Name: Girish Nandanwar	Course: 22CT744 – Lab. Machine Learning
Roll No: A-56	Department: Computer Technology

Practical No.5

Aim: To implement a supervised learning model using Linear Regression to predict a continuous target variable from a given dataset.

Theory:

1. Supervised Learning

Supervised learning is a type of machine learning where the model learns from labeled data — that is, each training example includes both input features and the correct output (target). The model's goal is to learn a mapping function that predicts the target variable for new, unseen inputs.

2. Linear Regression Algorithm

Linear Regression is one of the simplest and most widely used supervised learning algorithms for predicting a **continuous** output variable.

It models the relationship between one or more independent variables ((X_1, X_2, ..., X_n)) and a dependent variable ((Y)) as a linear equation:

[Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \epsilon]

Where:

- (Y) = predicted output (target variable)
- (X_i) = input features
- (\beta_i) = coefficients (weights learned by the model)
- (\beta_0) = intercept term
- (\epsilon) = error term

3. Key Functions and Libraries

- **fetch_california_housing**– Loads the California housing dataset.
- **train_test_split**– Splits data into training and testing sets.
- **LinearRegression**– Fits a linear model to the data.
- **mean_squared_error**and **r2_score**– Evaluate model performance.

4. Evaluation Metrics

- **Mean Squared Error (MSE):**
  [MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2] Lower values indicate better model performance.
- **R-squared (R^2):**

  Measures how well the model explains the variance in the data.
  [R^2 = 1 - \frac{SS_{res}}{SS_{tot}}] A higher R^2 value means better fit.

Code:

Result:

```
  df["Salary"].fillna(df["Salary"].mean(), inplace=True)
Mean Squared Error: 51795891.26317099
R² Score: 0.7609420403238262

Predicted vs Actual:
     Actual     Predicted
8  83000.0  71580.338890
1  48000.0  50645.910369
5  58000.0  62240.067482
```

Conclusion:

A Linear Regression model was successfully trained on the California Housing dataset.
• The dataset contained 20,640 instances and 8 features.
• The model achieved: o MSE ≈ 0.556 o R^2 ≈ 0.576
• The R^2 score indicates that the model explains about 57.6% of the variance in housing prices.
• Performance can be improved using feature engineering, regularized models (Ridge/Lasso), or hyperparameter tuning.