

Project - California Housing Price Prediction

Description :

The US Census Bureau has published California Census Data which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The dataset also serves as an input for project scoping and tries to specify the functional and nonfunctional requirements for it.

Background of the Problem Statement :

The project aims at building a model of housing prices to predict median house values in California using the provided dataset. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

Districts or block groups are the smallest geographical units for which the US Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). There are 20,640 districts in the project dataset.

Domain : Finance and Housing

Dataset Description :

Data Dictionary – Variable and Description

- **longitude** (signed numeric - float) : Longitude value for the block in California, USA
- **latitude** (numeric - float) : Latitude value for the block in California, USA
- **housing_median_age** (numeric - int) : Median age of the house in the block
- **total_rooms** (numeric - int) : Count of the total number of rooms (excluding bedrooms) in all houses in the block
- **total_bedrooms** (numeric - float) : Count of the total number of bedrooms in all houses in the block
- **population** (numeric - int) : Count of the total number of population in the block
- **households** (numeric - int) : Count of the total number of households in the block
- **median_income** (numeric - float) : Median of the total household income of all the houses in the block
- **ocean_proximity** (numeric - categorical) : Type of the landscape of the block
[Unique Values : 'NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND']
- **median_house_value** (numeric - int) : Median of the household prices of all the houses in the block

Dataset Size : 20640 rows x 10 columns

Questions to be answered with analysis :

1. Build a model of housing prices to predict median house values in California using the provided dataset.
2. Train the model to learn from the data to predict the median housing price in any district, given all the other metrics.
3. Predict housing prices based on median_income and plot the regression chart for it.

Project Guidelines :

1. Load the data :

- Read the “**housing.csv**” file from the folder into the program.
- Print first few rows of this data.
- Extract input (X) and output (Y) data from the dataset.

2. Handle missing values :

- Fill the missing values with the mean of the respective column.

3. Encode categorical data :

- Convert categorical column in the dataset to numerical data.

4. Split the dataset :

- Split the data into 80% training dataset and 20% test dataset.

5. Standardize data :

- Standardize training and test datasets.

6. Perform Linear Regression :

- Perform Linear Regression on training data.
- Predict output for test dataset using the fitted model.
- Print root mean squared error (RMSE) from Linear Regression.
[HINT: Import **mean_squared_error** from **sklearn.metrics**]

