Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística e Departamento de Informática
Programa de Especialização em *Data Science* e *Big Data*

João Batista de Andrade Junior

# Avaliação de modelos de regressão para previsão de preços de imóveis em Curitiba

Curitiba
2025

João Batista de Andrade Junior

# Avaliação de modelos de regressão para previsão de preços de imóveis em Curitiba

Artigo apresentado ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientadora: Prof. Dra Deisy Morselli Gysi

Curitiba

2025

# Avaliação de modelos de regressão para previsão de preços de imóveis em Curitiba

***Evaluation of regression models for predicting real estate prices in Curitiba***

João Batista de Andrade Junior[1], Deisy Morselli Gysi[2]

[1] Aluno do programa de Especialização em Data Science & Big Data, batist4joaoo@gmail.com

[2] Professora do Departamento de Estatística - DEST/UFPR, deisy.gysi@ufpr.br

**Resumo**

Este trabalho tem como objetivo analisar e comparar o desempenho de diferentes modelos de regressão aplicados na previsão de preços de imóveis residenciais na cidade de Curitiba - Paraná. A partir de uma base de dados, retirada do Kaggle, contendo 18.760 registros com características dos imóveis, como área, número de quartos, vagas de garagem, localização e outras comodidades, foram aplicados seis modelos de regressão: Regressão Linear, Regressão Ridge, Regressão Lasso, Random Forest, Gradient Boosting e Redes Neurais Artificiais. O processo incluiu etapas de limpeza dos dados, tratamento de outliers, transformação da variável dependente, criação de variáveis dummies e seleção de atributos relevantes. Os modelos foram avaliados com métricas como MAE, RMSE, $R^2$ e MAPE. Como resultado, observou-se que os modelos baseados em árvores, especialmente o Random Forest e o Gradient Boosting, apresentaram melhor desempenho na previsão dos preços em comparação aos modelos lineares tradicionais.
**Palavras-chave:** Data Science, Machine Learning, Previsão de Preços, Modelos de Regressão, Mercado Imobiliário.

**Abstract**

This study aims to analyze and compare the performance of different regression models applied to residential property price prediction in the city of Curitiba, Paraná. Using a dataset sourced from Kaggle containing 18.760 records with features such as area, number of rooms, garage spaces, location, and other amenities, six regression models were applied: Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Gradient Boosting, and Artificial Neural Networks. The process included data cleaning, outlier treatment, transformation of the dependent variable, creation of dummy variables, and selection of relevant features. The models were evaluated using metrics such as MAE, RMSE, $R^2$, and MAPE. As a result, tree-based models—particularly Random Forest and Gradient Boosting—demonstrated better performance in predicting prices compared to traditional linear models.
**Keywords:** Data Science, Machine Learning, Price Prediction, Regression Models, Real Estate Market.

## 1 Introdução

A precificação de imóveis é um desafio comum tanto para profissionais do mercado imobiliário quanto para compradores e investidores. O preço de um imóvel é definido por características físicas como o tamanho do imóvel, número de quartos, localização, infraestrutura disponível na região, e também por fatores macroeconômicos - estes mais difíceis de mensurar e prever, como taxas de juros, inflação, confiança do consumidor e oferta de crédito imobiliário [1]. O uso de tecnologias avançadas para auxiliar a entender a precificação de imóveis surge como uma poderosa ferramenta para lidar com essa complexidade e gerar estimativas mais precisas. Ao analisar os dados e identificar padrões de influência entre variáveis, essas abordagens oferecem suporte valioso na definição de preços de imóveis.Nesse âmbito, o uso de modelos estatísticos e de machine learning para previsão de preços tem crescido nos últimos anos, uma vez que permite maior precisão na tomada de decisão [2].

Este trabalho tem como objetivo aplicar e comparar diferentes modelos de regressão no contexto de imóveis da cidade de Curitiba, visando prever o preço de venda de imóveis a partir de suas características. O trabalho foi dividido em algumas seções. A Seção 2 apresenta uma discussão sobre os modelos de regressão empregados e as etapas de preparação dos dados. Na Seção 3, são detalhados os materiais e etapas de tratamento da base de dados. A Seção 4 detalha a etapa de seleção de variáveis. A Seção 5 traz com mais detalhes os métodos e os procedimentos adotados. A Seção 6 discute os resultados obtidos com os modelos. Por fim, a Seção 7 apresenta as conclusões e sugestões para trabalhos futuros.

