**INSTITUTO FEDERAL DO PARANÁ**

**CAMPUS PINHAIS**

**BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**Ciro Guilherme Nass**

**PREVENDO PREÇOS DE CASAS COM ANÁLISE PREDITIVA**

**relatório de atividade**

**pinhais**

**2025**

**Ciro Guilherme Nass**

**PREVENDO PREÇOS DE CASAS COM ANÁLISE PREDITIVA**

**relatório de atividade**

Relatório de Atividade desenvolvido no Curso de Bacharelado em Ciência da Computação, do Instituto Federal do Paraná – Campus Pinhais, para validação em atividade avaliativa na disciplina de Big Data.

Responsável:

Profa. Dra. Eliana Santos

**PINHAIS**

**2025**

**Sumário**

[**1**](#_7qbs31332mp8) **INTRODUÇÃO 3**

[**2**](#_uofty4fmxvv) **METODOLOGIA 4-5**

[**3**](#_1ryr1cbkl3sz) **QUESTÕES TEÓRICAS** [**6**](#_3nuhxyta8he9)

[**4**](#_3nuhxyta8he9) **RESULTADOS** [**7**](#_hlhnbrje1zu1)**-**[**8**](#_w75uqt7vmb8y)

[**5**](#_hlhnbrje1zu1) **DISCUSSÃO** [**9**](#_jftoh0kwma8p)

[**6**](#_w75uqt7vmb8y) **CONCLUSÃO 10**

[**REFERÊNCIAS**](#_jftoh0kwma8p) **11**

# INTRODUÇÃO

A previsão dos preços de casas é crucial no mercado imobiliário. Essa previsão ajuda investidores a investir com mais segurança e a entender quais características (features) de uma propriedade são mais valorizadas. A relevância está em transformar dados brutos em informação de valor financeiro para tomada de decisão.

O problema abordado é a Regressão, onde a meta é prever um valor contínuo: o preço final de venda (SalePrice).

Os dados utilizados são um *dataset* comum no campo de Machine Learning, contendo 1460 observações de casas e 81 colunas com diversas características (área, número de quartos, qualidade de acabamento, etc.). Com a análise, buscamos construir um modelo de Machine Learning que consiga prever o SalePrice com a maior precisão possível, com base nas características das casas.

# METODOLOGIA

O desenvolvimento do modelo preditivo seguiu um fluxo de trabalho estruturado, começando pela limpeza e preparação dos dados e culminando na construção e avaliação do modelo de regressão.

A fase inicial focou na Análise e Tratamento de Dados Ausentes/Nulos. Primeiramente, realizamos a identificação de colunas com grande volume de dados faltantes, como PoolQC e MiscFeature. Em seguida, aplicamos uma estratégia de remoção mais radical, eliminando colunas que possuíam menos de 1000 valores preenchidos. Com isso, o *dataset* teve seu número de colunas reduzido de 81 para 75. Após a remoção, a etapa de Imputação, ou seja, preenchimento, foi executada para os nulos restantes nas colunas numéricas, utilizando a mediana. A mediana foi a escolha feita por ser mais robusta a *outliers* do que a média.

Na Seleção e Preparação dos Dados, o *dataset* foi filtrado para incluir apenas as 37 colunas numéricas, simplificando o escopo inicial do modelo ao focar exclusivamente nesses tipos de *features*. Definimos o conjunto de *features* (X) e a variável alvo (y, SalePrice). Para a validação, realizamos a Separação Treino e Teste, dividindo o *dataset* em 80% para Treino e 20% para Teste (train\_test\_split com random\_state=42), garantindo que o modelo fosse avaliado em dados que ele nunca viu.

Por fim, na etapa de Construção e Treinamento do Modelo, optamos pelo poderoso algoritmo RandomForestRegressor. O modelo foi treinado com o conjunto de Treino. A Avaliação de desempenho foi feita no conjunto de Teste, utilizando as métricas Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R2) para mensurar sua precisão.

