PROJECT: California Housing Price Prediction

Purpose of the Document

The purpose of this document is to specify the requirements for the project "California Housing Price Prediction." Apart from specifying the functional and non-functional requirements for the project, it also serves as an input for project scoping.

Problem Statement

The purpose of the project is to predict median house values in Californian districts, given many features from these districts.

The project also aims at building a model of housing prices in California using the California census data. The data has metrics such as the population, median income, median housing price, and so on for each block group in California. 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.

Implementation details

Step 1: Load the dataset

- Loaded the dataset into the program using package pandas and extracted input (X) and output (y) data from the dataset.

Step 2: Handle missing values

- Assigned the missing values with 'mean' of the respective column using Imputer package.

Step 3: Encode categorical data

- Encoded categorical data i.e. ocean_proximity column from dataset using Label encoder and neutralizing the mathematical weightage using OneHotEncoder.

Step 4: Split the dataset

- Performed splitting of dataset into training data (80%) and testing data (20%) using train_test_split.

Step 5: Standardize data

- Standardized data using Standard Scaler.

Step 6: Perform Linear Regression

- Implemented Linear Regression on training data and printed the predicted the output.
- Printed root mean squared error (RMSE) from Linear Regression.

Step 7: Perform Decision Tree Regression

- Implemented Decision Tree Regression on training data and printed the predicted the output.
- Printed root mean squared error (RMSE) from Decision Tree Regression.

Step 8: Perform Random Forest Regression

- Implemented Random Forest Regression on training data and printed the predicted the output.
- Printed root mean squared error (RMSE) from Random Forest Regression.

Step 9: Perform Linear Regression with one independent variable

- Extracted the median_income column from the dataset and performed Linear Regression.
- Printed the predicted the output.
- Visualized the outcome of the Linear Regression on training data and testing data.