-world HR Employee dataset and compare their performance in predicting a continuous target variable using evaluation metrics such as Mean Squared Error (MSE) and R^2 score.

Theory

1. Regression

Regression is a supervised machine learning technique used to predict a continuous output variable based on one or more input features. It models the relationship between dependent and independent variables.

---

2. Multiple Linear Regression

Multiple Linear Regression is an extension of simple linear regression where multiple independent variables are used to predict a single dependent variable.

Mathematical Equation:

$y=\beta0+\beta1x1+\beta2x2+...+\beta nxny = \beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_{\square}x_{\square}y=\beta0+\beta1x1+\beta2x2+...+\beta nxn$

Where:

- y = predicted output
- β_0 = intercept
- $\beta_1, \beta_2 \dots \beta_{\square}$ = regression coefficients
- $x_1, x_2 \dots x_{\square}$ = independent variables

In this experiment, Multiple Linear Regression is used to predict MonthlyIncome based on employee attributes.

---

3. Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is a regularized regression technique that uses L1 regularization.

It:

- Adds a penalty equal to the absolute value of coefficients

- Can reduce some coefficients to zero
- Helps in feature selection
- Reduces model complexity

Lasso is useful when the dataset contains many features and some may not significantly influence the target variable.

---

4. Ridge Regression

Ridge Regression uses L2 regularization, which adds the square of coefficients as a penalty term.

It:

- Reduces large coefficient values
- Handles multicollinearity
- Prevents overfitting
- Keeps all features but shrinks their influence

Ridge Regression improves model stability and generalization.

Dataset Description

Dataset Name

HR Employee Attrition Dataset

Dataset Type

Real-world corporate HR dataset

Dataset Size

- Number of Records: Approximately 1470 employee records
- Number of Features: Multiple employee-related attributes

Independent Variables (Features)

The dataset contains various employee attributes such as:

- Age
- JobLevel
- YearsAtCompany
- TotalWorkingYears
- Education
- Department
- JobRole
- EnvironmentSatisfaction
- PerformanceRating
- and other HR-related factors

Categorical variables were converted into numerical form using Label Encoding before model training.

Target Variable (Dependent Variable)

Variable	Description
MonthlyIncome	Monthly salary of the employee (continuous value)

MonthlyIncome is continuous and suitable for regression analysis.

Dataset Characteristics

- Contains both categorical and numerical features
- Categorical features encoded into numeric values
- No significant missing values
- Target variable is continuous
- Suitable for applying Multiple Linear, Lasso, and Ridge Regression

---

Experimental Procedure

1. The HR dataset was uploaded and loaded into Google Colab using Pandas.
2. The dataset was explored to check shape and missing values.
3. Categorical variables were converted into numerical values using Label Encoding.
4. The dataset was divided into independent variables (X) and target variable (y).
5. Data was split into training and testing sets in an 80:20 ratio.
6. Multiple Linear Regression model was trained on the training data.
7. Lasso Regression model was trained using L1 regularization.
8. Ridge Regression model was trained using L2 regularization.
9. Predictions were made on the test dataset.
10. Model performance was evaluated using Mean Squared Error (MSE) and R^2 score.
11. The results of all three models were compared.

Performance Evaluation

The following evaluation metrics were used:

Mean Squared Error (MSE)

Measures the average squared difference between actual and predicted values.

Lower MSE indicates better model performance.

R^2 Score

Represents the proportion of variance explained by the model.

- R^2 close to 1 → Good model

- R^2 close to 0 → Poor model

Model Comparison

Model	Description
Multiple Linear Regression	Baseline regression model
Lasso Regression	Performs feature selection using L1 regularization
Ridge Regression	Shrinks coefficients using L2 regularization

---

Conclusion

In this experiment, Multiple Linear Regression, Lasso Regression, and Ridge Regression were successfully implemented on a real-world HR Employee dataset to predict MonthlyIncome based on employee attributes.

The dataset was preprocessed by encoding categorical variables into numerical format. The data was split into training and testing sets to ensure proper evaluation.

- Multiple Linear Regression provided a baseline model and captured the relationship between employee attributes and monthly income.

- Lasso Regression reduced the impact of less important features and simplified the model by shrinking some coefficients toward zero.

- Ridge Regression handled multicollinearity effectively and produced a more stable and generalized model.

Based on the comparison of MSE and R^2 scores, Ridge Regression generally performed slightly better or comparable to Multiple Linear Regression, indicating improved generalization. Lasso Regression, while slightly reducing performance in some cases, helped in reducing model complexity.

Overall, this experiment demonstrates that regularization techniques such as Lasso and Ridge improve model robustness and prevent overfitting in real-world datasets involving multiple correlated features.