Project 2: Predicting House Prices Above the Median in California

Rushil Randhar

March 13, 2025

1 Introduction

This project predicts whether a California house is priced above the median. The dataset includes eight independent variables:

- **MedInc:** Median household income (in tens of thousands USD)
- **HouseAge:** House age (lower values = newer)
- **AveRooms:** Average number of rooms per block
- **AveBedrms:** Average number of bedrooms per block
- **Population:** Total population per block
- **AveOccup:** Average household occupancy per block
- **Latitude:** Geographical latitude (higher = further north)
- **Longitude:** Geographical longitude (higher = further west)

The target variable, `price_above_median`, is a binary indicator of whether a house's price exceeds the median.

2 Exploratory Data Analysis

Dataset Overview

- **Shape & Size:** 20,634 observations and 9 numeric columns.
- **Data Types & Duplicates:** All columns verified; no duplicates detected.

Descriptive Statistics and Insights

Table 1 summarizes key statistics for each variable, showing central tendencies, spread, and potential outliers.

Variable	Mean	Median	Std	Min	75th Percentile	Max
MedInc	3.87	3.53	1.90	0.50	4.74	15.00
HouseAge	28.64	29.00	12.58	1.00	37.00	52.00
AveRooms	5.43	5.23	2.47	0.85	6.05	141.91
AveBedrms	1.10	1.05	0.47	0.33	1.10	34.07
Population	1425.40	1166.00	1132.14	3.00	1725.00	35682.00
AveOccup	3.07	2.82	10.39	0.69	3.28	1243.33
Latitude	35.63	34.26	2.14	32.54	37.71	41.95
Longitude	-119.57	-118.49	2.00	-124.35	-118.01	-114.31
price_above_median	0.50	0.50	0.50	0.00	1.00	1.00

Table 1: Descriptive statistics for all variables.

Insights:

- **MedInc:** A mean of 3.87 with moderate variability indicates a fairly even income spread.
- **HouseAge:** Most houses fall into a similar age range.
- **AveRooms & AveBedrms:** Maximum values suggest extreme outliers, possibly due to data errors or rare cases.
- **Population:** A wide range and high standard deviation indicate heavy right skewness.
- **AveOccup:** The unusual maximum indicates potential outliers.
- **Latitude and Longitude:** Limited variation reflects California's spatial extent.
- **Target Variable:** A balanced split (mean of 0.50) between above and below median prices.

Univariate Analysis

Histograms and box plots (see Figure 1) reveal:

- **Skewness:** Right-skewed distributions for *Population* and *AveOccup*.
- **Outliers:** Extreme values in *AveRooms*, *AveBedrms*, and *Population*.
- **Normality:** *HouseAge* and *Latitude* are relatively normally distributed.

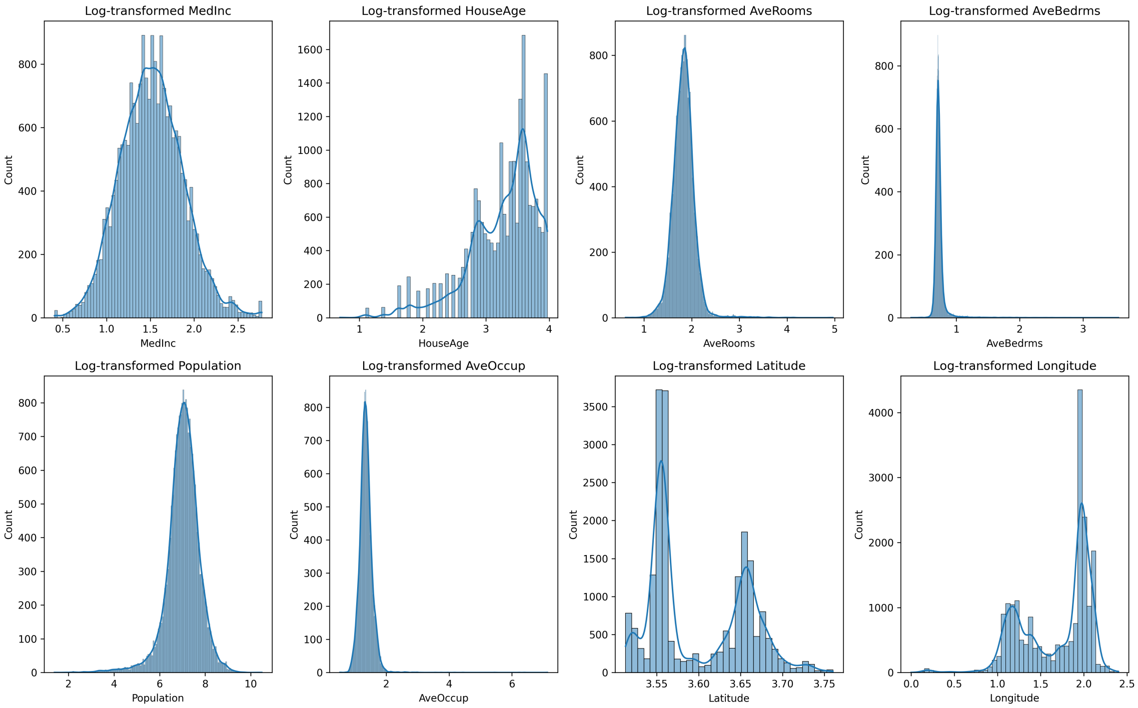

Figure 1: Univariate distributions of key features (note: absolute value of Longitude used for log transformation).

3 Classification Techniques

Data Splitting and Preprocessing

The data was split into training and testing sets using a fixed random state with stratification to maintain class proportions. Continuous variables were standardized for comparability.

