Relatório do Treinamento de Regressão - Boston Housing Dataset

O **Boston Housing Dataset** contém informações sobre imóveis na cidade de Boston, com 13 variáveis independentes (features) e o preço médio das casas (`MEDV`) como variável alvo.
link dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data

Descrição das Variáveis

Coluna	Significado
CRIM	Taxa de criminalidade por cidade
ZN	Proporção de terrenos residenciais com lotes > 25.000 sq.ft.
INDUS	Proporção de acres comerciais não varejistas por cidade
CHAS	Variável dummy Charles River (1 se próximo ao rio, 0 caso contrário)
NOX	Concentração de óxidos nítricos (partes por 10 milhões)
RM	Número médio de quartos por residência
AGE	Proporção de unidades ocupadas pelos proprietários construídas antes de 1940
DIS	Distâncias ponderadas até cinco centros de emprego em Boston
RAD	Índice de acessibilidade a rodovias radiais
TAX	Taxa de imposto sobre propriedade de valor total por $10.000
PTRATIO	Proporção aluno-professor por cidade
B	Proporção de população negra por cidade
LSTAT	Percentual da população de menor status
MEDV	Valor médio das casas (em milhares de dólares) – variável alvo

Pré-processamento

1. **Tratamento de valores ausentes**
    - Colunas numéricas preenchidas com a média.
    - Colunas categóricas preenchidas com a moda.
2. **Remoção de duplicados**
    - Linhas exatamente iguais foram removidas para evitar viés no modelo.
3. **Normalização**
    - Todas as features numéricas foram normalizadas (StandardScaler) para padronizar os valores antes do treinamento.

Comparativo de Modelos de Regressão

Foram testados três modelos:

Modelo	R^2	RMSE	MAE	Interpretação
Random Forest	88%	2.92	2.04	Melhor modelo, alta precisão e boa capacidade de explicar variação
KNN	72%	4.54	2.59	Médio, erra mais que Random Forest
Regressão Linear	67%	4.93	3.19	Mais simples, menos preciso, não captura bem relações não lineares

Conclusão: O **Random Forest Regressor** apresentou o melhor desempenho, sendo escolhido para otimização e seleção de features.

Seleção de Features

- Foi treinado um **Random Forest com GridSearch** para encontrar melhores hiperparâmetros (`n_estimators`, `max_depth`, `min_samples_split`).
- A importância das features foi avaliada, e um limiar de **0.01** foi usado para selecionar as mais relevantes.

Treinamento Final com Features Selecionadas

O modelo foi re-treinado usando **apenas as features selecionadas** e GridSearch para otimização.