California Housing Price Classification - Project Analysis Report

Introduction

This report presents the analysis of the California Housing dataset to predict whether house prices are above or below the median price. The analysis involved exploratory data analysis, implementing various classification models, and evaluating their performance.

Techniques Used to Train Models

For this classification task, four supervised learning algorithms were implemented:

1. **K-Nearest Neighbors (KNN)**: A non-parametric, instance-based learning algorithm that classifies data points based on the majority class of their k nearest neighbors. KNN was implemented with hyperparameter tuning to find the optimal number of neighbors using GridSearchCV.

2. **Decision Tree Classifier**: A tree-based method that creates a model resembling a flowchart-like tree structure where internal nodes represent tests on features, branches represent decision rules, and leaf nodes represent class labels. Hyperparameter tuning was performed to optimize tree depth, minimum samples split, and minimum samples leaf parameters.

3. **Random Forest Classifier**: An ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes from individual trees. This model was implemented with key parameters like n_estimators=100, max_depth=20, and min_samples_split=5.

4. **AdaBoost Classifier**: An ensemble boosting classifier that combines multiple weak classifiers to create a strong classifier. The model was implemented with n_estimators=100 and learning_rate=0.1.

All models were trained on a stratified training set (80% of data) and evaluated on a test set (20% of data) to maintain the class distribution of the target variable.

Data Analysis and Visualization

Before implementing the classification models, thorough exploratory data analysis was conducted to understand the dataset characteristics.

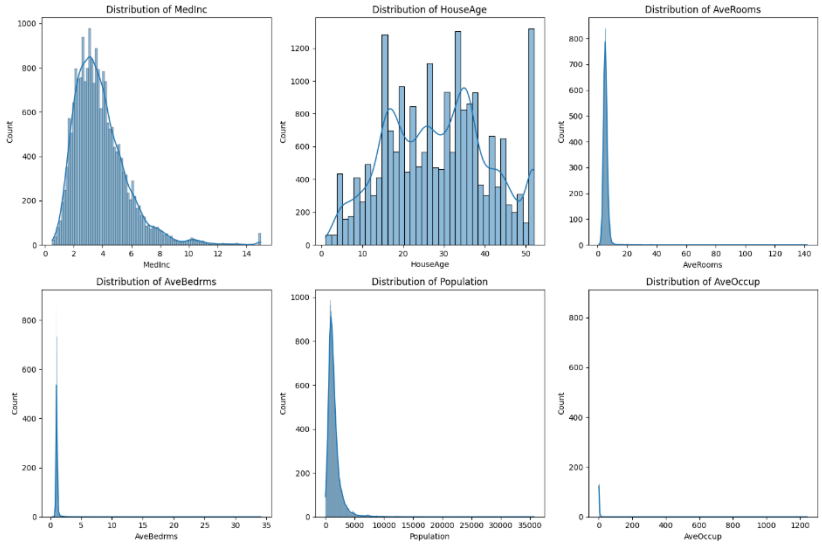

Figure 1: Feature Distributions

The univariate analysis of features revealed: - Median income (MedInc) shows a right-skewed distribution with most values between 2 and 6 - House age (HouseAge) has a multimodal distribution, suggesting different development periods - Average rooms (AveRooms) and bedrooms (AveBedrms) have right-skewed distributions with extreme outliers - Population and average occupancy show highly right-skewed distributions

