

“My First End-to-End ML Experience”

Working on the California housing dataset gave me a hands-on opportunity to build a complete machine learning pipeline—from data loading to model evaluation. It wasn't just about coding; it was about understanding how each step connects to the next and how small decisions can impact the final outcome.

I started by exploring the dataset, which included features like median income, house age, and population. Right away, I realized how important it is to understand the data before modeling. I didn't find any missing values, but I did notice outliers in the target variable, which threw off early predictions. To fix this, I used the interquartile range (IQR) method to filter out extreme values. That simple step made a big difference in model stability.

Training two models—Linear Regression and Decision Tree—taught me a lot about how different algorithms behave. The Linear Regression model performed better overall, with a lower Mean Squared Error and higher R^2 score. It captured the linear relationships in the data well. The Decision Tree, on the other hand, tended to overfit, especially in areas with high variance like population and average room count. Seeing that contrast helped me appreciate the importance of model selection and the trade-offs between simplicity and flexibility.

One of the biggest challenges I faced was visualizing the results in a way that made sense. Scatter plots of predicted vs. actual values helped me see where each model struggled—especially with extreme cases. That pushed me to think about future improvements, like using ensemble methods or log-transforming the target variable to reduce skew.

Beyond the technical side, this project helped me communicate results clearly. Writing summary cells and reflecting on model performance made me realize how important it is to explain insights—not just generate numbers.

If I were to apply this pipeline in the real world, I'd use it to predict rental prices in urban areas. With features like location, number of rooms, and proximity to public transport, a similar workflow could help renters find fair prices or help platforms offer dynamic pricing. The same steps—data cleaning, model comparison, and evaluation—would apply, making this project a solid foundation for practical applications.

Overall, this experience sharpened my skills and gave me confidence in building machine learning solutions that are both accurate and explainable. It also reminded me that every dataset tells a story—and it's up to me to listen, clean, model, and share it.