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 digits dataset from sklearn.datasets was loaded, containing 8x8 pixel images of handwritten digits (0–9). The dataset had no missing values, so we directly moved into exploring and preparing the data.

→ Basic structure and statistics were reviewed using .info() and .describe(), followed by visual checks of class distribution to ensure a balanced dataset. Sample digit images were plotted to confirm label integrity and get a sense of pixel-level patterns.

→ To understand feature relationships, a correlation heatmap was created using seaborn, and pixel intensity patterns were visualized through histograms and boxplots. A pairplot was also used on the first few features to observe how well they separate across classes.

→ The dataset was split into training and test sets (80-20) using stratified sampling to preserve class proportions.

→ For SVM, pixel features were standardized using StandardScaler since SVMs are sensitive to scale. For CNN, image data was normalized (divided by 16.0) and reshaped to (8, 8, 1) to match the expected input shape for convolutional layers.

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**