UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO,
ATUÁRIA E CONTABILIDADE - FEAAC
CURSO DE PÓS-GRADUAÇÃO EM ECONOMIA - CAEN
MESTRADO PROFISSIONAL EM ECONOMIA DO SETOR PÚBLICO MESP

VICTOR JUCÁ TÁVORA

MACHINE LEARNING INTERPRETÁVEL NA AVALIAÇÃO DE IMÓVEIS EM MASSA: APLICAÇÃO DO SHAP

FORTALEZA
2025

VICTOR JUCÁ TÁVORA

MACHINE LEARNING INTERPRETÁVEL NA AVALIAÇÃO DE IMÓVEIS EM MASSA: APLICAÇÃO DO SHAP

Dissertação apresentada ao Programa de Pós-Graduação em Economia do Setor Público da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Economia. Área de concentração: Economia do Setor Público.

Orientador: Prof. Dr. Diego Rafael Fonseca Carneiro

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

T237m Távora, Victor Jucá.
Machine Learning interpretável na avaliação de imóveis em massa: aplicação do SHAP / Victor Jucá Távora. – 2025.
64 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Mestrado Profissional em Economia do Setor Público, Fortaleza, 2025.
Orientação: Prof. Dr. Diego Rafael Fonseca Carneiro.

1. Avaliação em massa. 2. aprendizagem de máquinas. 3. SHAP. I. Título.

CDD 330

VICTOR JUCÁ TÁVORA

MACHINE LEARNING INTERPRETÁVEL NA AVALIAÇÃO DE IMÓVEIS EM MASSA: APLICAÇÃO DO SHAP

Dissertação apresentada ao Programa de Pós-Graduação em Economia do Setor Público da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Economia. Área de concentração: Economia do Setor Público.

Aprovada em 17/11/2025.

BANCA EXAMINADORA

_____

Prof. Dr. Diego Rafael Fonseca Carneiro (Orientador)
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Andrei Gomes Simonassi
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Francisco Gildemir Ferreira da Silva
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

A Deus, pela vida e tudo o que dela decorre.

Ao Professor Raimundo Nonato Távora Costa, meu pai e primeiro professor, ativo na Universidade Federal do Ceará há mais de 50 anos, por, mesmo sem precisar falar, conseguir, por meio de seu exemplo, me ensinar sobre ser estudante, servidor público e, sobretudo, humano.

À Maria Angélica Jucá Távora, minha mãe, por ter me ensinado o que é amor, com sua dedicação incondicional a mim e ao meu irmão.

À Maria Thereza Leite, por seu amor, companhia, carinho e paciência.

Ao meu irmão, Luiz Eurico Jucá Távora, pela parceria de sempre.

Ao Prof. Diego Rafael Fonseca Carneiro, orientador, pelas contribuições ao longo da elaboração do trabalho.

À Secretaria Municipal das Finanças, pelo incentivo na participação neste Mestrado, por meio do Programa Permanente de Formação do Servidor Fazendário Municipal - PFORMS.

Aos antigos colegas de trabalho que se tornaram amigos ao longo da trajetória profissional, Antonio Rinaldo, Igor Gabriel e João Augusto, que me moldaram como profissional e pessoa. Também ao amigo Paulo Mateus, pela sua amizade e parceria.

À Fernandinha e à Tati, pelo acolhimento sensacional no Cadastro Imobiliário do Município e pela gestão humana e profissional com que conduzem o setor.

Aos companheiros de setor Augusto, Sandro, Luan, Klinsman e Luis, pela dedicação inspiradora no trabalho e por fomentarem um ambiente de constante busca por novos conhecimentos e inovação.

Por fim, agradeço aos professores membros da banca examinadora, Prof. Andrei e Prof. Gildemir, pela disponibilidade e atenção com que avaliaram o trabalho.