## 2 Discussão

O mercado imobiliário apresenta características que tornam o problema de previsão de preços particularmente desafiador: a influência de múltiplos fatores, tanto estruturais (como área, número de quartos, presença de garagem) quanto locacionais (bairro, proximidade de serviços, segurança, entre outros). Além disso, o comportamento dos preços pode ser influenciado por não linearidades, outliers e interações entre variáveis. Então, para o desenvolvimento deste trabalho, foram considerados seis modelos de regressão amplamente utilizados em problemas de previsão de preços, cada modelo oferece abordagens diferentes, com vantagens e limitações específicas para este tipo de problema.

### *2.1 Regressão Linear*

A Regressão Linear é um dos métodos estatísticos mais antigos e amplamente utilizados para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. O objetivo é estimar os coeficientes da equação linear que melhor se ajusta aos dados, minimizando a soma dos quadrados dos erros [3]. No caso mais simples, com apenas uma variável independente, o modelo assume a forma $y = \beta0 + \beta1x + \varepsilon$, onde β0 é o intercepto, β1 o coeficiente angular e ε o termo de erro. Apesar de sua simplicidade, a regressão linear é uma base conceitual importante para métodos mais avançados, como regressão regularizada e modelos não lineares [4]. O seu desempenho, no entanto, tende a ser limitado, especialmente porque o mercado imobiliário frequentemente apresenta relações não lineares. Por exemplo, o aumento da área de um imóvel nem sempre resulta em um aumento proporcional no preço — existem efeitos de escala e localização que tornam essa relação mais complexa. Neste trabalho, para melhorar seu desempenho, foi aplicada a técnica de seleção de variáveis via stepwise (RFECV), permitindo reduzir a presença de variáveis irrelevantes e melhorar a robustez do modelo.

### *2.2 Regressão Ridge*

A Regressão Ridge é uma técnica de regressão linear regularizada que adiciona um termo de penalização à soma dos quadrados dos coeficientes, com o objetivo de reduzir a variância do modelo e mitigar problemas de multicolinearidade [5]. A penalização é dada pela soma dos quadrados dos coeficientes multiplicada por um parâmetro λ. Ao aumentar o valor de λ, os coeficientes tendem a se aproximar de zero, mas sem nunca se tornarem exatamente nulos, preservando todas as variáveis no modelo. Essa característica torna a Regressão Ridge particularmente útil quando há muitas variáveis correlacionadas [6]. Ao desenvolver o trabalho voltado para o mercado imobiliário, ficou perceptível que algumas características dos imóveis estão altamente correlacionadas. Por exemplo, imóveis maiores tendem a ter mais quartos, mais banheiros e, muitas vezes, mais vagas de garagem. Isso gera o problema de multicolinearidade, que prejudica a estabilidade dos coeficientes na regressão linear. A decisão pela regressão Ridge veio justamente por abordar esse problema por meio de uma penalização que reduz o valor dos coeficientes, tornando o modelo mais estável e menos sensível às interdependências entre as variáveis, isso significa que o modelo consegue distribuir de forma mais equilibrada o peso das variáveis correlacionadas, melhorando a generalização e evitando super ajustes aos dados de treinamento.

### *2.3 Regressão Lasso*

No caso do uso da Regressão Lasso para prever valores de imóveis de Curitiba, onde os preços dos imóveis podem ser influenciados por uma variedade de atributos — desde características físicas até informações sobre localização e serviços no entorno —, o Lasso auxilia na abordagem de forma mais enxuta e interpretável. A Regressão Lasso é uma variação da regressão linear que, assim como a Regressão Ridge, aplica regularização aos coeficientes, mas utilizando a penalização da soma dos valores absolutos desses coeficientes [7]. Essa penalização incentiva que alguns coeficientes sejam exatamente zero, resultando em um modelo esparso que realiza seleção automática de variáveis.

### *2.4 Random Forest*

O Random Forest é um modelo de aprendizado de máquina baseado em uma coleção de árvores de decisão, onde cada árvore é construída a partir de uma amostra aleatória dos dados e um subconjunto aleatório das variáveis. Essa abordagem visa aumentar a precisão e reduzir o sobreajuste comum em modelos de árvore única. Segundo Breiman (2001) [8], criador do método, o Random Forest melhora a generalização ao agregar os resultados de diversas árvores não correlacionadas por meio da votação majoritária (no caso de classificação) ou da média (para regressão). O modelo se destaca no problema de previsão de preços de imóveis por ser um modelo capaz de capturar relações não lineares e interações complexas entre variáveis, o que é bastante comum no mercado imobiliário. Características como metragem ou número de suítes podem impactar o preço de forma diferente dependendo da localização, o Random Forest é um modelo que lida bem com esse tipo de complexidade, além de ser robusto contra outliers. Por exemplo, imóveis extremamente caros e fora do padrão não afetam tanto o desempenho do modelo quanto afetariam uma regressão linear. A sua desvantagem é a menor interpretabilidade, sendo mais difícil entender

exatamente como cada variável contribui para a previsão do preço.