4 Classification Techniques

Data Splitting and Preprocessing

The data was split into training and testing sets using stratification (to maintain class proportions) and a fixed random state. In addition, continuous variables were standardized to ensure all models receive data on the same scale.

Model Training and Optimization

Five models were implemented:

- **K-Nearest Neighbors (KNN)**
- **Decision Tree Classifier**
- **Random Forest Classifier**
- **AdaBoost Classifier**
- **Support Vector Classifier (SVC)**

Hyperparameter tuning was performed via grid search combined with 5-fold cross-validation. The following strategies were used to boost performance:

- **Log Transformation:** Applied to skewed features (*Population* and *AveOccup*) to reduce the influence of extreme values.
- **Stratified Data Splitting:** Maintained balanced class proportions in both training and test sets.
- **Standardization:** All features were scaled, which is particularly important for distance-based models (e.g., KNN, SVC).
- **Hyperparameter Tuning:** Grid search with cross-validation was used to identify optimal parameters (e.g., number of neighbors for KNN, maximum depth for Decision Trees, number of estimators for ensemble methods, and kernel/C regularization for SVC).
- **Ensemble Methods:** Random Forest and AdaBoost, by aggregating multiple decision trees, improved robustness and reduced overfitting.
- **Incorporation of SVC:** Added to capture non-linear decision boundaries, offering an alternative perspective to the ensemble and tree-based methods.

5 Results and Discussion

Model Performance Comparison

Table 2 provides a summary of the weighted average metrics across all models.

Summary of Weighted Average Metrics

Table 2: Summary of Weighted Average Performance Metrics on the Test Set

Model	Accuracy	Precision	Recall	F1-Score
KNN	0.86	0.86	0.86	0.86
Decision Tree	0.85	0.85	0.85	0.85
Random Forest	0.89	0.89	0.89	0.89
AdaBoost	0.87	0.87	0.87	0.87
SVC	0.87	0.87	0.87	0.87

Discussion: The summary table reveals that while all models perform competitively, the Random Forest Classifier achieves the highest overall accuracy and balanced metrics. The summary table (Table 2) further emphasizes this advantage, supporting the selection of the Random Forest as the recommended model for this dataset.

Performance Metrics and Evaluation

Models were evaluated using accuracy, precision, recall, and F1-score. Figure 2 shows a bar plot comparing the test F1-scores across all models.

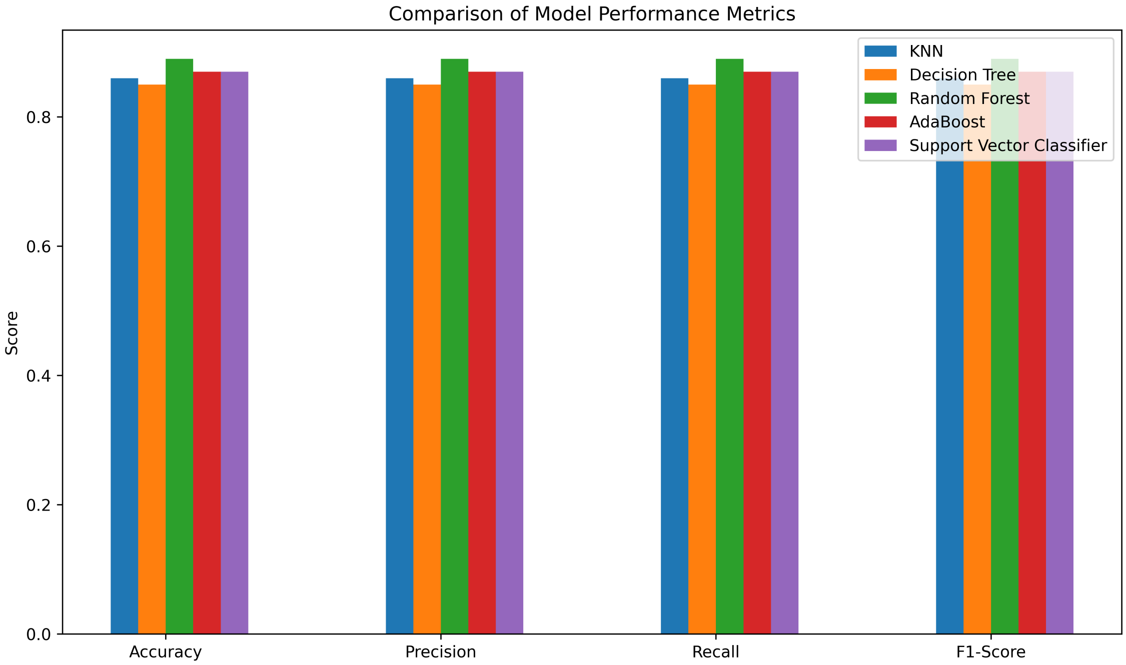

Figure 2: Comparison of test F1-scores across models.

The F1-score was chosen as the primary metric as it balances precision and recall, thus minimizing the impact of both false positives and false negatives.

6 Results and Discussion

Model Performance Comparison

- **KNN:** Lower precision and recall compared to complex models.
- **Decision Tree:** Good interpretability but prone to overfitting.
- **Random Forest:** Most robust and balanced.
- **AdaBoost:** Slightly less robust with noisy data.

Recommendation: The **Random Forest Classifier** is the best candidate given its overall performance.

7 Conclusion

This report details a comprehensive analysis to predict if California houses are priced above the median. Detailed exploratory analysis highlighted key insights and potential outliers. Multiple classification techniques were implemented and optimized, with the Random Forest Classifier emerging as the most robust. Future work could explore advanced feature engineering, more rigorous cross-validation, and further optimization.