https://doi.org/10.47456/bjpe.v11i1.47469 ARTIGO ORIGINAL | OPEN ACCESS

NÚCLEO DE PESQUISA EM GESTÃO DE SISTEMAS DE PRODUÇÃO - NP GSP
Brazilian Journal of Production Engineering
BJPE - Revista Brasileira de Engenharia de Produção
Disponível em: https://periodicos.ufes.br/bjpe
ISSN: 2447-5580
Campus São Mateus
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Pró-Reitoria de Pesquisa e Pós-Graduação
Universidade Federal do Espírito Santo

PREÇOS DE HABITAÇÃO NA CALIFÓRNIA: UMA ABORDAGEM PARA PREVISÕES NO SETOR IMOBILIÁRIO

HOUSING PRICES IN CALIFORNIA: AN APPROACH TO FORECASTING IN THE REAL ESTATE SECTOR

PRECIOS DE VIVIENDA EN CALIFORNIA: UN ENFOQUE PARA PRONÓSTICOS EN EL SECTOR INMOBILIARIO

Christian Gianelli da Silva [1]

[1] Universidade Federal de Ouro Preto (UFOP)
[1] christiangianelli63@gmail.com

ARTIGO INFO.

Recebido: 27.01.2025
Aprovado: 18.02.2025
Disponibilizado: 17.03.2025

PALAVRAS-CHAVE: Machine Learning; Predição de Preços, Setor Imobiliário; Random Forest; Regressão Linear.
KEYWORDS: Machine Learning; Price Prediction; Real Estate Sector; Random Forest; Linear Regression.
PALABRAS CLAVE: *Aprendizaje Automático; Predicción de Precios; Sector Inmobiliario; Bosque Aleatorio; Regresión Lineal.*

***Autor Correspondente**: **Silva, C. G. da.**

RESUMO

Este artigo apresenta uma análise comparativa de técnicas de aprendizado de máquina aplicadas à previsão de preços no setor imobiliário da Califórnia. Foram investigados os modelos de Regressão Linear Múltipla, Regressão Polinomial, Regressão Robusta (RANSAC) e Floresta Aleatória (Random Forest), sendo cada um avaliado com base em métricas estatísticas como Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2). O conjunto de dados utilizado foi obtido do repositório StatLib e contém informações sobre características dos imóveis, localização e perfil socioeconômico da população. Os resultados indicam que, apesar da Floresta Aleatória apresentar melhor desempenho preditivo, há indícios de overfitting, sugerindo que um aumento no número de amostras poderia melhorar a generalização do modelo. Por outro lado, os modelos de Regressão Linear e Regressão Polinomial demonstraram maior estabilidade e capacidade de generalização, ainda que com leve perda de precisão. Este estudo contribui para a compreensão da aplicabilidade dessas técnicas na modelagem de preços imobiliários e discute os impactos do tamanho da amostra na acurácia dos modelos.

ABSTRACT

This paper presents a comparative analysis of machine learning techniques applied to price forecasting in the California real estate sector. The models investigated were Multiple Linear Regression, Polynomial Regression, Robust Regression (RANSAC) and Random Forest, each one being evaluated based on statistical metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination (R^2). The dataset used was obtained from the StatLib repository and contains information on property characteristics, location and socioeconomic profile of the population. The results indicate that, although Random Forest presents better predictive performance, there are signs of overfitting, suggesting that an increase in the number of samples could improve the generalization of the model. On the other hand, the Linear Regression and Polynomial Regression models demonstrated greater stability and generalization capacity, although with a slight loss of accuracy. This study contributes to the understanding of the applicability of these techniques in real estate price modeling and discusses the impacts of sample size on model accuracy.

RESUMEN

Este artículo presenta un análisis comparativo de técnicas de aprendizaje automático aplicadas a la previsión de precios en el sector inmobiliario de California. Se investigaron los modelos de regresión lineal múltiple, regresión polinomial, regresión robusta (RANSAC) y Random Forest, cada uno de los cuales se evaluó en función de métricas estadísticas como el error absoluto medio (MAE), el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). El conjunto de datos utilizado se obtuvo del repositorio StatLib y contiene información sobre las características de la propiedad, la ubicación y el perfil socioeconómico de la población. Los resultados indican que, aunque Random Forest presenta un mejor desempeño predictivo, existen signos de sobreajuste, lo que sugiere que un aumento en el número de muestras podría mejorar la generalización del modelo. Por otro lado, los modelos de Regresión Lineal y Regresión Polinomial demostraron mayor estabilidad y capacidad de generalización, aunque con una ligera pérdida de precisión. Este estudio contribuye a la comprensión de la aplicabilidad de estas técnicas en el modelado de precios inmobiliarios y analiza los impactos del tamaño de la muestra en la precisión del modelo.