### *2.5 Gradient Boosting*

O Gradient Boosting é um método de ensemble baseado em árvores de decisão que constrói o modelo de forma sequencial, onde cada nova árvore é treinada para corrigir os erros residuais do conjunto anterior. A técnica utiliza gradiente descendente para minimizar uma função de perda, ajustando os modelos de forma iterativa para melhorar a precisão. Diferentemente do Random Forest, que constrói árvores de forma independente e combina seus resultados, o Gradient Boosting adiciona cada nova árvore levando em conta os erros das anteriores, o que o torna altamente eficaz, mas também mais suscetível ao sobreajuste se não houver regularização adequada [9]. Isso resulta em um modelo extremamente poderoso para dados estruturados, como os dados do mercado imobiliário. É um modelo altamente eficaz para capturar efeitos não lineares e interações complexas e, assim como o Random Forest, é menos interpretável.

### *2.6 Rede Neural Artificial (MLP)*

As Redes Neurais Artificiais são modelos computacionais inspirados na estrutura e funcionamento do cérebro humano, compostos por unidades de processamento chamadas neurônios artificiais, organizados em camadas interconectadas. Cada conexão possui um peso que é ajustado durante o treinamento, permitindo que a rede aprenda padrões complexos a partir dos dados [10]. O processo de aprendizado das Redes Neurais Artificiais é geralmente realizado por meio de algoritmos de retropropagação do erro (backpropagation), que ajustam os pesos com base no gradiente descendente, buscando minimizar uma função de perda [11]. Essas redes são capazes de modelar relações não lineares e são amplamente utilizadas em tarefas como classificação, regressão e reconhecimento de padrões. No contexto imobiliário, uma rede neural pode aprender, por exemplo, que o impacto da metragem depende não só do tamanho absoluto, mas também da configuração interna do imóvel, da localização etc. A sua principal desvantagem é a baixa interpretabilidade, tornando difícil entender claramente como cada variável impacta o preço previsto.

## 3 Preparação dos dados

### *3.1 Conjunto de Dados*

Para o trabalho utilizamos como base um conjunto de dados públicos disponibilizado na plataforma Kaggle. O dataset utilizado foi publicado por Fernando Wittmann em 2022 e contém informações detalhadas sobre imóveis à venda na cidade de Curitiba (PR). As 18 variáveis contemplam diversas características dos apartamentos, como área privativa, número de quartos, banheiros, vagas de garagem, presença de atributos como elevador, piscina, salão de festas, além da localização por bairro, além disso, possui 18.760 registros. O dataset está disponível para acesso público na plataforma Kaggle [12]. As colunas são descritas na tabela abaixo:

| Coluna | Tipo | Descrição |
|---|---|---|
| usableAreas | int64 | Área útil |
| totalAreas | float64 | Área total |
| suites | float64 | Quantidade de suítes |
| bathrooms | int64 | Quantidade de banheiros |
| bedrooms | int64 | Quantidade de quartos |
| parkingSpaces | float64 | Vagas de estacionamento |
| amenities | object | Comodidades |
| description | object | Descrição do anúncio |
| title | object | Título do anúncio |
| zipCode | int64 | Cep |
| lon | float64 | Longitude |
| lat | float64 | Latitude |
| street | object | Rua |
| neighborhood | object | Bairro |
| poisList | object | Pontos de interesse próximos |
| yearlyIptu | float64 | IPTU anual |
| monthlyCondoFee | float64 | Aluguel |
| price | int64 | Preço |

**Tabela 1:** Descrição das variáveis do dataset.

### *3.2 Transformação da variável dependente*

A variável alvo do estudo 'price' apresentou inicialmente uma distribuição assimétrica à direita, o que é comum em dados relacionados a preços, pois há uma concentração de valores nas faixas inferiores e uma cauda longa com imóveis de alto valor. Esse tipo de assimetria pode comprometer a performance de modelos de regressão que assumem normalidade dos resíduos, como os modelos lineares [13].

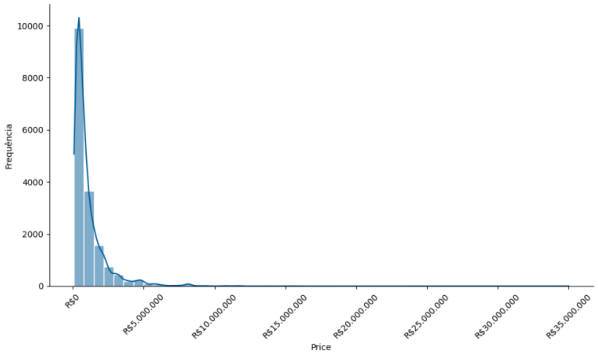


**Figura 1:** Distribuição original da variável dependente

Para mitigar a influência de valores extremos e aproximar a distribuição de uma normal, foi aplicada uma transformação logarítmica de base 10 à variável 'price'. O resultado dessa transformação é uma nova variável, que chamaremos de 'log_price', que apresenta uma distribuição mais simétrica e com formato próximo ao da distribuição normal.

