**IMPLEMENTATION AND PLANNING OF CALIFORNIA HOUSING PRICE DATASET**

**Business Idea**: - Predict house price of houses in California

Supervised Regression Project

Domain – Real Estate Domain

**Implementation**

Data Shape – (20640, 9)  
  
**Data Analysis and Cleaning:  
  
Correlation Matrix Analysis:  
  
1. Housing Median Age and Latitude and Longitude :** The housing median age is not strongly correlated with latitude and longitude, which suggests that the age of housing units is not significantly dependent on their geographical location.

**2. Median House Value and Median Income:** There is a strong positive correlation between the median house value and median income. So this is of high priority variable.

**3. Total Rooms and Total Bedrooms:** The total number of rooms and total number of bedrooms are highly positively correlated so we can drop one.

**4. Households and Population:** The number of households and population are highly positively correlated so we can drop one.  
  
**Preprocessing:  
  
Handling missing values in total\_bedrooms:**

The missing values are filled with median as values don’t follow a normal distribution.

**Drop total rooms as total rooms and total bedrooms high positive correlation**

**Outlier Handling**

Handling outliers in housing\_median\_age, total\_bedrooms, population, median income

**One Hot Encoding:**

Perform one hot encoding on ocean proximity as it is a nominal variable.

**Modelling**

* Linear Regression
* Random Forest Regressor
* Gradient Boost Regressor

**MODEL ANALYSIS**

|  |  |
| --- | --- |
| **MODEL** | **Root Mean Squared Error (RMSE)** |
| Linear Regressor | 78,000 |
| Random Forest Regressor | 51,000 |
| Gradient Boost Regressor | 58,000 |

**INFERENCE**

As per the results and model analysis Random Forest Regressor is giving the best prediction.