Report: Predicting California Housing Prices Using Linear Regression

The Linear Regression model trained on the California Housing dataset achieved a Mean Squared Error (MSE) of approximately 0.5547 on the test set. This indicates that, on average, the squared difference between the predicted and actual median house values is 0.53 (in units of 100,000 USD). While the model provides a reasonable estimate of house prices, there is some error, likely due to the fact that Linear Regression assumes strictly linear relationships and may not capture complex, non-linear patterns in the data.

Analysis of the model coefficients reveals that median income (MedInc) is the strongest predictor of housing prices, followed by average rooms (AveRooms) and average occupancy (AveOccup). This aligns with expectations, as areas with higher incomes and larger homes tend to have higher housing values. A key challenge observed is that certain features, such as population density or housing age, have weaker linear relationships with house prices, which may limit the model’s predictive accuracy. Overall, Linear Regression provides useful insights into the factors influencing California housing prices, but more advanced models might capture non-linear relationships more effectively.