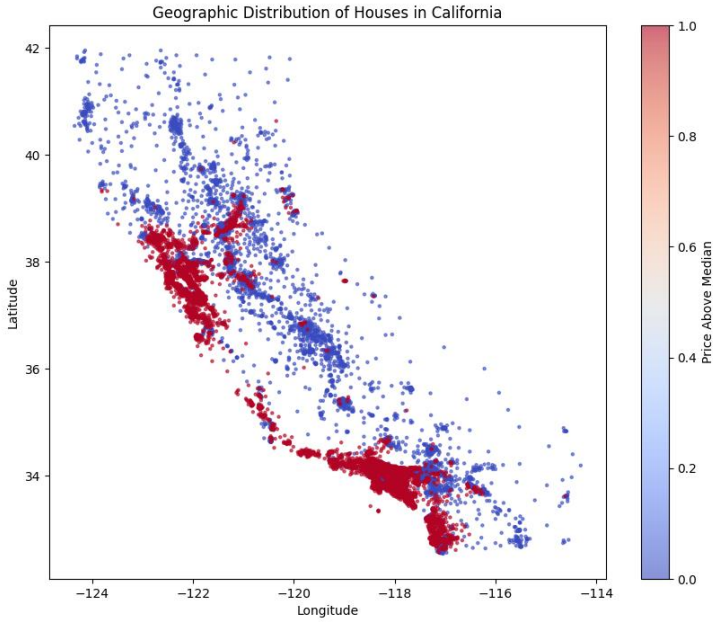

Figure 2: Geographic Distributions

The geographic distribution visualizes housing prices across California, with red points representing above-median prices and blue points representing below-median prices. Clear

clusters of higher-priced homes appear along coastal areas, particularly around San Francisco and Los Angeles.

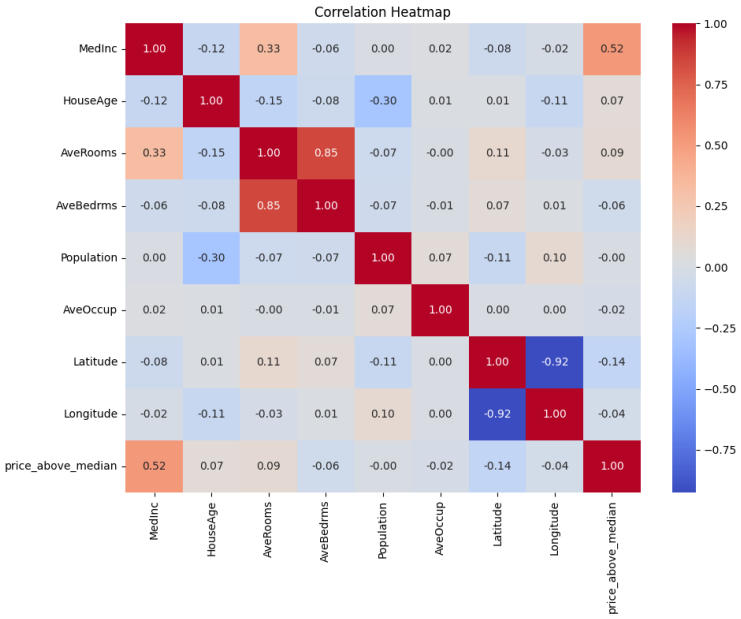

Figure 3: Feature Distributions

The correlation analysis shows that: - Median income has the strongest correlation (0.52) with the target variable (price_above_median) - There's a strong negative correlation (-0.92) between Latitude and Longitude, reflecting California's geography - AveRooms and AveBedrms are highly correlated (0.85)

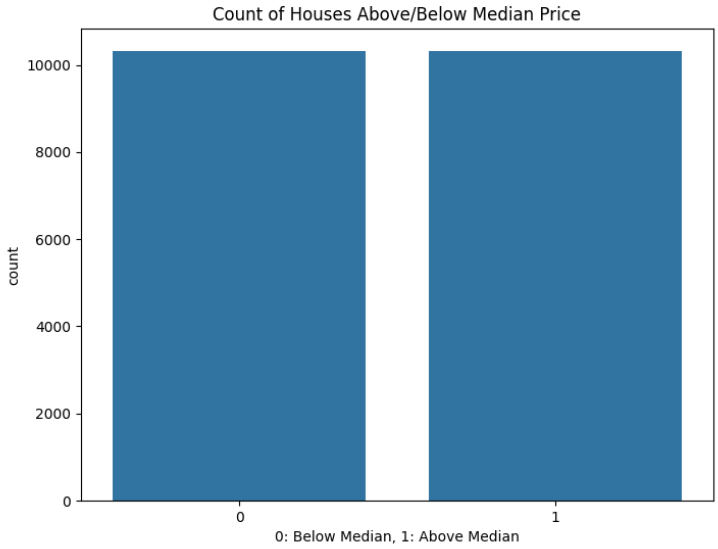

Figure 4: Target Distribution

The target variable is perfectly balanced with 50% of houses priced above the median and 50% below, which is ideal for classification tasks.

