TASK 2

SUBMITTED BY:

Abinash Kumar Chhatai

1)Which features gave best R²?

After testing various combinations, the best R² was obtained using:

* MedInc (Median Income)
* AveRooms (Average Rooms)
* HouseAge (Housing Age)
* AveOccup (Average Occupancy)

These features showed strong correlation and helped the model explain a significant portion of the variance in house prices.

2 ) What could improve the model?

* We can improve model by using Random Forest .
* Apply feature scaling or transformation .
* Remove outliers or apply clustering to segment similar neighborhoods.

3) Summary of Project

This project aimed to predict California housing prices using a linear regression model built on the California Housing dataset. Exploratory Data Analysis (EDA) revealed that features like median income, average number of rooms, housing age, and average occupancy had significant influence on house prices, with median income showing the strongest correlation. After testing multiple feature combinations, the best model was achieved using four features—MedInc, AveRooms, HouseAge, and AveOccup—resulting in an R² score of approximately 0.62 and a mean squared error (MSE) of around 0.52. The actual vs predicted plot showed a positive linear trend, indicating that the model was able to capture the general pattern of housing prices, although some variance remained unexplained. To improve the model, more advanced techniques like polynomial regression, tree-based models (e.g., Random Forest or XGBoost), and the inclusion of geographic location data could be explored. Additionally, feature scaling and outlier removal may enhance model performance further.

4) Graph of prediction vs actual Predicted House Prices:

