# Exploratory Data Analysis and Machine Learning on California Housing Dataset

**Team Members**

- Durgasi, Ranjitha
- Ettam, Harshitha
- Korapati Murali, Harshitha
- Mengane, Dhawalshree Ashok
- Nallanagula, Nikhitha Reddy

### Abstract

This project conducts Exploratory Data Analysis (EDA) and applies Machine Learning techniques to the California Housing dataset to understand the key factors influencing house prices. A range of regression and classification models are utilized, including Linear Regression, Decision Tree, Random Forest, and Logistic Regression. The analysis aims to uncover relationships between demographic/geographic features and housing prices, and to assess the performance of various ML models in predicting housing value.

### Introduction

The real estate market is an essential component of any economy, and understanding what drives housing prices is crucial for buyers, sellers, and policymakers. In this project, we explore the California Housing dataset to identify important features influencing housing prices and predict both continuous house values and categorical value ranges using supervised machine learning techniques.

### Dataset

The dataset used is the California Housing dataset, which includes variables such as:

- Longitude, Latitude
- Housing Median Age
- Total Rooms, Total Bedrooms
- Population
- Households
- Median Income
- Ocean Proximity (categorical)
- Median House Value (target)

Dataset source: Nugent, C. (2017). California Housing Prices Dataset. Retrieved from https://www.kaggle.com/datasets/camnugent/california-housing-prices
(File used: housing.csv, based on the 1990 California Census)

**Preprocessing**

- Null values were dropped to ensure clean training data.
- Categorical feature ocean_proximity was converted into dummy variables using one-hot encoding.
- Features were scaled using StandardScaler.

- For classification, the continuous median_house_value was categorized into three bins: Low, Medium, and High.

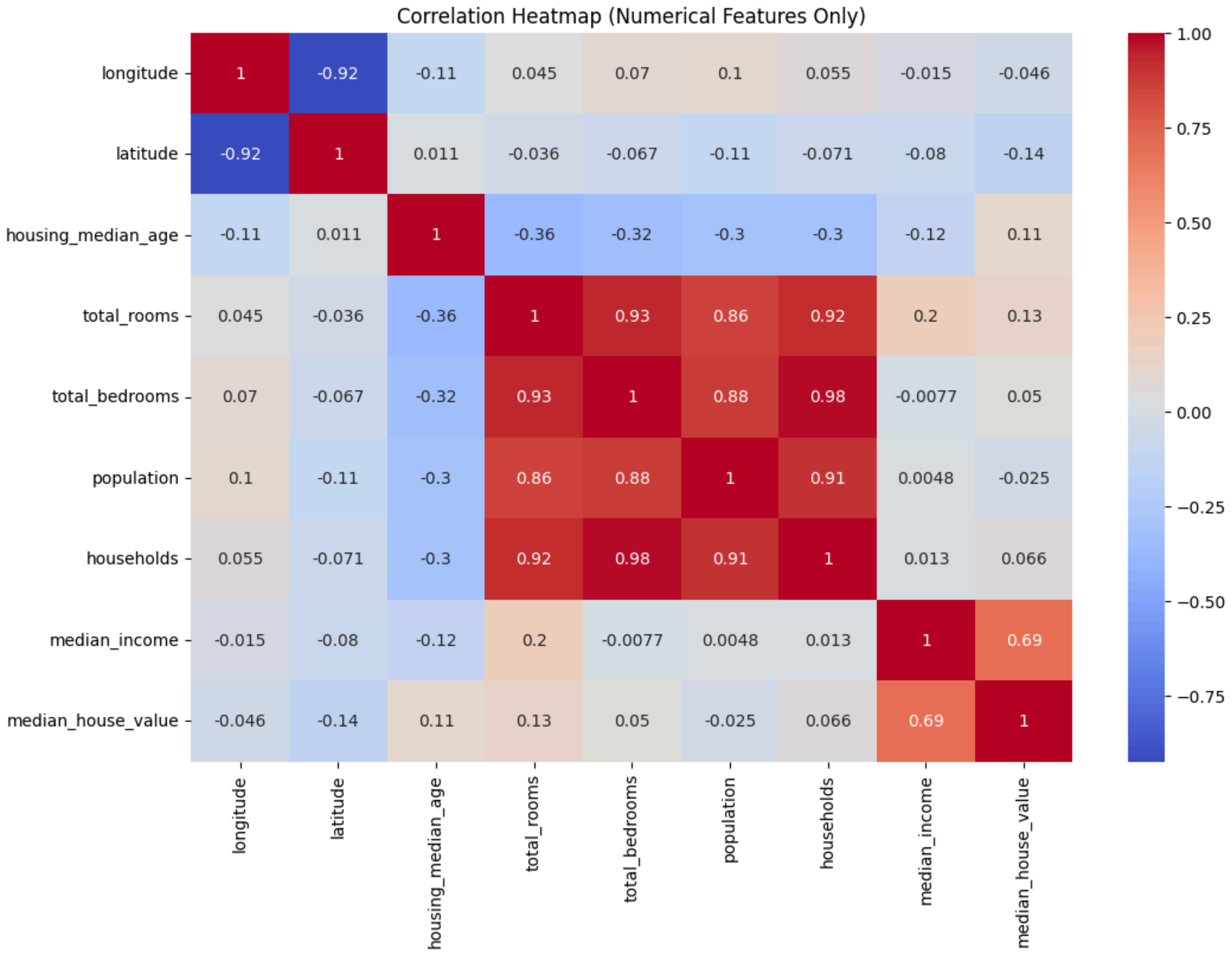
**Literature Review**

Numerous studies have shown that location and income levels are significant indicators of housing prices. Regression models like Linear Regression and ensemble methods such as Random Forest are frequently used in predicting real estate prices due to their ability to model complex interactions.