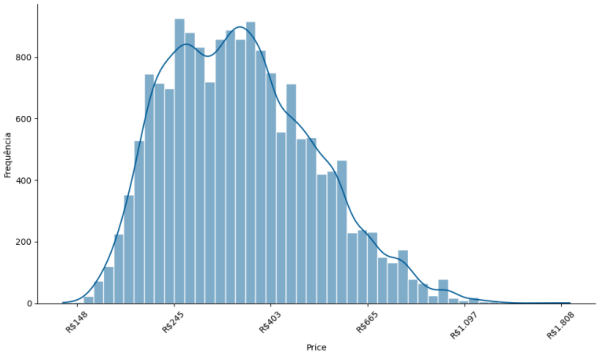


**Figura 2:** Distribuição ajustada da variável dependente

### *3.3 Criação de variáveis dummies*

Uma das variáveis mais ricas em informação do conjunto de dados era **'amenities'**, que listava as comodidades presentes em cada imóvel — como elevador, piscina, academia, entre outras. Como essa variável estava em formato de lista, foi necessário transformá-la em um formato que os modelos conseguissem interpretar. Para isso, aplicamos a técnica conhecida como one-hot encoding, que cria uma nova coluna para cada tipo de comodidade. Cada uma dessas colunas indica com 1 ou 0 se o imóvel possui ou não aquela característica. Após essa transformação, o número total de colunas geradas chegou a 173. No entanto, percebemos que várias dessas novas colunas representavam comodidades muito raras, presentes em menos de 1% dos imóveis. Manter essas colunas no modelo poderia trazer mais ruído do que informação. Por isso, fizemos uma análise de frequência e removemos as colunas com presença inferior a 1%. Essa limpeza ajudou, posteriormente, a reduzir a complexidade dos modelos, diminuir o risco de sobreajuste e tornar a base de dados mais enxuta e eficiente.

Outra variável de grande relevância foi a 'poisList', que continha, em formato de lista, os pontos de interesse mais próximos de cada imóvel. Entre os pontos de interesse estavam itens como pontos de ônibus, terminais de transporte, supermercados, escolas, restaurantes, farmácias, academias, padarias, entre outros. Esses elementos representam a infraestrutura e os serviços disponíveis no entorno do imóvel, que têm impacto direto na sua atratividade e, consequentemente, no seu valor de mercado. Para tornar essa informação utilizável pelos modelos preditivos, a **'poisList'** foi processada de forma a identificar a presença de determinados pontos de interesses considerados mais relevantes do ponto de vista da precificação. A partir disso, foram criadas nove variáveis dummies, cada uma representando um tipo de ponto de interesse. Cada coluna indica, com valor 1 ou 0, se aquele tipo de ponto está presente no entorno do imóvel. Essas novas features auxiliam a considerar não apenas onde o imóvel está e suas características físicas, mas também o que há ao seu redor.

### *3.4 Incorporação de dados externos*

Como forma de enriquecer a base de dados e considerar características geográficas mais amplas, foi realizada a incorporação de dados externos por meio do agrupamento da variável **'neighborhood'** (bairro) em 10 macrorregiões administrativas da cidade de Curitiba. Essas macrorregiões são delimitadas pela prefeitura municipal como uma forma de organização territorial e de planejamento urbano, e a associação entre cada bairro e sua respectiva região foi obtida diretamente do site oficial da Prefeitura de Curitiba. Essa decisão teve como objetivo reduzir a granularidade da variável de localização, possibilitando uma análise mais robusta sobre o comportamento dos preços por grandes zonas da cidade — o que seria difícil de capturar utilizando apenas os bairros de forma isolada, devido à grande quantidade de categorias distintas. Após essa correspondência, foram criadas 10 novas variáveis do tipo dummy, correspondentes a cada macrorregião. Essa abordagem permite que os modelos considerem a localização como um fator categórico generalizado, facilitando a identificação de padrões espaciais amplos, como zonas de valorização ou desvalorização imobiliária. Além disso, contribui para evitar o excesso de colunas com baixa representatividade, comum quando se utiliza bairros individuais.