Na etapa de imputação, optamos pelo uso da Mediana para Preencher Valores Ausentes em vez da média. Essa escolha se justifica porque o *dataset* de preços de casas é altamente suscetível a valores extremos (*outliers*), como imóveis de altíssimo luxo. A média seria puxada para cima por esses valores atípicos, distorcendo a representação do preço típico e resultando em uma imputação menos precisa. A mediana, por ser mais robusta a esses extremos, forneceu uma estimativa de valor ausente mais segura.

## Foi selecionado o RandomForestRegressor (Regressão por Floresta Aleatória) para a modelagem. Este modelo de *ensemble* foi escolhido por sua versatilidade e poder preditivo. O *Random Forest* é conhecido por entregar um bom desempenho "out-of-the-box", sem a necessidade de otimização excessiva. Além disso, ele é inerentemente robusto ao *overfitting* (em comparação com árvores de decisão isoladas) e lida eficientemente com a não-linearidade e a complexidade inerente aos dados de preços de casas.

## 

## 

# QUESTÕES TEÓRICAS

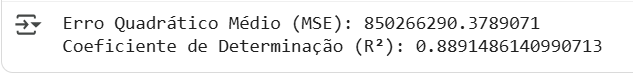
Para início, é importante destacar sobre como foi decidida a escolha do valor para o preenchimento dos nulos. Foi definido a partir de um valor central capaz de representar o todo. A média não foi escolhida porque ela obtida pela soma de todos os números dividida pela contagem total, representando o centro de gravidade do conjunto. Já a mediana foi mais adequada, porque é o valor que divide um conjunto de dados ordenado exatamente ao meio, com 50% dos dados sendo menores e 50% maiores que ela. Por ser uma medida de posição, a mediana é considerada mais segura quando há valores extremos (*outliers*) nos dados. Isso ocorre porque a média é diretamente influenciada por valores muito grandes ou muito pequenos, sendo "puxada" na direção desses extremos, enquanto a mediana permanece estável e representa melhor o valor típico do conjunto.

Em Machine Learning, trabalhamos com variáveis categóricas, que são aquelas que representam categorias, como "Tipo de Garagem" ou "Bairro". O desafio é que a maioria dos algoritmos de ML exige entradas numéricas. Para contornar isso, o tratamento mais comum é o One-Hot Encoding (transformar cada categoria em uma nova coluna binária, com valores 0 ou 1) ou o Label Encoding (atribuir um número inteiro a cada categoria).

No processo de avaliação do modelo, é crucial testá-lo com dados que ele nunca viu antes. Isso garante que o modelo não apenas memorizou os dados de treinamento que é um fenômeno conhecido como overfitting (sobreajuste), mas que ele realmente aprendeu a tendência geral e é capaz de fazer previsões precisas em novas situações. O *overfitting* ocorre quando um modelo aprende demais os detalhes e o ruído específicos do conjunto de treino, resultando em um desempenho quase perfeito no treino, mas um desempenho muito ruim no mundo real. A métrica obtida no conjunto de teste é, portanto, a melhor estimativa do desempenho real do modelo após sua implantação.

