California Housing Price Prediction Using Linear and Ridge Regression

1. Introduction

The rapid growth of urban areas has increased the importance of accurately predicting house prices. Housing price prediction helps governments, real estate companies, and individuals make informed decisions. Machine Learning techniques are widely used for this purpose due to their ability to learn patterns from large datasets.

This project focuses on predicting median house values in California using Linear Regression and Ridge Regression models. The California Housing dataset is used to analyze relationships between various housing features and house prices.

2. Objective

The main objectives of this project are:

- To clean and preprocess the California Housing dataset.
- To analyze feature relationships using correlation analysis.
- To build and train Linear Regression and Ridge Regression models.
- To evaluate and compare the performance of both models using RMSE and R^2 score.

3. Dataset Description

The dataset contains information about housing blocks in California. Each represents aggregated data for housing district.

Key Features:

- longitude – Longitude of the location
- latitude – Latitude of the location
- housing_median_age – Median age of houses
- total_rooms – Total number of rooms
- total_bedrooms – Total number of bedrooms
- population – Population of the district

- households – Number of households
- median_income – Median income of residents
- ocean_proximity – Distance from ocean (encoded)
- median_house_values – Target variable (house price)

4. Data Preprocessing

Data preprocessing is a crucial step to improve the model performance.

Steps performed:

- Removed missing values and inconsistencies.
- Converted categorical feature ocean_proximity into numerical from using encoding.
- Saved the cleaned dataset as cleaned_california_housing.csv.
- Split the dataset into features (X) and target (y).
- Divided data into training (80%) and testing (20%) sets.
- Applied StandardScaler to normalize the features.

5. Exploratory Data Analysis (EDA)

A correlation matrix heatmap was used to understand the relationship among features.

Observations:

- Strong correlation exits among total_rooms, total_bedrooms, households, and population
- median_income shows strong positive correlation with house value
- Location-based features (latitude and longitude) influence pricing trends

EDA helped identify features and multicollinearity in the dataset

6. Model Building

Two regression models were implemented:

6.1. Linear Regression

Linear Regression models the relationship between independent variables and the target using a linear equation. It serves as a baseline model.

6.2. Ridge Regression

Ridge Regression is a regularized version of Linear Regression. It add an L2 penalty term to reduce overfitting caused by multicollinearity.

7. Model Evaluation

The models were evaluated using:

- Root Mean Squared Error (RMSE) – Measures prediction error
- R^2 Score – Measures variance explained by the model

Results

Model	RMSE	R^2 Score
Linear Regression	≈ 0.70	≈ 0.63
Ridge Regression	≈ 0.71	≈ 0.63

Interpretation:

- Both models perform similarly
- Ridge Regression slightly controls multicollinearity
- Around 62-63% of variance in house prices is explained

8. Conclusion

This project successfully demonstrated the application of regression models for housing price prediction. Linear Regression provided a strong baseline, while Ridge Regression helped handle multicollinearity without significant performance loss.

The results indicate that income and location-based features play major role in determining house prices. The model can be further improved using advanced techniques such as Lasso Regression, Polynomial Regression, or ensemble methods.

9. Future Scope

- Hyperparameter running for Ridge Regression

- Use of Lasso and ElasticNet models
- Feature engineering and dimensionality reduction
- Applying advanced models like Random Forest and XGBoost
- Deployment as a web application

10. Tools and Technologies Used

Programming Language: Python

Libraries: Pandas, NumPy, Matplotlib, Seaborn

Machine Learning: Scikit-learn

Environment: Jupyter Notebook

11. References

- Scikit-learn Documentation – https://scikit-learn.org
- California Housing Dataset – https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html
- GeeksforGeeks Machine Learning – https://www.geeksforgeeks.org/machine-learning/