INTRODUÇÃO

A busca por ferramentas de precificação de imóveis mais ágeis e precisas, ganhou um grande destaque após a crise econômica de 2008, ano no qual em um adendo ao acordo de Basiléia II (que regulamenta o funcionamento de bancos e instituições financeiras) foram reafirmadas em suas diretrizes, indicando para que essas instituições monitorassem anualmente os valores dos imóveis de garantia de hipotecas (Hong et al., 2020). Isso consequentemente, aumentou a frequência, os custos e as responsabilidades do processo de precificação.

Os modelos de precificação tradicionais, também chamados de hedônicos, possuem dificuldade de equiparar em relação ao aumento dessa demanda e crescimento exponencial na velocidade com que novos dados são disponibilizados. Diante disso, houve uma maximização na utilização e desenvolvimento de modelos de Aprendizado de Máquina (AM) capazes de capturar, processar e analisar informações de alta complexidade voltadas para o setor (Ho et al., 2021). A definição de AM se baseia na ciência (e a arte) da programação de computadores para que eles possam aprender com diversos tipos de dados. Em outras palavras, diz-se que um programa de computador aprende pela experiência e em relação a algum tipo de tarefa T e alguma medida de desempenho P se o seu desempenho em T, conforme medido por P, melhora com a experiência (Mitchell, 1977). Ainda, pode-se considerar AM como um subconjunto da Inteligência Artificial (IA) que visa treinar máquinas através da entrada de informações (dados, imagens, valores numéricos etc.), podendo ser categorizado em Aprendizado Supervisionado, Aprendizado não Supervisionado e Semi-supervisionado (Ho et al., 2021).

Tratando-se desse contexto, é possível destacar o tipo de aplicação para previsão de preços do setor imobiliário utilizando Aprendizado Supervisionado, onde os dados de treinamento são disponibilizados ao algoritmo para alcançar as previsões desejadas (rótulos). Para prever uma tarefa típica para um alvo de valor numérico (ex: preço de um imóvel em determinado bairro) utiliza-se uma tarefa de regressão (Geron, 2019). Portanto, este trabalho visa utilizar os seguintes algoritmos (Regressão Linear Aleatória).

As escolhas dos algoritmos foram fundadas a partir dos resultados de previsão mais assertivos encontrados na literatura para previsão de valores de imóveis no estado da Califórnia (Geron, 2019). O conjunto de dados foi retirado do repositório chamado StatLib e possui informações referente a preços do setor imobiliário. Por fim, também possui objetivo de disponibilizar o desempenho dos métodos aplicados entre ambos, argumentando sua *performance*, considerando algumas características multivariadas.

METODOLOGIA

A abordagem tradicional de avaliação de bens origina-se da hipótese de Lancaster de 1966, a qual resulta da atribuição de um valor determinado de acordo com suas características, demonstrando uma relação entre o valor desse bem e suas características peculiares (Ferreira, 2010). Apesar da clara necessidade de garantir a acuidade das previsões, a determinação desse índice continua sendo um problema. Essa avaliação do padrão se baseia na agregação das vendas de forma temporal e geográfica, o valor médio dessas negociações serve de

indicador para os demais imóveis. Com isso, influenciado por oscilações de mercado onde imóveis que são vendidos abaixo ou acima do valor esperado acabam influenciando o método de avaliação que precisa ser corrigido considerando ações do tempo (Barr et al., 2015). Diante dessas dificuldades, tem-se cada vez mais buscado metodologias alternativas que visem garantir uma maior rapidez e eficiência na previsão desses valores. Dentre esses novos métodos, o modelo de AM vem ganhando cada vez mais espaço devido sua rapidez e robustez em lidar com uma expressiva quantidade de dados (Park, 2015) realiza uma comparação entre o modelo tradicional (Hedônico) e técnicas de AM. Seu trabalho identificou que o uso dos algoritmos de AM aumentou a previsibilidade dos preços e melhorou sua assertividade, demonstrando-se como uma potencial alternativa em relação aos modelos tradicionais devido à redução no tempo de execução que permite tomar decisões de forma mais rápida e eficiente. Compreendendo o que se deseja fazer com os dados, é preciso determinar os requisitos adicionais para solução do problema aplicando um modelo alinhado ao melhor algoritmo. Se tratando das implementações dos modelos propostos, incluindo o algoritmo de Random Forest, foi utilizada a linguagem de programação Python, conhecida pela eficiência no processamento de dados e ampla gama de bibliotecas voltadas à modelagem preditiva (Geron, 2019). Foi adatada a biblioteca Scikit-Learn para a implementação do algoritmo, devido à sua *interface* intuitiva e otimizações computacionais para modelos de Aprendizado Supervisionado (Supervised Learning). Nesse prisma, o desenvolvimento e a aplicação do modelo envolveram as seguintes bibliotecas:

- NumPy: Utilizada para manipulação eficiente de arrays multidimensionais e operações matemáticas;
- Pandas: Empregada para leitura, estruturação e pré-processamento do conjunto de dados;
- Matplotlib e Seaborn: Aplicadas para a construção de visualizações gráficas que auxiliam na análise exploratória dos dados e na avaliação do desempenho do modelo;
- Scikit-Learn: Responsável pela implementação dos algoritmos de aprendizado de máquina, incluindo a classe Random Forest Regressor para modelagem preditiva.

