**Project Proposal**: Predicting Housing Prices in California

**Data Source**: <https://www.kaggle.com/datasets/shibumohapatra/house-price>

**Group Members**: - Suwaiba, Shavez, Gwendoline Grenu, Arfan Hassan

**Overview**:

The objective of this project is to develop a housing price prediction model for California, leveraging machine learning techniques. Using the provided dataset sourced from Kaggle ("California Housing Prices" by Shibumohapatra), our aim is to build a model that can accurately predict the median house values based on various features. The features include the total number of rooms, number of bedrooms, median house age, longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, and ocean proximity.

**Aim of the project**

1. Can we accurately predict housing prices based on the given features?

2. Is there a correlation between median household income and living space?

**Project Phases:**

**1. Data Exploration**

- Utilize Python and Pandas to explore the dataset.

- Check for missing values and understand the distribution of each feature.

**2. Data Preprocessing**

- Handle missing values and outliers appropriately.

- Normalize or scale numerical features if necessary.

- Preprocess data by splitting it into training and testing sets.

- Standardize features using StandardScaler for uniform scale, crucial for distance-based models like Linear Regression.

**3. Data Visualization**

- Use Matplotlib or Plotly to visualize relationships between features and the target variable ("median\_house\_value" column).

**4. Machine Learning Model**

In our predictive modeling approach, the central objective is to estimate the median house value in California based on various features provided in the dataset. This process involves the following key elements:

- The target variable, or the dependent variable, is the "median\_house\_value." This is the variable we aim to predict using our machine learning models. It represents the median value of houses in a specific geographic location

- Features: A set of factors influencing housing prices, including geographical data, housing characteristics, and socioeconomic factors (Longitude, Latitude, Housing median age, Total rooms, Total bedrooms, Population, Household‘sMedian income, Ocean proximity). These factors would be used in training the model.

- Linear Regression: Employed for initial analysis to establish a linear relationship between features and the target variable.

- RandomForestRegressor: Explored for capturing non-linear relationships and improving predictive performance.

**5. Evaluation Metrics**

- Assess model performance using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) for accuracy and insights into model performance.

**6. SQL Integration**

- Integrate an SQLite database for storage and retrieval of data.

- Train models directly on data stored in the database.

- Demonstrate training a Linear Regression model using data retrieved from the database.

**7. Alternative Model**

- Implement a RandomForestRegressor using the same SQL-retrieved data for comparison.

- Evaluate whether a more complex model yields better predictive power.

**9. Web Interface**

- Create a user-friendly dashboard with dropdown menus.

- Allow users to input feature values and receive predicted house prices.

- Use HTML, CSS, Bootstrap for styling, and JavaScript for handling user interactions.

This project will showcase our ability to handle end-to-end machine learning projects, from data exploration to model deployment in a user-friendly web interface.