A) Problem Statement

To predict median house values in California using both a traditional linear regression model and an ensemble learning model (AdaBoost Regressor with decision trees as base estimators), evaluating and comparing their performance on the California Housing dataset.

B) EDA and Data Pre-processing Steps Taken

→ The fetch_california_housing dataset from sklearn.datasets was loaded as a DataFrame. It includes features such as population and house age, with the target being median house value.

→ Dataset structure was reviewed using .info() and .describe() to check feature types and distributions. A check for missing values confirmed a clean dataset.

→ Visual EDA was conducted through correlation heatmaps, histograms, and boxplots to identify relationships, outliers, and skewed distributions. Scatter plots were also plotted to observe how each feature individually correlates with the target.

→ Outliers were removed using the IQR method for all features. Highly skewed features (absolute skew > 0.75) were log-transformed using np.log1p() to reduce their influence and normalize distributions.

→ The dataset was split into training and test sets using an 80-20 split. A StandardScaler was applied to scale the features since linear models and ensemble methods often benefit from normalized data.

C) Model Training and Evaluation

→ A simple Linear Regression model was trained on the scaled features. Its predictions were evaluated using regression metrics such as R^2 Score, RMSE, MAE, Median Absolute Error, and Explained Variance.

→ A Tuned AdaBoost Regressor was implemented with a base estimator of DecisionTreeRegressor(max_depth=4) and configured with 200 estimators and a lower learning rate (0.1). It was trained on the same scaled features and evaluated using the same metrics for fair comparison.

→ To visually compare how each model performs: Actual vs Predicted plots were generated side-by-side to assess prediction accuracy visually. Residual histograms for both models were plotted to analyze error distributions.

→ Both models were trained using the same training/test split for consistency, and random seeds were fixed for reproducibility.

D) Challenges Faced

→ The AdaBoostRegressor originally used the parameter base_estimator, which caused an error in newer versions of scikit-learn. It was updated to estimator to match the latest API.

→ Outlier removal using the IQR method led to reduced sample size, which slightly impacted model generalizability. Care was taken to apply transformations consistently.

→ When comparing residuals, it was important to ensure predictions were on the same scale, especially after transformations and standardization.

→ Initially, accuracy was added as a custom metric for regression, but later removed for interpretational clarity, as it's not standard for regression models.

E) Summary

Traditional Model Used: Linear Regression

Performance → R^2 Score ~**0.62**

RMSE ~**0.56**

MAE ~**0.42**

Ensemble Model Used: AdaBoost Regressor

Performance → R^2 Score ~**0.57**

RMSE ~**0.6**

MAE ~**0.49**