Após essa análise, foi possível definir os algoritmos que melhor se encaixam na resolução desse problema, conforme a seguir; Regressão Linear: a regressão linear múltipla também é uma técnica multivariada cuja finalidade principal é obter uma relação matemática entre uma das variáveis (a variável dependente) e o restante das variáveis que descrevem o sistema (variáveis independentes). Sua principal aplicação, após relação matemática é produzir valores para a variável dependente quando se têm as variáveis independentes. Ou seja, ela pode ser usada na predição de resultados (Moita Neto, 2004). Optou-se pelo uso também da técnica de Regressão Robusta por ser considerada uma técnica robusta não somente com respeito aos *outliers*, e, sim, porque quanto maior o número de variáveis de um modelo, mais difícil se torna a identificação de *outliers* com o uso das técnicas de regressão clássicas (S-PLUS 2017).

Outra técnica utilizada neste trabalho é a Regressão Polinomial para problemas de relações não lineares a partir das variáveis existentes, são geradas novas variáveis polinomiais e com elas terá sua capacidade elevada proporcionalmente ao grau do polinômio criado (S-PLUS 4, 1998). É importante ressaltar que o algoritmo de regressão linear não muda.

Floresta Aleatória: O modelo de Floresta Aleatória (RF do inglês Random Forest) desenvolve uma combinação de árvores de decisão através da seleção aleatória das variáveis de composição, essas variáveis são estruturadas na forma de nós de decisão (nós internos) até chegar aos nós folha (objetivo), o resultado do método é a consolidação dos dados oriundos das precisões de todas as árvores geradas (Breiman, 2001). Esse modelo pode ser utilizado tanto em tarefas de classificação como de Regressão. O Conjunto de dados trabalhado neste artigo é oriundo de um repositório chamado StatLib e possui informações referente a preços do setor imobiliário levantadas no ano de 1990 no estado da Califórnia, é composto por 20.640 objetos, com 10 atributos, que descrevem as propriedades quanto a suas características, habitantes e localização. Possui também 207 valores ausentes, todos da variável Total de Quartos. A estruturação dos dados e conhecimento de suas características é uma etapa importante para aplicação das técnicas de AM. Para este conjunto de dados, temos:

- Localização: latitude, longitude e proximidade com oceano, sendo os dois primeiros atributos quantitativos intervalares e o último qualitativo nominal;
- Caracterização do Imóvel: total cômodos, total quartos e idade média casas, onde os dois primeiros são atributos quantitativos discretos racionais, o último quantitativo contínuo racional;
- Caraterização dos Habitantes: população, famílias e renda média, onde os dois primeiros são atributos quantitativos discretos racionais e o último quantitativo contínuo racional;
- Atributo Alvo: valor médio das casas (atributo quantitativo contínuo racional).

O conhecimento de técnicas voltadas para implementação de modelos de AM gera maior confiabilidade e *performance* no desenvolvimento do projeto de acordo com suas previsões almejadas. Dentre algumas técnicas de AM foi definido para aplicação deste trabalho, a técnica de Aprendizado Supervisionado. Utiliza-se essa técnica quando necessita de um aprendizado a partir de resultados pré-definidos (ex.: preços de imóveis), utilizando os valores passados da variável (Alvo) para aprender quais devem ser seus resultados de saída. Para a previsão dos preços dos imóveis, foram selecionados e comparados quatro modelos estatísticos e de aprendizado de máquina:

- Regressão Linear Múltipla (MLR - Multiple Linear Regression): Escolhida por sua interpretação simples e por ser amplamente utilizada em estudos de precificação imobiliária (Hastie, Tibshirani & Friedman, 2009);
- Regressão Polinomial (PR-Polynomial Regression): Utilizada para identificar relações não lineares entre as variáveis, aumentando a capacidade do modelo em representar padrões mais complexos (James et al., 2013);
- Regressão Robusta (RANSAC - Random Sample Consensus): Minimizadora da influência de *outliers* nos coeficientes da regressão linear, melhorando a robustez do modelo (FISCHLER; Bolles, 1981);

- Floresta Aleatória (RF - Random Forest): Algoritmo de aprendizado de máquina baseado em árvores de decisão, capaz de capturar interações complexas entre variáveis e reduzir o risco de *overfitting* por meio de *ensemble learning* (Breiman, 2001).