### *3.5 Tratamento de erros e duplicidade*

Uma etapa essencial no processo de preparação dos dados foi a identificação de registros duplicados e de

valores inconsistentes. Como o conjunto de dados continha milhares de registros provenientes de anúncios online, era comum encontrar imóveis repetidos, divulgados mais de uma vez com pequenas variações — por exemplo, diferença no título ou descrição — mas com as mesmas características estruturais. Para lidar com esse problema, foi criada uma chave única de identificação para cada imóvel, por meio da concatenação das colunas 'usableAreas','totalAreas','suites','bathrooms','bedrooms','lon','lat' e 'price'. Essa combinação foi pensada de forma a capturar a identidade física e geográfica do imóvel, além do seu valor anunciado, o que permitiu detectar e remover registros duplicados com maior precisão. Além disso, foi realizado um tratamento específico para lidar com valores atípicos, especialmente no que diz respeito à variável 'price'. Durante a análise exploratória, foram identificados imóveis com preços extremamente baixos ou excessivamente altos, que destoavam completamente do restante da base. Muitos desses casos eram claramente erros de digitação — por exemplo, imóveis de alto padrão com preço de apenas alguns reais — ou confusões entre valores de venda e aluguel, algo comum em bases coletadas por scraping ou APIs.Imóveis com valores fora de faixas realistas foram, portanto, removidos ou corrigidos. Esse cuidado foi importante para evitar distorções nas métricas e nos modelos, já que valores extremos podem afetar desproporcionalmente o ajuste e a acurácia das previsões. Ao final das correções de erros e valores duplicados e criação das variáveis dummies, ficamos com um dataset contendo 17.086 registros e 104 colunas.

## 4 Seleção de variáveis

Para a seleção de variáveis, inicialmente foi removido do dataset as variáveis categóricas, mantendo apenas variáveis numéricas. Posteriormente, os dados foram divididos em treino (70%) e teste (30%), garantindo uma avaliação adequada dos modelos. Foram utilizadas duas abordagens principais e complementares para a seleção de variáveis, descritas a seguir.

### *4.1 Random Forest*

A seleção foi baseada na importância das variáveis calculadas pelo modelo. Inicialmente, foi plotado um gráfico de RMSE versus número de estimadores, a fim de identificar a quantidade ideal de árvores na floresta. Optou-se por utilizar 200 estimadores, uma vez que, a partir desse ponto, o RMSE apresentou estabilidade, indicando um bom equilíbrio entre performance e custo computacional. Após o ajuste, o modelo Random Forest forneceu os graus de importância de cada variável, permitindo a seleção das quatro variáveis com os maiores valores de importância, consideradas, portanto, as mais relevantes para o problema de previsão de preços.

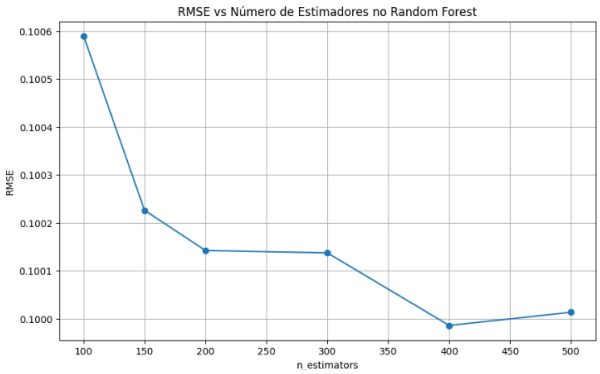


**Figura 3:** Gráfico de quantidade de estimadores pelos valores de RMSE

### *4.2 Regressão Lasso*

Na abordagem com Lasso, foi utilizado o método de validação cruzada para ajuste do hiperparâmetro alpha, que controla o grau de penalização aplicado aos coeficientes. O gráfico de erro de validação cruzada versus alpha indicou que o valor ideal seria 0.000658, ponto onde o erro se estabilizou em um patamar mais baixo. A regressão Lasso tem a capacidade de realizar seleção de variáveis, uma vez que força coeficientes menos relevantes a zero. Para este trabalho, optamos por manter as variáveis que apresentaram coeficiente absoluto superior a 0.01, totalizando 20 variáveis selecionadas. Este critério foi adotado para garantir que apenas variáveis com impacto relevante no modelo fossem mantidas.