RESUMO

Este trabalho tem como objetivo avançar com o emprego de algoritmos de *machine learning* no contexto de avaliações de imóveis em massa, ao explorar e validar ferramentas que se prestam a tornar esses modelos mais interpretáveis e explicáveis. Utilizando uma amostra de 6.660 apartamentos do município de Fortaleza/CE, provenientes da Secretaria Municipal das Finanças, comparou-se o modelo tradicional de regressão linear múltipla com os algoritmos *XGBoost* e *Random Forest*. O *XGBoost* apresentou o melhor desempenho em todas as métricas avaliadas (MAPE, RMSE, R^2, COD e PRD). Para lidar com o desafio de interpretabilidade, aplicou-se a técnica SHAP (*Shapley Additive Explanations*), que permite analisar a contribuição individual de cada variável e compreender, de forma global, as relações capturadas pelo modelo. A comparação entre gráficos de dependência parcial do modelo clássico e do *XGBoost*, elaborados com a biblioteca SHAP, mostrou que este capturou uma relação não linear entre idade e preço unitário, além da interação entre idade e renda. Também se pôde observar a diversidade de efeitos em variáveis dicotômicas. A análise local revelou que a contribuição das variáveis varia significativamente entre imóveis de diferentes contextos. A técnica SHAP demonstrou-se útil para promover transparência e facilitar a compreensão dos modelos por técnicos e gestores públicos, contribuindo para a justiça fiscal e a comunicação com os contribuintes. A visualização por meio de *waterfall plots* mostrou-se eficaz para apresentar predições individuais sem exigir conhecimento aprofundado sobre os algoritmos. O estudo reforça a viabilidade do uso de inteligência artificial interpretável em avaliações fiscais e propõe sua adoção para fortalecer e ampliar o uso de algoritmos baseados em árvore no contexto tributário.

Palavras-chave: Avaliação em massa; aprendizagem de máquinas; SHAP.

ABSTRACT

This study aims to advance the use of machine learning algorithms in mass real estate appraisal context by exploring and validating tools designed to enhance model interpretability and explainability. Using a dataset of 6,660 apartments from Fortaleza, Brazil, provided by the Municipal Finance Department, the performance of a traditional multiple linear regression model was compared with that of modern machine learning algorithms, namely XGBoost and Random Forest. XGBoost achieved the best performance across all evaluated metrics (MAPE, RMSE, R^2, COD, and PRD). To address the challenge of interpretability, SHAP (Shapley Additive Explanations) method was employed, enabling the assessment of each variable's individual contribution and a global understanding of relationships captured by the model. Comparison between partial dependence plots from the classical model and SHAP-based plots for XGBoost revealed a non-linear relationship between building age and unit price, as well as an interaction between age and neighborhood income levels. Diverse effects were also observed in binary variables. The local analysis showed that the influence of predictors varies substantially across properties with different characteristics. SHAP proved valuable for enhancing transparency and improving model interpretability for both technical and public administration purposes, contributing to greater tax fairness and better communication with taxpayers. Waterfall plots, in particular, demonstrated effectiveness in explaining individual predictions without requiring in-depth technical knowledge of the algorithms. This study highlights the feasibility of integrating explainable artificial intelligence into fiscal mass appraisal workflows and recommends its adoption to strengthen and expand the use of tree-based algorithms in property taxation.

Keywords: Mass appraisal; machine learning; SHAP.

LISTA DE FIGURAS

Figura 1 - Série temporal da taxa Selic (2000-2024).....12
Figura 2 - Série histórica da participação da Construção Civil no PIB (2000-2024).....13
Figura 3 - Série histórica da arrecadação de ITBI do Município de Fortaleza (2009-2024). Valores corrigidos para 2024.....13
Figura 4 - Exemplo de árvore de decisão 18
Figura 5 - Ilustração do efeito de viés e variância.....19
Figura 6 - Gráfico do dilema viés variância 20
Figura 7 - Esquema de funcionamento de modelos do tipo *bagging* 21
Figura 8 - Exemplo de cálculo de valor de Shapley. 29
Figura 9 - Exemplo de *force plot*.....31
Figura 10 - Gráfico de valores SHAP de cascata 32
Figura 11 - Dependência parcial com SHAP.....33
Figura 12 - *Summary plot* 33
Figura 13 - *Heatmap plot* 33
Figura 14 - *Heatmap plot ordenado* 33
Figura 15 - Mapa de renda do chefe de família, elaborado a partir dos dados do Censo do IBGE, de 2022 36
Figura 16 - Localização dos dados utilizados, com os respectivos valores.....39
Figura 17 - Matriz de correlação 40
Figura 18 - Correlações com a variável alvo.....41
Figura 19 - Dispersão: valores preditos (eixo y) *vs* valores observados (eixo x).....45
Figura 20 - *Summary plot* (*XGBoost*).....46
Figura 21 - *SHAP values* de ambos os modelos para a variável *of* 48
Figura 22 – SHAP *values* de ambos os modelos para a variável *idade*.....49
Figura 23 - Gráficos de dependência parcial para *distbm XGboost* (à esquerda) e clássico (à direita).....50
Figura 24 - Gráficos de dependência parcial para *idade*. 50
Figura 25 - Gráfico de dependência parcial da *idade*, relacionado com *renda* 51
Figura 26 - Análise local com *waterfall plot* – GI952431.....53
Figura 27 - Análise local com *waterfall plot* - GI947836 54