Cada modelo foi avaliado utilizando técnicas estatísticas e métricas de desempenho, conforme descrito nas próximas seções. Esses mesmos valores servem como supervisão dessas previsões, permitindo o ajuste nas previsões com base nos erros. Outra técnica de importante aplicação no processo de geração e implementação de um modelo de AM, é a seleção da medida de desempenho. Importante tarefa para previsão de erros gerados pelo sistema em suas previsões. Diante disso, foram escolhidas três métricas típicas de desempenho para modelos de regressão linear e floresta aleatória. A equação da Raiz do Erro Quadrático Médio (RMSE) proporciona uma ideia da quantidade de erros gerados pelo sistema em suas previsões, com um peso maior para grandes erros (Geron, 2019) (Equação 1), Raiz do Erro Quadrático Médio (RMSE):

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2} \quad (1)$$

Então,

m é o número de instâncias no conjunto de dados;

$x^{(i)}$ é um vetor de todos os valores da característica da i − ésima instância do conjunto de dados, e $y^{(i)}$ é seu valor desejado de saída para aquela instância;

h é a função de previsão do sistema (hipótese);

RMSE (X, h) é hipótese h, a função de custo medida no conjunto de exemplos utilizado.

O valor do coeficiente de determinação é obtido na parte superior da fração onde calcula-se a soma residual dos erros quadrados, representando o valor real da amostra e o valor previsto pelo modelo ajustado. No Coeficiente de Determinação R^2 também conhecido como R − quadrado, tem-se a soma dos quadrados da diferença entre as amostras e o valor médio das amostras (Equação 2), R-quadrado (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (2)$$

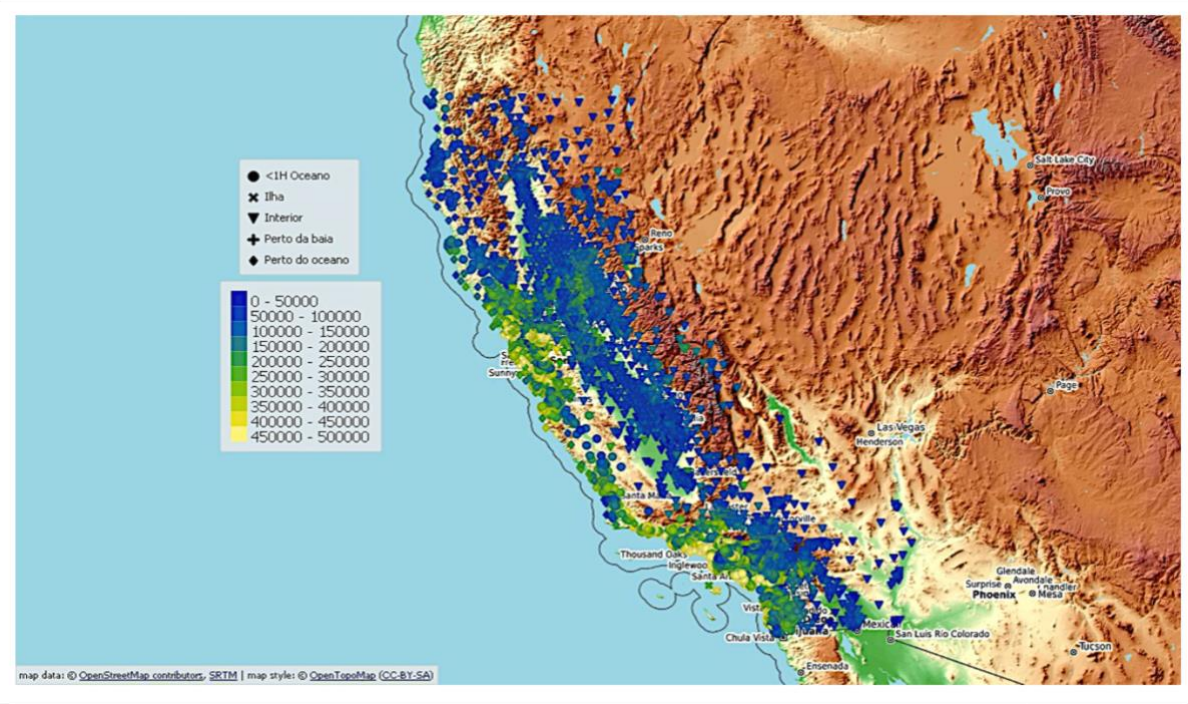
A terceira métrica escolhida foi a equação do Erro Médio Absoluto onde realiza a medida da distância entre dois vetores, o vetor das previsões e o vetor do alvo, Equação (3), Erro Médio absoluto (MAE):

$$MAE(X,h) = \frac{1}{m}\sum_{i=1}^{m}\left|h\left(x^{(i)}\right) - y^{(i)}\right| \quad (3)$$

Em que;

Quanto maior o índice da norma, mais ela se concentra em valores maiores e negligência os pequenos. Por este motivo é possível afirmar que a RMSE é mais sensível a *outliers*. Após visualização dos dados foi implementado de forma geográfica a distribuição dos preços dos imóveis em relação à localização. É possível identificar que os preços do setor imobiliário mais altos estão relacionados com a localização Próximo do Oceano e a densidade populacional (Figura 1).