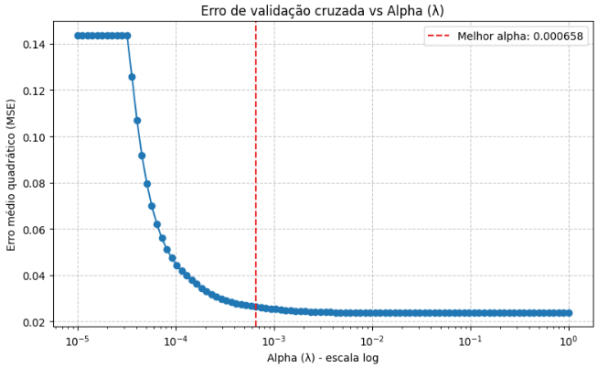


**Figura 4:** Gráfico de erro de validação cruzada pelo valor de Alpha

Após a aplicação dos dois métodos, foram cruzadas as variáveis selecionadas por ambos, priorizando aquelas em comum, por representarem um consenso entre os dois critérios distintos — um baseado em árvores de decisão (Random Forest) e outro em regularização linear (Lasso). Adicionalmente, foi necessária a

inclusão manual das variáveis relativas às regionais da cidade, uma vez que nem todos os modelos selecionaram todas elas. Essa etapa foi essencial para garantir que os modelos fossem capazes de realizar previsões para imóveis localizados em todas as regiões da cidade. Caso contrário, ao manter apenas uma parte das regionais selecionadas pelos métodos, os modelos se tornariam incapazes de prever preços para imóveis de regiões não representadas nas variáveis escolhidas. Esse processo de seleção buscou equilibrar dois objetivos principais: reduzir a dimensionalidade do dataset, mitigando riscos como a multicolinearidade e o overfitting, e, ao mesmo tempo, preservar as variáveis que capturam de maneira mais efetiva os fatores que impactam o valor dos imóveis na cidade de Curitiba.

## 5 Métodos e Ferramentas

### *5.1 Processo de modelagem*

O processo de modelagem preditiva consistiu em diversas etapas sequenciais, iniciando pelo pré-processamento dos dados, etapa fundamental para garantir a qualidade e integridade das informações utilizadas nos modelos e que foi melhor detalhada anteriormente. Após isso, a base de dados foi dividida em duas partes: conjunto de treino (70%) e conjunto de teste (30%). Essa separação permitiu que os modelos aprendessem os padrões presentes nos dados históricos e, posteriormente, tenham seu desempenho avaliado em dados não vistos durante o treinamento, garantindo uma avaliação mais realista da sua capacidade preditiva. Foram ajustados e avaliados seis modelos de regressão: Regressão Linear, Ridge Regression, Lasso Regression, Random Forest Regressor, Gradient Boosting e Rede Neural Artificial (RNA). A escolha desses modelos visa contemplar abordagens tanto lineares quanto não lineares, além de métodos tradicionais e modelos mais sofisticados baseados em árvores e redes neurais. Para avaliar o desempenho dos modelos, utilizamos as métricas estatísticas a seguir.

#### 5.1.1 *MAE (Mean Absolute Error)*

Segundo Géron (2019) [14], o Erro Médio Absoluto (MAE – Mean Absolute Error) é uma medida de erro que quantifica a média das diferenças absolutas entre os valores previstos e os valores reais. Por ser uma medida intuitiva, é bastante útil na avaliação da magnitude média dos erros de previsão, análise de séries temporais e regressão, independentemente da direção. Quanto menor o valor do MAE, melhor é a precisão do modelo.

#### 5.1.2 *RMSE (Root Mean Squared Error)*

O RMSE é a raiz quadrada da média dos erros ao quadrado. Ao elevar os erros ao quadrado, a métrica dá um peso maior a erros maiores, o que a torna sensível a outliers ou a previsões com grandes desvios. O fato de extrair a raiz quadrada no final faz com que o valor do RMSE esteja na mesma unidade da variável que está sendo prevista, o que facilita a interpretação do resultado, um valor de RMSE igual a zero indica um ajuste perfeito do modelo, enquanto valores mais altos indicam que o modelo está menos preciso [15].

#### 5.1.3 *$R^2$ (Coeficiente de Determinação)*

De acordo com Bussab (2017) [16], o coeficiente de determinação, $R^2$, é uma medida da porcentagem de variação total em Y que é explicada pela regressão de Y em X. Um $R^2$ igual a 0,85, por exemplo, significa que 85% da variabilidade da variável dependente é explicada pelas variáveis independentes do modelo. É importante ressaltar que um valor alto de $R^2$ não garante que o modelo seja o melhor. O $R^2$ tende a aumentar à medida que novas variáveis são adicionadas ao modelo, mesmo que elas não sejam estatisticamente significativas. Por isso, é comum utilizar o $R^2$ ajustado, que penaliza a inclusão de variáveis desnecessárias.

#### 5.1.4 *MAPE (Mean Absolute Percentage Error)*

O Erro Percentual Médio Absoluto (MAPE) expressa a precisão de um modelo como uma porcentagem. Essa métrica é popular por ser fácil de interpretar, pois o resultado é uma porcentagem que indica a magnitude média do erro de previsão em relação ao valor real. Por exemplo, um MAPE de 10% significa que, em média, a previsão se desvia 10% do valor real [17].

