ML Lab Question EXP-7: Regression Models Evaluation with Data Preprocessing and Visualization

Objective:

Apply various regression models to the California Housing Dataset, preprocess the data, and compare the performance of the models using appropriate metrics and visualizations.

Dataset:

You are required to use the California Housing Dataset available in the https://www.kaggle.com/datasets/camnugent/california-housing-prices.

This dataset contains information about various housing attributes in California and is suitable for regression tasks.

Task:

1. Data Preparation and Preprocessing (25 Marks)

- Load the California Housing Dataset.
- Perform Exploratory Data Analysis (EDA) to understand the dataset. This includes:
  - Summary statistics and distribution of features.
  - Identification and handling of missing values.
  - Correlation analysis between features.
- Preprocess the data by:
  - Handling any missing values.
  - Normalizing/standardizing features as needed.
  - Encoding categorical variables (if applicable).
  - Splitting the dataset into training and testing sets.

2. Data Visualization (15 Marks)

- Create visualizations to explore the dataset and the results of the models:
  - Histograms or density plots of feature distributions.
  - Scatter plots to analyze relationships between features and the target variable.

- Correlation heatmap to visualize the relationships between features.
- Residual plots for each regression model to evaluate model performance.
- Bar charts or box plots comparing the performance metrics of the different models.

3. Model Implementation (30 Marks)

- Implement and evaluate the following regression models:
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
  - Decision Tree Regression
  - Random Forest Regression
  - Support Vector Regression (SVR)
- Train each model on the training set and make predictions on the test set.

4. Performance Evaluation (20 Marks)

- Evaluate the performance of each regression model using the following metrics:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R-squared (R^2)
- Present the performance metrics in a tabular format and visualize them using bar charts or other appropriate plots.

5. Discussion and Conclusion (15 Marks)

- Discuss the performance of each model based on the metrics obtained and the visualizations.
- Explain which model performed the best and why, considering the characteristics of the dataset and the models used.
- Provide recommendations for potential improvements or further analysis.

Instructions:

- Use Python and relevant libraries such as `pandas`, `numpy`, `sklearn`, `matplotlib`, and `seaborn` for implementation and visualization.

- Submit a Jupyter Notebook or Python script with well-commented code.

- Include Markdown cells to explain your approach, the results, and any observations or conclusions.

- Ensure that your code is properly formatted, organized, and reproducible.

Evaluation Criteria:

- Thoroughness and accuracy of data preparation and preprocessing (25%)

- Quality and clarity of data visualizations (15%)

- Correct implementation and training of regression models (30%)

- Accurate performance evaluation and presentation of results (20%)

- Depth of discussion and insightful conclusions (15%)