Figura 1. Preços dos imóveis de acordo com sua localização



Fonte: Autor (2025).

Após algumas análises foi possível identificar na exploração dos dados que, todos os atributos são numéricos, exceto Proximidade com Oceano seu tipo é objeto. Então, supõe-se que, poderia conter qualquer tipo de objeto. Para esse banco de dados foi considerado como um objeto de texto. Analisando as primeiras instâncias foi possível constatar que os valores atribuídos para essa coluna são categóricos por serem repetitivos. Observando o resumo de cada atributo transparece que, os valores nulos são ignorados e 25% dos bairros possui Idade Média das Casas menor que 18, enquanto 50% são menores que 29, e 75% são menores que 37. Em outra denotação 25º percentil (ou 1º quartil), média (2º quartil) e o 75º percentil (ou 3º quartil).

A detecção da presença de *outliers* (dados que se diferenciam drasticamente de todos os outros) considerando os valores de quartil (Linf: Q1-1,5*QR; LSup: Q3+1,5*QR), mostra que todas as variáveis apresentam valores acima do limite máximo, com exceção dos atributos longitude e latitude (que não são calculadas) e Idade Média das casas (na qual os dados estão dentro dos limites). Através dessas informações e por ter um conjunto de dados mediano foi possível calcular o coeficiente de correlação padrão (ou r de Person). Esse coeficiente varia de (−1 a 1).

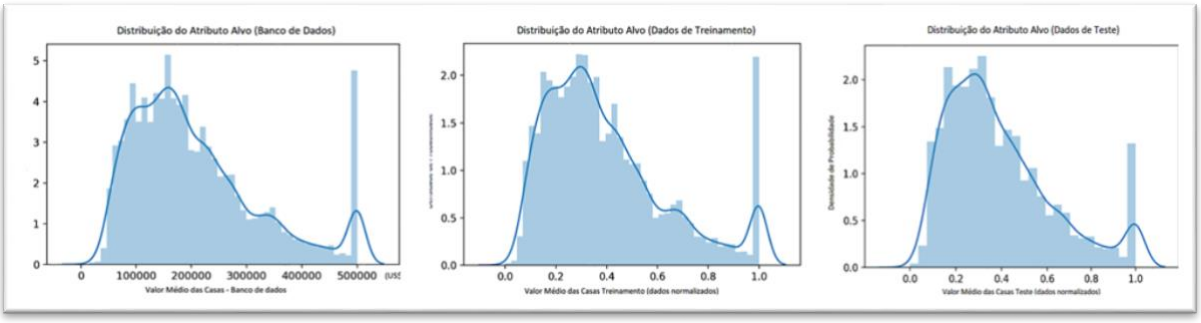
Quando apresenta um comportamento próximo de 1, entende-se que existe uma forte correlação positiva; por exemplo, Valor Médio da Habitação assumindo o valor de 1 tende a subir proporcionalmente a elevação da Renda Média (0.687). Tratando-se do valor assumido de −1, significa que existe uma forte correlação negativa; Latitude (−0.142), Longitude (−0.047) e População (−0.026). Com isso podemos reconhecer que, os preços tendem a diminuir quando deseja-se alocar no norte do estado. Por fim, os coeficientes próximos de 0 (Famílias) representam uma correlação baixa.

RESULTADOS E DISCUSSÃO

A etapa de pré-processamento consiste em uma série de processos que visam preparar os dados antes de sua aplicação. Esse processo possui objetivo de organizar, estruturar e limpar os dados de forma a melhorar o desempenho dos modelos a serem utilizados. Algumas Técnicas de AM possuem restrição em relação ao tipo dos dados de entrada quanto a sua escala, ausência de dados e a presença de *outliers*. Para os dados ausentes existentes no banco de dados foram realizadas as transformações necessárias e os 207 valores ausentes da variável Total Quartos foram imputados considerando valor médio da variável (537, 870).

Alguns algoritmos possuem restrição quanto ao tipo de dados dependendo da sua aplicação, considerando o atributo próximo do oceano, por ser um atributo qualitativo nominal foi realizado a transformação dos seus valores para uma escala numérica (0 e 1). Os dados foram normalizados (Figura 2).