Usar essas diferentes métricas juntas ajuda a ter uma visão mais completa sobre o desempenho dos modelos. Cada uma delas analisa o erro por um ponto de vista diferente — seja em valores absolutos, em porcentagem ou na capacidade de explicar a variação dos preços. Isso é importante porque, em um problema como a previsão de preços de imóveis, nenhuma métrica sozinha conta toda a história. Combinando todas, conseguimos entender melhor como os modelos estão se saindo e tomar decisões mais seguras sobre qual deles funciona melhor.

### *5.2 Ferramentas*

#### 5.2.1 *Pandas e Numpy*

Usadas para manipulação, organização e análise dos dados. Permitiram ler o dataset, tratar valores ausentes, transformar colunas e realizar cálculos estatísticos de forma eficiente.

#### 5.2.2 *Scikit Learn*

A principal biblioteca para criação e avaliação de modelos de machine learning. Foi utilizada para aplicar os modelos de regressão linear, ridge, lasso, random forest e gradient boosting, além de fornecer ferramentas para validação, escalonamento e divisão da base de dados.

#### 5.2.3 *TensorFlow e Keras*

Usadas para a construção e treinamento da rede neural artificial. O Keras, integrado ao TensorFlow, facilitou a definição da arquitetura da rede, o ajuste de camadas ocultas, funções de ativação e o controle do processo de treinamento.

#### 5.2.4 *Matplotlib e Seaborn*

São bibliotecas de visualização que ajudaram na criação de gráficos e na análise exploratória dos dados, como a distribuição dos preços, correlação entre variáveis e comparação dos resultados dos modelos.

## 6 Resultados

Após testar os seis modelos de regressão para prever os preços dos imóveis, foi possível observar diferenças importantes no desempenho de cada um, como demonstra a seguinte tabela.

| Modelo | MAE | RMSE | $R^2$ | MAPE |
|---|---|---|---|---|
| Regressão Linear (Stepwise) | 0.11 | 0.14 | 0.86 | 1.87% |
| Ridge | 0.10 | 0.13 | 0.87 | 1.79% |
| Lasso | 0.10 | 0.13 | 0.87 | 1.78% |
| Random Forest | 0.07 | 0.11 | 0.91 | 1.26% |
| Gradient Boosting | 0.09 | 0.12 | 0.90 | 1.49% |
| Rede Neural | 0.09 | 0.13 | 0.89 | 1.60% |

**Tabela 2:** Resultados

Os modelos lineares, como Regressão Linear (com seleção stepwise), Ridge e Lasso, apresentaram resultados consistentes, com $R^2$ entre 0,86 e 0,87. Eles são úteis principalmente por sua facilidade de interpretação, o que permite entender melhor a influência de cada variável no preço dos imóveis. No entanto, por serem modelos lineares, têm limitações na hora de lidar com relações mais complexas entre as variáveis. Para melhorar seu desempenho, foram aplicadas estratégias adicionais de pré-processamento. Entre elas, a discretização de variáveis numéricas em faixas e o uso de one-hot encoding para as variáveis categóricas. Também foi feito um diagnóstico dos resíduos para avaliar onde os erros eram maiores, o que ajudou a ajustar os modelos e entender melhor seu comportamento.

Já o Random Forest e Gradient Boosting tiveram os melhores desempenhos gerais. O Random Forest se destacou como o melhor modelo, com MAE de 0,07, RMSE de 0,11 e $R^2$ de 0,91, enquanto o Gradient Boosting também obteve bons resultados ($R^2$ = 0,90). Esses modelos conseguem lidar melhor com interações e não linearidades presentes nos dados, o que explica sua performance superior. Durante os testes, foram avaliadas diferentes quantidades de estimadores (n_estimators) para ajustar o desempenho dos modelos. Isso foi essencial para equilibrar a complexidade e o tempo de execução, buscando sempre o melhor resultado nas métricas de erro.

A rede neural utilizada também teve um bom desempenho, com $R^2$ = 0,89, superando os modelos lineares. Para chegar nesse resultado, foi necessário testar diversas arquiteturas diferentes, alterando o número de camadas, neurônios, funções de ativação, taxa de dropout e outros hiperparâmetros. A arquitetura que apresentou o melhor equilíbrio entre desempenho e estabilidade foi a seguinte: Arquitetura: [32, 12] (duas camadas ocultas com 32 e 12 neurônios, respectivamente), função de ativação: elu, dropout rate: 0.2 (para evitar overfitting). Embora a rede neural não tenha superado os modelos de árvores, seu resultado foi bastante competitivo, ao mesmo tempo exigindo mais tempo de ajuste e apresentando maior sensibilidade aos hiperparâmetros, o que demandou mais cuidado na construção do modelo.