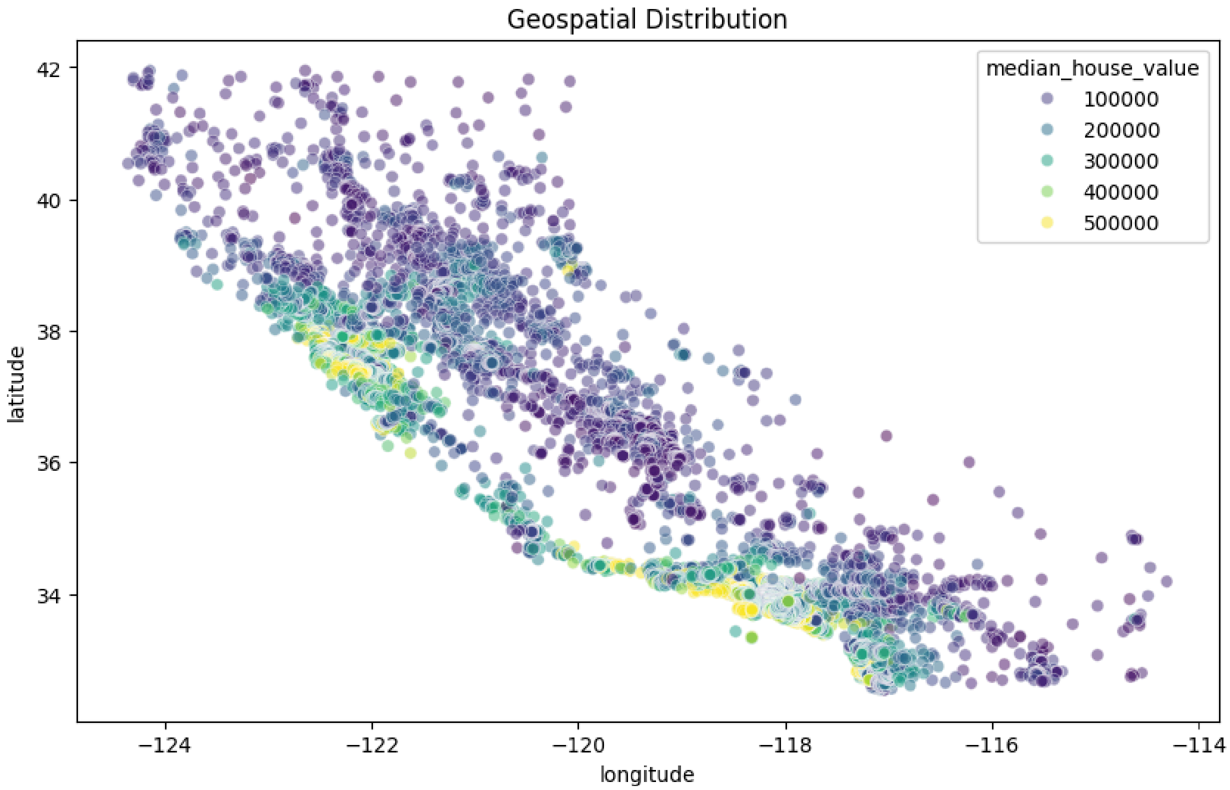
**Architecture / Methodology**

1. **Exploratory Data Analysis (EDA):**

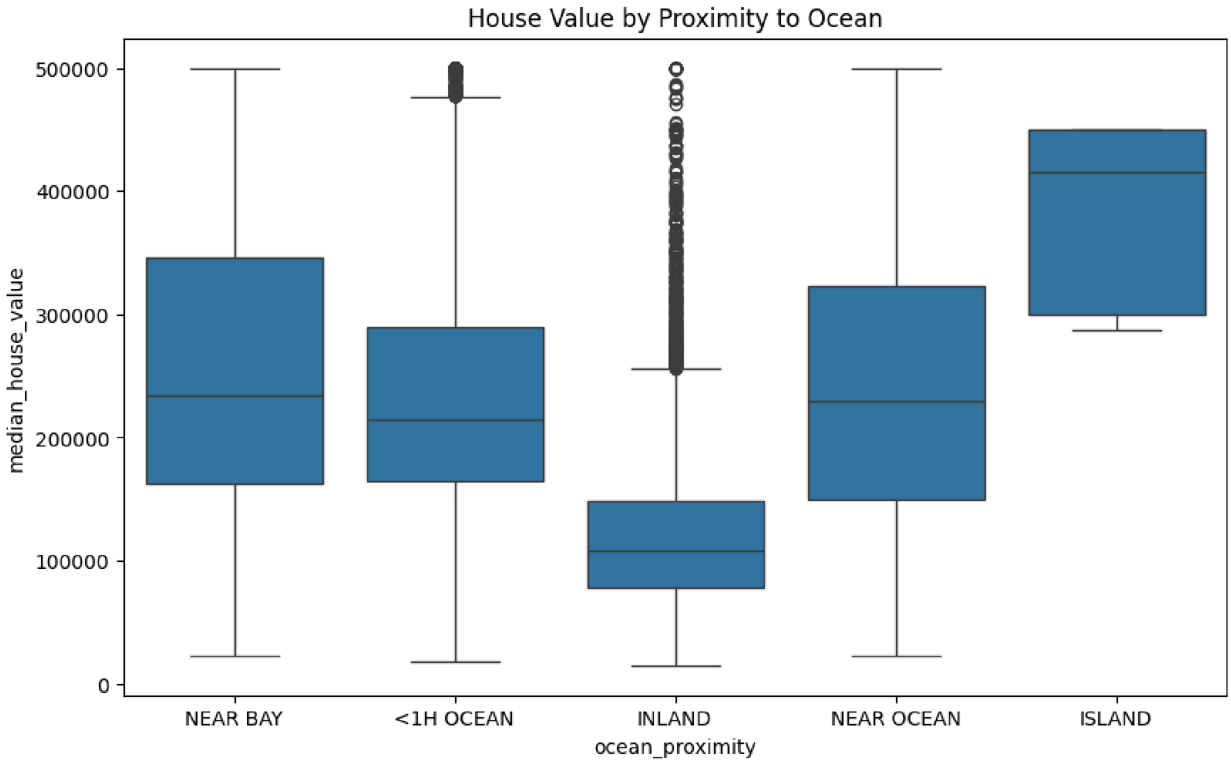
- **Correlation Heatmap:** A heatmap of numerical features revealed that `median_income` has the strongest positive correlation with `median_house_value`, indicating that income is a key factor in housing prices.



- **Geospatial Scatter Plot:** A latitude-longitude scatter plot with color-coded house values demonstrated geographical trends, showing that coastal areas typically have higher house values.



- **Ocean Proximity vs House Value:** A boxplot analysis showed a clear difference in median house values depending on proximity to the ocean, with properties closer to the ocean generally having higher values.



2. **Regression Models:**
   - Linear Regression
   - Decision Tree Regressor (max_depth=10)
   - Random Forest Regressor (n_estimators=100, max_depth=10)
   - Performance was evaluated using R² score and scatter plots of predicted vs actual values.
3. **Classification Models:**
   - Logistic Regression
   - Decision Tree Classifier (max_depth=10)
   - Random Forest Classifier (n_estimators=100, max_depth=10)
   - Performance was evaluated using classification reports and confusion matrices.

## Results

- **Regression Results (R² Score):**
  - Linear Regression: 0.6488

  - Decision Tree: 0.7219
  - Random Forest: 0.7860

- **Classification Results (Precision / Recall / F1-Score):**
  **Logistic Regression:**
  - Accuracy: 0.75
  - Precision: High: 0.80, Low: 0.81, Medium: 0.69
  - Recall: High: 0.59, Low: 0.78, Medium: 0.79

- **Decision Tree:**
  - Accuracy: 0.79
  - Precision: High: 0.82, Low: 0.84, Medium: 0.74
  - Recall: High: 0.67, Low: 0.81, Medium: 0.82
- **Random Forest:**
  - Accuracy: 0.80
  - Precision: High: 0.91, Low: 0.86, Medium: 0.72
  - Recall: High: 0.57, Low: 0.83, Medium: 0.87

- Random Forest showed the strongest performance across both regression and classification tasks, indicating its robustness in modeling complex relationships within the housing data.

## Conclusion

The EDA revealed meaningful patterns and correlations in the California Housing dataset. Among regression models, Random Forest achieved the best performance. For classification, Random Forest again showed strong results. The project's findings demonstrate the predictive power of ML models in real estate data and highlight the importance of median income and location-based variables in determining housing value.

## References

1. Scikit-learn Documentation - https://scikit-learn.org/
2. Seaborn Documentation - https://seaborn.pydata.org/
3. Nugent, C. (2017). California Housing Prices Dataset. https://www.kaggle.com/datasets/camnugent/california-housing-prices

## GitHub Repository for Code

https://github.com/rd39257n/Python-Final