Techniques Used to Optimize Model Performance

Several techniques were employed to optimize model performance:

1. **Data Standardization**: Features were standardized using StandardScaler for the KNN algorithm, which is sensitive to the scale of the data. This ensures that all features contribute equally to the distance calculations.

2. **Hyperparameter Tuning**: GridSearchCV was used with 5-fold cross-validation to find optimal hyperparameters for the KNN and Decision Tree models. For Random Forest and AdaBoost, reasonable default parameters were used to reduce computation time while maintaining performance.

3. **Stratified Sampling**: The dataset was split using stratified sampling to ensure that both training and test sets had similar distributions of the target variable (approximately 50% for each class).

4. **Feature Importance Analysis**: For tree-based models, feature importance was analyzed to understand which features contributed most to the predictions.

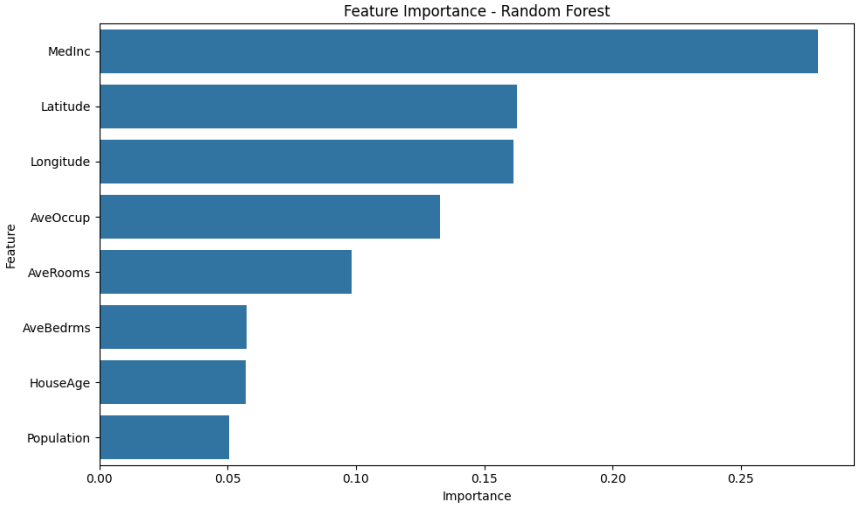

Figure 5: Feature Importance

5. **Metrics Evaluation**: Multiple evaluation metrics (accuracy, precision, recall, F1-score) were used to thoroughly assess model performance beyond just accuracy.