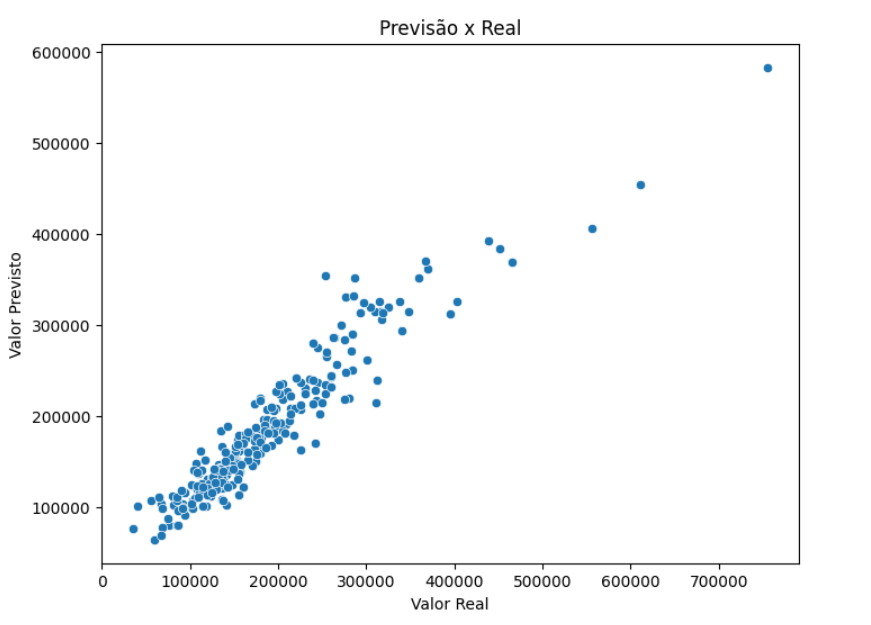
# RESULTADOS

O modelo RandomForestRegressor treinado e testado nos dados numéricos forneceu as seguintes métricas de avaliação:



O valor de R2 de 0.889 indica que o modelo consegue explicar aproximadamente 89% da variância total nos preços das casas no conjunto de teste.

O gráfico de dispersão abaixo compara os valores previstos pelo modelo com os valores reais de venda:



O gráfico mostra que a maioria dos pontos está concentrada muito próxima da linha diagonal ideal (y=x), o que visualmente confirma o alto valor de R2. Isso significa que o modelo, na maioria das vezes, acerta a previsão ou erra por uma margem pequena.

No entanto, há alguns pontos de desvio significativo, especialmente para casas com valores reais mais altos (acima de $500.000). Nesses casos, o modelo tende a subestimar o preço real, indicando que pode estar falhando em capturar as características que tornam essas casas de luxo mais caras.

# DISCUSSÃO

O modelo alcançou um excelente desempenho inicial, com um R2 de quase 89%. Seu ponto forte reside na escolha do Random Forest, que se provou um algoritmo robusto, complementado por um tratamento de dados inicial eficaz. Entretanto, podemos observar limitações críticas. A principal delas é que o modelo foi treinado apenas com 37 *features* numéricas, ignorando variáveis categóricas altamente preditivas para preços de casas, como Bairro e Qualidade Geral. Além disso, o modelo ainda demonstra dificuldade em prever preços muito altos (os *outliers*).

Para melhorar os resultados e buscar um R2 superior a 90%, sugiro: Incluir Variáveis Categóricas aplicando One-Hot Encoding para adicionar o contexto de localização e acabamento; Otimização de Hiperparâmetros utilizando métodos como Grid Search para refinar o *RandomForestRegressor*; Tratamento de Outliers na variável alvo (SalePrice); e Engenharia de Recursos (*Feature Engineering*) para criar *features* úteis, como a idade da casa (AnoVenda - AnoConstrucao).

# CONCLUSÃO

Esta atividade proporcionou uma visão completa e prática do pipeline de análise preditiva. Essa tarefa ensinou uma lição muito importante, a relevância do Pré-processamento de Dados para uma análise e previsão mais precisa. A atividade reforçou a importância da avaliação crítica, onde foi essencial interpretar o gráfico de dispersão para entender *onde* e *por que* o modelo está falhando. Essa análise tem aplicações práticas diretas, podendo ser o motor por trás de um Sistema de Avaliação Imobiliária Automatizada (AVM) é uma ferramenta essencial para otimização de investimento, ajudando investidores a identificar propriedades subvalorizadas ou supervalorizadas no mercado imobiliário e contribuir para uma melhor decisão na hora de investir.

# REFERÊNCIAS

SCIKIT-LEARN. **Scikit-learn documentation**. Disponível em:<https://scikit-learn.org/stable/>. Acesso em: 05 out. 2025.

PANDAS. **Pandas documentation**. Disponível em:<https://pandas.pydata.org/docs/>. Acesso em: 05 out. 2025.

KAGGLE. **House Prices - Advanced Regression Techniques**. Disponível em:<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>. Acesso em: 05 out. 2025.