Figura 2. Comparação da distribuição do Atributo Alvo (Valor Médio das Casas) na base de dados considerando o conjunto de treinamento e teste



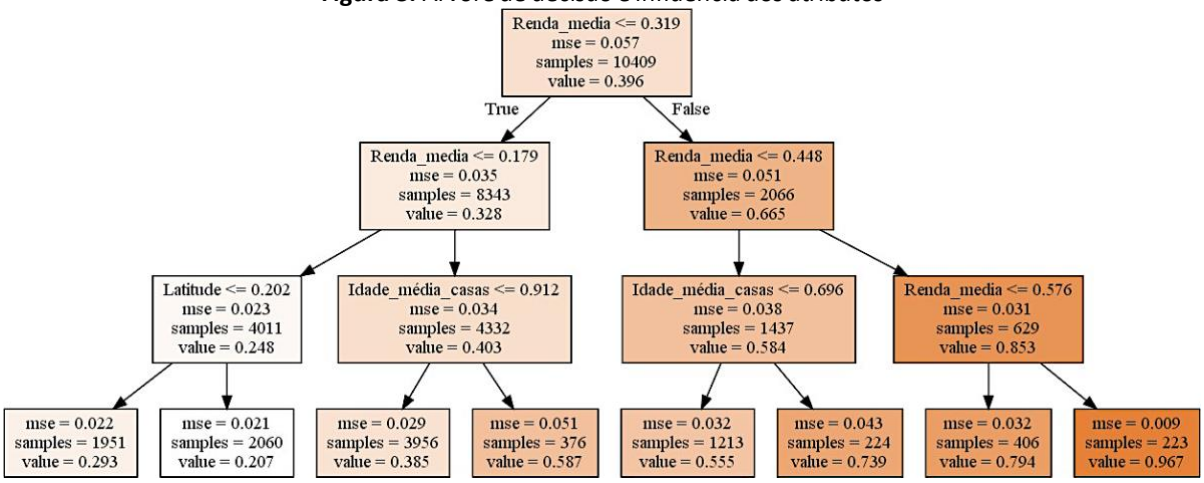
Fonte: Autor (2025).

O modelo de Floresta Aleatória (RF0) foi inicialmente treinado com a configuração padrão básica considerando todos os atributos do banco de dados utilizando a técnica de validação cruzada. A divisão do conjunto para treinamento/teste foi dividida em:

- 80% dos objetos para o conjunto de Treinamento;
- 20% dos objetos para o conjunto de Teste.

Se tratando da validação cruzada, os dados de treinamento e teste possuem a mesma distribuição espacial para o atributo alvo Valor Médio das Casas (Figura 2). Através da análise de correlação dos atributos latitude, idade média das casas e renda média foram definidos sendo os atributos que possuem potencial para influenciar o desempenho do modelo (Figura 3).

Figura 3. Árvore de decisão e Influência dos atributos



Fonte: Autor (2025).

O modelo RF1 foi implementado considerando apenas os 3 atributos indicados anteriormente e as medidas de desempenho foram novamente calculadas. Uma terceira versão do modelo, indicada como RF2, foi testada considerando os melhores parâmetros através do GridSearchCv. Para este trabalho foram consideradas as medidas de desempenho MAE, R-quadrado e RMSE.

O desempenho dos modelos de AM (Figura 4) demonstram que RF0 e RF1 sofreram uma pequena defasagem entre si, tendo resultados quase idênticos, indicando que a exclusão dos demais atributos não influenciam o desempenho do modelo. Já o modelo RF 2 indica uma melhora considerável em relação ao desempenho, que pode ser visivelmente observado.

Figura 4. Comparação dos desempenhos dos modelos de Floresta Aleatória considerando os dados de treinamento e de teste

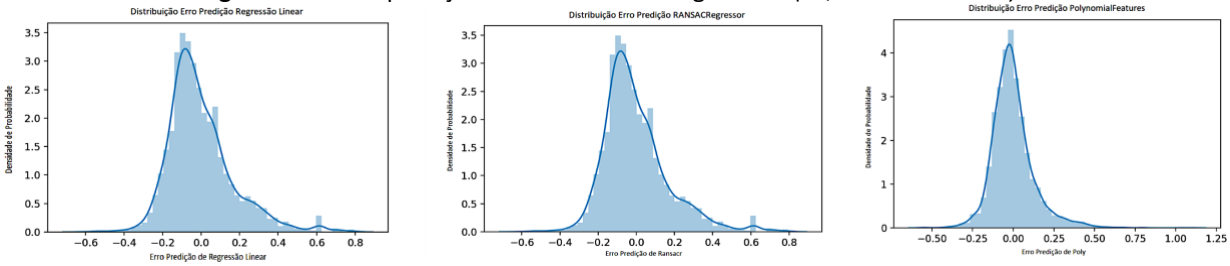
	Model	Descrição	Treinamento			Teste		
			MAE	RMSE	R2	MAE	RMSE	R2
0	RF 0	Random Forest Regressor	0.122651	0.164569	**0.523996**	0.127276	0.170435	**0.475984**
1	RF 1	M. Atributos	0.122556	0.164628	**0.523656**	0.127117	0.170364	**0.476418**
2	RF 2	M. Parâmetros	0.066230	0.097881	**0.831612**	0.102824	0.146254	**0.614129**

Fonte: Autor (2025).