## 7 Conclusão

Este trabalho teve como objetivo comparar diferentes modelos de regressão na tarefa de prever os preços de imóveis residenciais na cidade de Curitiba. Os resultados obtidos mostraram que os modelos baseados em árvores, especialmente o Random Forest e o Gradient Boosting, foram os que apresentaram melhor desempenho geral, superando os modelos lineares em todas as métricas avaliadas (MAE, RMSE, $R^2$ e MAPE). Esses modelos mostraram maior capacidade de capturar relações complexas entre as variáveis, como não linearidades e interações, sendo, portanto, mais eficazes no cenário analisado.

A aplicação da transformação logarítmica na variável preço foi um passo importante para melhorar a qualidade das previsões. Com essa transformação, os modelos apresentaram métricas mais equilibradas e previsões mais consistentes, reduzindo a influência de valores extremos e estabilizando a variância dos dados.

Os modelos lineares, como Regressão Linear, Ridge e Lasso, embora apresentem menor desempenho preditivo, mostraram-se úteis por sua interpretabilidade e simplicidade. Foram aplicadas estratégias como discretização de variáveis numéricas e codificação de variáveis categóricas (one-hot encoding) para tentar melhorar sua performance, além da análise dos resíduos para entender melhor os padrões de erro. Mesmo com esses ajustes, esses modelos não conseguiram competir com os modelos mais complexos em termos de precisão. A rede neural, por sua vez, mostrou-se promissora, com desempenho próximo ao dos modelos de árvores. Apesar disso, o modelo exigiu mais tempo de treinamento e sensibilidade no ajuste de hiperparâmetros, o que aumenta a complexidade do seu uso em comparação aos modelos baseados em árvore.

Como possibilidades de melhorias para trabalhos futuros, destacam-se:

- Explorar modelos de redes neurais mais profundas (Deep Learning), que podem captar padrões ainda mais complexos nos dados;
- Avaliar outros tipos de pré-processamento e melhorar o ajuste dos modelos lineares, buscando aumentar sua competitividade com os modelos não lineares;
- Incorporar variáveis econômicas e indicadores do mercado imobiliário;
- Testar abordagens de aprendizado em tempo real (online learning) ou modelos baseados em séries temporais, que considerem a evolução do mercado imobiliário ao longo do tempo.

Dessa forma, este estudo contribui para mostrar que, com um bom tratamento dos dados e escolha adequada do modelo, é possível obter previsões precisas e confiáveis dos preços de imóveis, o que pode apoiar decisões em diversos contextos, como financiamentos, avaliações e investimentos no setor imobiliário.

## 8 Agradecimentos

Agradeço à minha orientadora Deisy Morselli Gysi, pelo suporte, orientações e paciência durante o desenvolvimento deste trabalho. Agradeço também aos professores do curso de Especialização em Data Science e Big Data da UFPR, aos colegas de turma, amigos e familiares que me apoiaram durante toda essa jornada.

## 9 Referências

[1] FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Análise de dados: modelagem multivariada para tomada de decisões**. 2. ed. Rio de Janeiro: Elsevier, 2019.

[2] GHOSH, Sanjeet Kumar; DEY, Lopamudra. Predictive modeling of real estate property value using machine learning. *Materials Today: Proceedings*, v. 46, p. 10428-10431, 2021.

[3] MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. 5. ed. Hoboken: Wiley, 2012.

[4] JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: with applications in R**. 2. ed. New York: Springer, 2021.

[5] HOERL, A. E.; KENNARD, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, v. 12, n. 1, p. 55-67, 1970.

[6] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. New York: Springer, 2009.

[7] TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 58, n. 1, p. 267-288, 1996.

[8] BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001.

[9] FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, v. 29, n. 5, p. 1189-1232, 2001.

[10] HAYKIN, S. **Neural networks and learning machines**. 3. ed. Upper Saddle River: Pearson, 2009.

[11] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

[12] WITTMANN, F. Curitiba apartment prices. *Kaggle*, 2023. Disponível em: https://www.kaggle.com/datasets/wittmannf/curitiba-apartment-prices/data. Acesso em: 9 ago. 2025.

[13] GUJARATI, Damodar N.; PORTER, Dawn C. **Econometria básica**. 5. ed. Porto Alegre: AMGH, 2011.

[14] GÉRON, Aurélien. Estudo das técnicas de previsão da demanda aplicado em uma fabricante de medicamentos. Revista FT, 2019. Acesso em: 08/08/2025.

[15] WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, v. 30, n. 1, p. 79-82, 2005.

[16] BUSSAB, W. O. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.

[17] HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, v. 22, n. 4, p. 679-688, 2006.