Escopo do Projeto

1. Definição do Problema

- Descrição: o objetivo é treinar um modelo de machine learning para prever uma variável contínua com base em um conjunto de variáveis preditoras usando o dataset “California Housing” do UCL Repository que contém informações sobre características de casas e seus preços.

- Premissas e Hipóteses: Acredito que o preço da casa será influenciado pelas características da localidade, número de quartos e a proximidade de centros urbanos.

- Restrições do Dataset:

- Longitude: coordenada geográfica;
- Latitude: coordenada geográfica;
- Housing_median_age: idade média das casas;
- Total_rooms: número total de quartos;
- Total_bedrooms: número total de dormitórios;
- Population: população de residências;
- Median_income: renda média;
- Median_house_value: preço médio das casas.

2. Preparação do Dados

- Carregar o dataset: utilizei o dataset “California Housing” que foi obtido do UCI Repository;

- Análise Exploratória: parte inicial do dataset, procurando por dados ausentes e/ou valores discrepantes;

- Separação entre treino e teste: separei os dados em 80% para treinamento e 20% para teste;

- Validação Cruzada: utilizei deste método para poder estimar o desempenho do modelo, pois ajuda a evitar o overfitting dando uma avaliação mais confiável para validar os resultados.

- Transformação de Dados: padronizei os dados para que todas as variáveis tivessem uma escala comparável.

3. Modelagem e Treinamento

Selecionei os algoritmos de Regressão Linear (para ter uma linha de base simples), Random Forest Regressor (para buscar relações não-lineares) e o por último o Gradien Boosting Regressor (para melhorar o desemprenho dos resultados).
Os modelos foram usados com os hiperparâmetros padrão, logo, tentei otimizar os parâmetros com a validação cruzada, sendo este ajuste realizado através do GridSearchCV para obter um melhor desempenho dos modelos.

4. Avaliação do REsultados:

Para poder avaliar os modelos, utilizei o MSE (para avaliar o erro quadrático médio) e o MAE (para medir o erro absoluto médio)

5. Conclusão: a

o rodar o código pude observar que o melhor resultado foi da Regressão Linear, sendo este a melhor opção para o modelo final, pois teve um desempenho significativo em relação ao Random Forest e ao Gradient Boosting.