Já o modelo RF2 indica uma melhora considerável em relação ao desempenho, que pode ser visivelmente observado na Figura 4. Porém, existe uma grande variação entre o desempenho observado nos dados de teste em relação aos dados de treinamento, o que indica a possibilidade de ocorrência de *overfitting*, quando o modelo apresenta um alto ajuste aos dados de treinamento. Isso ocorre devido ao número de objetos reduzido, para a correção seria necessário aumentar o número de objetos de modo a melhorar o desempenho do modelo.

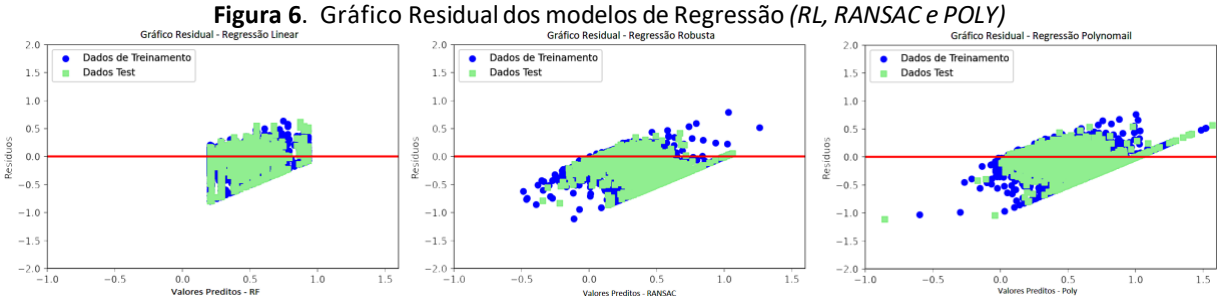
Partindo para o contexto do algoritmo de Regressão Linear, foram testados 3 modelos, o de Regressão linear (RF), Regressão Robusta (RANSAC) e Regressão Polinomial (POLY). Para essa implementação foi realizada a análise dos erros de predição (Figura 5) e Resíduos (Figura 6).

Figura 5. Erro da predição dos modelos de Regressaão *(RL, RANSAC e POLY)*



Fonte: Autor (2025).

Figura 6. Gráfico Residual dos modelos de Regressão *(RL, RANSAC e POLY)*



Fonte: Autor (2025).

Já a Figura 7 apresenta os resultados de desempenho dos modelos de Regressão utilizados considerando os dados de treinamento e teste. Através deste resultado e possível identificar o melhor desempenho do modelo POLY usando a medida de desempenho R-quadrado (0.68) nos dados de teste, sendo a melhor *performance* entre os modelos testados.

Figura 7. Resultados de desempenho dos modelos de Regressão

	Model	Descrição	Treinamento			Teste		
			MAE	RMSE	R2	MAE	RMSE	R2
3	RL	Regressão linear	0.104866	0.143448	**0.638340**	0.105142	0.143998	**0.625943**
4	RAN	RANSAC	0.120869	0.176651	**0.451542**	0.120209	0.175565	**0.443964**
5	Poly	Polyfeatures	0.092779	0.130241	**0.701867**	0.093389	0.131884	**0.686228**

Fonte: Autor (2025).

Para os modelos de regressão a redução dos atributos não apresentou melhoras no desempenho, portanto foram descartadas. Os resultados observados no geral apontam para uma divergência em relação aos dados observados na literatura estudada, onde os modelos de Floresta Aleatória apresentaram resultados superiores aos de Regressão. Uma das justificativas desse comportamento é a baixa quantidade de objetos do banco de dados escolhido, que prejudica o desempenho dos modelos.

CONSIDERAÇÕES FINAIS

Este banco de dados possui atributos referentes a características de localização, caracterização do imóvel, caracterização dos habitantes e tendo como atributo alvo o valor médio das casas. Dos 6 modelos analisados, os que apresentaram melhor *performance* considerando as medidas de desempenho observadas foram os algoritmos de Regressão linear e de Regressão Polinomial, que apresentaram um desempenho de 0.63 e 0.68, respectivamente, sendo considerado os modelos que melhor se ajustaram ao problema.

