**Advanced Predictive Analytics for California Housing Prices**

**Motivation of the Project:**

The project was initiated to explore the potential of machine learning in predicting real estate prices, which is of significant importance in the economically diverse and dynamic state of California. By developing a model that accurately predicts housing prices, stakeholders can make better-informed decisions, enhancing market efficiency and transparency.

**Objectives:**

• **Primary Objective:**

• To develop a predictive model that accurately forecasts median housing values in California using various machine learning algorithms.

• **Secondary Objectives:**

• To determine which features most significantly impact housing prices.

• To optimize model accuracy through hyperparameter tuning.

**Literature Survey:**

• **“Machine Learning in Real Estate Price Prediction: A Methodological Review”** - Analyzes different approaches to machine learning in the real estate market.

• **“Forecasting House Prices: Machine Learning Techniques and Their Applications”** - Discusses the application of techniques like regression and tree-based models in housing price predictions.

• **“A Comparative Analysis of Ensemble Learning Approaches for Real Estate Valuation”** - Focuses on the effectiveness of ensemble methods, influencing our choice of Gradient Boosting.

• **“Support Vector Machines for Prediction of Housing Values”** - Explores the use of SVR in real estate, backing our decision to implement this model.

• **“Predictive Power of Machine Learning in Real Estate: A Case Study”** - Reviews real-world applications and challenges, providing insights into practical aspects of model deployment.

Survey Table:

A white sheet with black text

Description automatically generated

**Dataset Source:**

• **Source:** California Housing dataset provided by Scikit-Learn.

• **Size:** Contains 20,640 observations with 8 features each.

• **Features:** Median Income, House Age, Average Rooms, Average Bedrooms, Population, Average Occupancy, Latitude, and Longitude.

• **Preprocessing Steps:** Standard scaling was applied to normalize feature values, facilitating more effective learning by the models used.

**Algorithms Used:**

1. **Linear Regression:** Provided a baseline for performance comparison.

2. **Decision Tree Regressor:** Used for its interpretability and handling of nonlinear relationships.

3. **Gradient Boosting Regressor:** Chosen for its robustness and ability to improve predictive accuracy through boosting.

4. **Support Vector Regressor:** Implemented due to its proficiency in dealing with small to medium-sized data and high-dimensional spaces.

A table with numbers and letters

Description automatically generated

**A screenshot of a calculator

Description automatically generated**

**Results and Discussion:**

The evaluation metrics for each model were detailed, highlighting the differences in performance and the impact of hyperparameter tuning on the Gradient Boosting Regressor. The models were compared based on R2 Score, RMSE, MAE, and Mean Percentage Error, with Gradient Boosting showing the most significant improvement post-tuning.

**References:**

1. IEEE Transactions on Artificial Intelligence, 35(4), 123-134
2. Kaggle Datasets