**Housing Data Analysis Project Report**

**Executive Summary**

This report details the findings and methodologies used to predict housing prices in California. Utilizing machine learning techniques, we aimed to create a robust model that accurately predicts housing prices based on various features such as median income, house age, and location details. The project involved data exploration, preprocessing, model training, and evaluation. The final model achieved a satisfactory prediction accuracy and can be integrated into the business process for better investment decisions.

**1. Introduction**

**Objective**  
The objective of this project is to predict housing prices in California to aid in real estate investment decisions. Accurate predictions will help the company allocate resources effectively and identify potential investment opportunities.

**Dataset**  
The dataset includes features such as median income, house age, total rooms, total bedrooms, population, households, and geographical location details (longitude and latitude).

**2. Data Exploration**

**Objective**  
Understand the dataset's structure and identify any initial patterns or anomalies.

**Approach**

* **Descriptive Analysis**: Used **describe()** and **info()** functions to obtain summary statistics and data types.
* **Graphical Analysis**: Plotted histograms for numerical features to visualize their distributions.

**Findings**

* The dataset contains 20,640 instances with 10 attributes.
* Some features, such as **total\_bedrooms**, had missing values.
* The distribution of **median\_income** is right-skewed, which might affect the model performance.

**3. Splitting Data into Training and Test Sets**

**Objective**  
Ensure robust evaluation by separating data into training and test sets to prevent data leakage.

**Approach**

* **Avoid Snooping Bias**: Split the data early to avoid any bias in model evaluation.
* **Hash-Based Splitting**: Used unique identifiers to ensure consistent data splits across different runs.
* **Stratified Sampling**: Ensured that the training and test sets have a similar distribution of important features like **median\_income**.

**4. Data Visualization**

**Objective**  
Identify relationships between features and uncover any underlying patterns.

**Approach**

* **Correlation Matrix**: Calculated correlation coefficients to identify relationships between features.
* **Heat Maps**: Visualized correlations to easily spot strong relationships.
* **Data Analysis**: Identified potential data issues and insights, such as high correlation between **median\_income** and **median\_house\_value**.

**5. Feature Engineering and Transformation**

**Objective**  
Prepare the data for modeling by handling missing values and transforming features.

**Approach**

* **Data Cleaning**:
  + **Missing Values**: Used **SimpleImputer** to fill missing values with the median.
  + **Categorical Values**: Converted categorical features using one-hot encoding.
* **Custom Transformers**: Created reusable transformers for adding new features, ensuring consistency and reusability.
* **Additional Features**: Added new features like **rooms\_per\_household** and **population\_per\_household** to improve model performance.

**6. Feature Scaling**

**Objective**  
Standardize the features to improve model performance.

**Approach**

* **Standardization**: Scaled features to have a mean of 0 and a standard deviation of 1 to handle outliers effectively.
* **Normalization**: Applied min-max scaling to features to bring them within a 0-1 range.

**7. Model Selection and Training**

**Objective**  
Train various models and select the one with the best performance.

**Approach**

* **Cross-Validation**: Used cross-validation to evaluate models and ensure their robustness.
* **Stratified K-Folds**: Split data into k folds while maintaining the distribution of **median\_income**.

**Models Trained**

* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor

**8. Fine-Tuning**

**Objective**  
Optimize the model performance through hyperparameter tuning.

**Approach**

* **Grid Search**: Conducted an exhaustive search over specified hyperparameter values to find the best combination.
* **Model Evaluation**: Evaluated models using metrics like RMSE (Root Mean Squared Error) to find the best-performing model.

**Best Model**

* Random Forest Regressor: Achieved the lowest RMSE after hyperparameter tuning.

**9. Evaluate on Test Set**

**Objective**  
Assess the final model's performance on the test set to estimate its real-world effectiveness.

**Approach**

* **Final Evaluation**: Used the test set to evaluate the final model, ensuring it performs well on unseen data.

**Results**

* **Final RMSE**: The Random Forest Regressor achieved an RMSE of 47,730.2 on the test set, indicating good predictive performance.

**Conclusion and Recommendations**

**Conclusion**  
The Random Forest Regressor was identified as the best model for predicting housing prices in California. The model's performance on the test set indicates that it can provide accurate predictions, which can be used to make informed investment decisions.

**Recommendations**

* **Integration**: Integrate the model into the business process for real-time housing price predictions.
* **Monitoring**: Continuously monitor the model's performance and update it with new data to maintain accuracy.
* **Future Work**: Explore additional features and advanced models to further improve prediction accuracy.