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 nonfunctional 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.

Bonus Exercise: Predict housing prices based on median_income and plot the regression chart.

Project Guidelines

#	Step	Process
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 “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)*
7)	Perform Decision Tree Regression	Perform Decision Tree Regression on training data. Predict output for test dataset using the fitted model. Print root mean squared error from Decision Tree Regression.
8)	Perform Random Forest Regression	Perform Random Forest Regression on training data. Predict output for test dataset using the fitted model. Print RMSE (root mean squared error) from Random Forest Regression.

9)	Bonus exercise: Perform Linear Regression with one independent variable	Extract just the median_income column from the independent variables (from X_train and X_test). Perform Linear Regression to predict housing values based on median_income. Predict output for test dataset using the fitted model. Plot the fitted model for training data as well as for test data to check if the fitted model satisfies the test data.