Os resultados obtidos têm importantes implicações para profissionais do setor imobiliário, incluindo corretores, investidores, construtoras e instituições financeiras. Algumas das principais aplicações incluem:

- Precificação mais precisa de imóveis;
- Modelos baseados em Floresta Aleatória podem ser úteis para empresas que precisam de estimativas rápidas e precisas de preços de imóveis com base em características estruturais e socioeconômicas. No entanto, esses modelos devem ser complementados com técnicas que reduzam o *overfitting*, como regularização ou aumento do conjunto de dados;
- Identificação de padrões de valorização;
- A Regressão Polinomial demonstrou que os preços dos imóveis não seguem uma relação linear simples com as variáveis explicativas. Isso significa que, para identificar tendências de valorização, os profissionais do setor podem utilizar modelos não lineares para prever a evolução dos preços em determinadas regiões;
- Tomada de decisão para investimentos;
- Empresas e investidores podem usar essas abordagens para prever quais bairros ou cidades tendem a se valorizar mais, baseando-se em fatores como infraestrutura, crescimento populacional e renda média da população.

Entretanto, existe uma fragilidade evidente neste estudo em relação ao tamanho da amostra e na quantidade limitada de variáveis disponíveis. O conjunto de dados, obtido do repositório StatLib, contém 20.640 amostras, o que pode ser considerado um volume reduzido para problemas de modelagem de preços imobiliários, especialmente em um mercado dinâmico e multivariado como o da Califórnia. Trabalhos anteriores apontam que a eficácia de modelos de aprendizado de máquina pode ser significativamente afetada pelo número de observações disponíveis, sendo que bases de dados maiores tendem a reduzir o impacto de *overfitting* e melhorar a confiabilidade dos resultados (Yilmazer & Kocaman, 2020). Entretanto a base de dados utilizada não inclui algumas variáveis relevantes que poderiam aprimorar a capacidade preditiva dos modelos, como taxa de criminalidade, infraestrutura urbana, proximidade a centros comerciais e oferta de transporte público. Estudos anteriores indicam que a incorporação de tais variáveis pode aumentar significativamente a acurácia na predição de preços imobiliários (Park & Bae, 2015). A ausência desses fatores pode ter limitado a capacidade dos modelos de capturar plenamente os padrões subjacentes ao comportamento do mercado imobiliário.

Já em relação ao algoritmo de Floresta Aleatória, apesar do desempenho superior considerando os melhores parâmetros, existem evidências de *overfitting*, sobre a capacidade de generalização e precisão de predição dos resultados obtidos, que não dependem apenas da qualidade do conjunto de dados, mas também do número de objetos. Dado o conjunto de dados usado neste artigo, nossa principal conclusão é que os modelos de Regressão Linear e Regressão Polinomial são capazes de gerar estimativas de preços comparativamente precisas com erros de predição mais baixos, em comparação com os resultados do modelo de Floresta Aleatória.

Com isso, a complexidade de se obter um modelo e o impacto da qualidade do banco de dados no desempenho dos modelos. Por fim, recomenda-se como trabalhos futuros inserir mais atributos que possam ter correlações positivas para precificações do setor imobiliário próximo da realidade e disponibilizar um banco de dados maior para implementações de novos modelos.

REFERÊNCIAS

Barr, J., Ellis, E., Kassab, A., Redfearn, C., Srinivasan, N., & Voris, K. (2015). Home price index: A machine learning methodology. *International Journal of Semantic Computing*, 11(1), 111-133. https://doi.org/10.1142/S1793351X17500015

Breiman, L. (2001). Random forests. *Machine Learning*,45(1), 532. https://doi.org/10.1023/A:1010933404324

Ceh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *International Journal of Geo-Information*, 7(5), 168. https://doi.org/10.3390/ijgi7050168

Ferreira, S. F. & Filho, R. (2010). Aplicação do método de preços hedônicos na precificação de atributos raros de peças filatélicas e construção de carteiras eficientes. *Estudos Econômicos*, 40(2), 469-498. https://doi.org/10.1590/S0101-41612010000200008

Geron, A. (2019). *Mãos à obra: Aprendizado de máquina com Scikit-Learn e TensorFlow* (1ª ed.). Alta Books.

Ho, K., Tang, B., & Wong, S. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48-70. https://doi.org/10.1080/09599916.2020.1832558.

Hong, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, 24(3), 140-152. https://doi.org/10.3846/ijspm.2020.11544

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy,* 74(2), 132-157. https://doi.org/10.1086/259131

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Moita Neto, J. M. (2004). Introdução à análise multivariada para as ciências sociais. *Editora UFRN.*

Park, B. & Bae, J. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications,* 42(1), 292-303. https://doi.org/10.1016/j.eswa.2014.11.040

Park, B. & Bae, J. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(1), 292-303. https://doi.org/10.1016/j.eswa.2014.11.040

S-PLUS (1998). Guide to statistical and mathematical analysis. *Insightful Corporation.*

Yilmazer, S. & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy*,99, 104889. https://doi.org/10.1016/j.landusepol.2020.104889