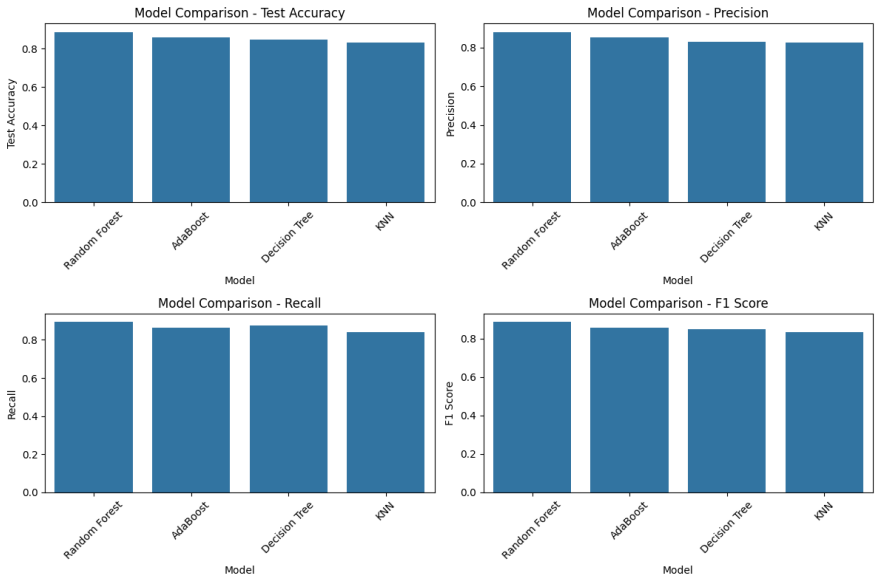

Figure 6: Metrics Evaluation Histograms

Comparison of Model Performance

The confusion matrices for each model provide visual confirmation of their predictive capabilities:

Confusion Matrix - KNN Confusion Matrix - Decision Tree Confusion Matrix - AdaBoost

The performance comparison of all four classification models is summarized below:

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
Random Forest	0.995	0.889	0.887	0.891	0.889
Decision Tree	0.911	0.847	0.828	0.875	0.851
KNN	0.864	0.833	0.827	0.842	0.834
AdaBoost	0.815	0.810	0.821	0.792	0.806

The Random Forest classifier outperformed all other models across all metrics, achieving a test accuracy of 88.9%. This superior performance can be attributed to the ensemble learning approach that reduces overfitting by averaging predictions from multiple decision trees.

The Decision Tree classifier ranked second with a test accuracy of 84.7%, followed by KNN (83.3%) and AdaBoost (81.0%). It’s worth noting that the Random Forest model showed signs of overfitting with a large gap between training accuracy (99.5%) and test accuracy (88.9%), but it still maintained the best generalization performance.

Recommended Model

Based on the comprehensive evaluation, the **Random Forest** model is recommended for this dataset for the following reasons:

1. **Highest Performance**: It achieved the best results across all metrics (accuracy, precision, recall, and F1-score).
2. **Feature Importance Insights**: The model provides valuable insights about feature importance, indicating that median income and geographic location (latitude and longitude) are the most important predictors of housing prices in California.
3. **Robustness**: Random Forest models are generally robust to outliers and non-linear data, which is beneficial for housing data that often contains these characteristics.
4. **Balance Between Classes**: The model performed well for both classes (houses below and above median price), as evidenced by the confusion matrix and balanced precision/recall scores.

If computation time and model simplicity are concerns, the Decision Tree model could be an alternative, as it achieved respectable performance (84.7% accuracy) with a simpler structure.

Most Important Metric for This Dataset

For the California housing price classification problem, the **F1-score** is the most important evaluation metric for the following reasons:

1. **Balance Between Precision and Recall**: The F1-score is the harmonic mean of precision and recall, providing a balance between these two metrics. This is important because in housing price prediction, both false positives (incorrectly predicting above median) and false negatives (incorrectly predicting below median) have consequences.
2. **Class Balance Consideration**: Even though our dataset is balanced with equal distribution of both classes, in real-world applications of housing price prediction, the cost of misclassification might vary. The F1-score helps in assessing the overall effectiveness of the model in such scenarios.
3. **Real-world Application**: In practical applications such as real estate investment or mortgage approval, a balanced measure of the model's ability to correctly identify both above-median and below-median properties is crucial. An investor or lender would want to minimize both:
    - False positives: Investing in properties that are actually below median value but predicted as above
    - False negatives: Missing opportunities in properties that are actually above median value but predicted as below

The Random Forest model achieved the highest F1-score of 0.889, further supporting its recommendation as the best model for this classification task.

Conclusion

This analysis of the California housing dataset demonstrates that the Random Forest classifier is the most effective model for predicting whether house prices are above or below the median. The

most influential factors in this prediction are median income and geographic location, which aligns with real-world understanding of housing markets.

Future work could include engineering additional features such as distance to major cities, trying more advanced models like XGBoost or neural networks, and addressing any class imbalance with techniques like SMOTE.

Note: ChatGPT was used to debug code issues, format the notebook, and reduce runtime by simplifying hyperparameter tuning, but all analysis and model implementation was performed independently.