InternsElite Major Project-

California House Price Prediction

1. Introduction:

Accurately predicting house prices is crucial in real estate, guiding buyers, sellers, and investors. This project uses the California Housing Prices Dataset to predict house prices based on various features like income, house age, and location. We employed Linear Regression and Random Forest Regressor, which are widely used machine learning techniques, to build robust predictive models.

Machine Learning Techniques Used:

- **Linear Regression**: A statistical method that models the relationship between the target (house prices) and input features. It is simple and interpretable, making it a good baseline for understanding basic relationships.
- **Random Forest Regressor**: An advanced model that combines multiple decision trees to improve accuracy and manage complex feature interactions. It's effective for capturing non-linear relationships between variables, making it ideal for this dataset.

How It Helps: These techniques help in identifying the key factors that influence house prices and in developing reliable prediction models. They also provide valuable insights into the real estate market by analysing feature importance and patterns.

2. Problem Statement

The project aims to predict house prices in California using various attributes like location, income, and house characteristics. Accurate predictions can support market analysis, set property values, and guide investment decisions.

3. Methodology

Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the dataset's structure, identify patterns, and determine the relationships between features.

- **Key Findings from EDA**:
    - **Income Matters**: median_income showed a strong positive correlation with house prices, indicating higher incomes are associated with more expensive houses.
    - **Location Impact**: Proximity to desirable locations (indicated by latitude and longitude) significantly affected prices.

Data Preprocessing

- **Data Cleaning**: Missing values were handled, and features were scaled to improve model performance.
- **Scaling**: Key numerical features were scaled to standardize the data, ensuring consistency during model training.

Model Building

- **Training Models**: Both Linear Regression and Random Forest models were trained on the data. Hyperparameters for the Random Forest were optimized to enhance accuracy.
- **Evaluation**: Models were evaluated using RMSE and R2 Score, comparing their performance to determine the best approach.

4. Results and Discussions

Model Performance

- **Linear Regression**:
    - RMSE: Approximately 70,000
    - R2 Score: 0.64
    - **Discussion**: Linear Regression provided a baseline understanding but struggled with complex patterns in the data.
- **Random Forest Regressor**:
    - RMSE: Approximately 50,000
    - R2 Score: 0.82
    - **Discussion**: Random Forest performed significantly better, capturing intricate relationships between features and house prices.

Feature Importance

The Random Forest model highlighted median_income, total_rooms, and housing_median_age as key factors influencing house prices, emphasizing the importance of income and house characteristics.

Visual Insights

- **Scatter Plot**: Showed that higher incomes are generally linked with higher house prices, reinforcing the correlation between wealth and real estate value.
- **Correlation Heatmap**: Helped identify strong positive and negative relationships between features, providing a deeper understanding of factors affecting house prices.

5. Conclusions

The project successfully demonstrated how data analysis and machine learning can predict house prices with significant accuracy. Random Forest emerged as the most effective model, providing robust predictions by effectively handling complex data interactions. The insights gained highlight the importance of income, house characteristics, and location in determining property values, offering valuable guidance for stakeholders in the real estate market.

6. References

1. **California Housing Prices Dataset on Kaggle**: https://www.kaggle.com/datasets/camnugent/california-housing-prices
2. **Scikit-learn Documentation**: https://scikit-learn.org/
3. **Hastie, T., Tibshirani, R., & Friedman, J. (2009)**. The Elements of Statistical Learning. Springer**.**