California Housing Price Prediction

Introduction:
The California Housing Price Prediction program aims to predict median house values in California using machine learning techniques. The dataset used in this program, '1553768847_housing.xlsx', contains various metrics such as population, median income, housing age, and more for different block groups in California. By training and evaluating regression models, this program provides insights into the factors influencing housing prices and predicts the median house value based on the given metrics.

Data Loading and Preprocessing:
The program begins by loading the dataset using Pandas and displays the first few rows for initial exploration. The input data (X) consists of all the columns except the target variable 'median_house_value', which is stored separately as the output data (Y). Missing values in the input data are filled with the column mean to ensure data completeness. Categorical data in the 'ocean_proximity' column is encoded using one-hot encoding for further analysis.

Model Training and Evaluation:
The program splits the dataset into training and test sets using the train_test_split function. The input data is standardized using the StandardScaler to ensure feature scaling and prevent any particular feature from dominating the model. The program then trains three regression models:

1. Linear Regression:
The program performs linear regression on the standardized training data and predicts the median house values for the test data. The root mean squared error (RMSE) is calculated to evaluate the model's performance.

2. Decision Tree Regression:
A decision tree regression model is trained using the standardized training data and applied to the test data for prediction. The RMSE is computed to assess the model's accuracy in predicting housing prices.

3. Random Forest Regression:
The program employs random forest regression, which consists of multiple decision trees, to predict housing prices. The RMSE is calculated to measure the accuracy of this ensemble model.

Bonus Exercise: Linear Regression with Median Income:
A separate linear regression model is trained using only the 'median_income' feature. The model predicts house values based on this single independent variable. A scatter plot is generated to visualize the fitted model for both the training and test data, highlighting the relationship between median income and median house value.

Results and Analysis:

The program outputs the RMSE values for each regression model, providing an assessment of their predictive performance. Lower RMSE values indicate better accuracy in predicting median house values. The scatter plot visualizes the fitted linear regression model for the 'median_income' feature, showing how changes in income correspond to changes in housing prices.

Conclusion:

The California Housing Price Prediction program demonstrates the application of regression models to predict median house values based on various metrics. By comparing the performance of linear regression, decision tree regression, and random forest regression, it offers insights into the most effective models for this specific dataset. The bonus exercise focusing on the 'median_income' feature emphasizes its influence on housing prices. This program provides a valuable tool for understanding and predicting housing prices in California, enabling informed decision-making in the real estate market. Future enhancements may involve exploring additional features and evaluating different regression algorithms to further refine the prediction accuracy.