SUMÁRIO

1. INTRODUÇÃO 8
2. REVISÃO DE LITERATURA 11
2.1 O mercado de imóveis 11
2.2 Avaliação em massa e métricas de performance 14
2.3 Abordagens tradicionais 16
2.4 Modelos de aprendizagem de máquinas 17
2.4.1 Árvore de decisão 17
2.4.2 Métodos Ensemble 20
2.4.3 Aplicações de machine learning para avaliação em massa 23
2.5 Técnicas de interpretação 26
2.5.1 Permutation Feature Importance 26
2.5.2 Gráfico de dependência parcial 27
2.5.3 SHAP 28
3. MATERIAL E MÉTODOS 35
3.1 Caracterização e análise exploratória dos dados 35
3.2 Modelo de regressão linear 41
3.3 Random Forest e XGBoost 42
4. RESULTADOS E DISCUSSÃO 43
4.1 Performance 43
4.2 Análise global com SHAP 45
4.3 Análise local com SHAP 52
5. CONCLUSÃO 57
REFERÊNCIAS 59
APÊNDICE A – HISTOGRAMAS DAS VARIÁVEIS DA AMOSTRA 63

1. INTRODUÇÃO

Imóveis são objeto de tributos sob diferentes modalidades, como em transferências (ITBI ou ITCMD), propriedade ou posse (IPTU ou ITR) e ganho de capital (IR). As receitas advindas desses tributos imobiliários constituem uma parcela relevante dos orçamentos dos entes federados - cerca de 1/3 da receita própria de impostos, para Fortaleza, de acordo com a Lei Orçamentária de 2025 (Fortaleza, 2025) - e são desvinculadas, podendo ser aplicadas para gastos correntes ou investimentos. Em geral, essas cobranças têm em comum o fato de o valor venal servir como base de cálculo.

No caso dos tributos imobiliários recorrentes, os quais geralmente incidem sobre posse ou propriedade, é adequado que a cobrança tenha como referência o valor de mercado do imóvel, para que se evitem distorções, como nos casos em que imóveis de menor valor de mercado são tributados em valor proporcionalmente maior que os de maior valor, fenômeno conhecido como regressividade.

Já para os tributos incidentes sobre transferências de imóveis, sejam elas onerosas ou não, é fundamental para o fisco conhecer os valores do bem, para julgar se os preços declarados pelos contribuintes estão alinhados com os valores praticados no mercado e, assim, decidir pelo seu uso como base de cálculo. Essa prática contribui para coibir subdeclarações e, consequentemente, reduzir a sonegação fiscal.

O processo de avaliar os diversos imóveis existentes em uma região é conhecido como “avaliação em massa” e podem ser aplicadas abordagens variadas (IAAO, 2013). As técnicas mais tradicionais se baseiam na teoria de preços hedônicos, segundo as quais são aplicados modelos de regressão linear múltipla. Conforme a teoria, o bem é virtualmente particionado, e cada atributo do imóvel tem sua importância relativa atribuída. O valor global do bem é explicado pelo somatório dos valores atribuídos a cada característica. Em uma casa, por exemplo, esses atributos poderiam ser a área construída do imóvel, área do terreno, número de quartos e idade do imóvel.

Embora esse método seja tradicional e amplamente utilizado na avaliação imobiliária, a literatura não define rigorosamente uma forma funcional para os modelos (González e Formoso, 1994), cabendo aos avaliadores adaptá-los conforme o cenário exige. Ao final do processo, o técnico poderá avaliar a importância relativa de cada atributo do imóvel por meio da análise dos parâmetros, bem como analisar o atendimento às suposições fundamentais de um modelo econométrico.

Essa facilidade de checagem e compreensão é uma vantagem importante do modelo de preços hedônicos. No entanto, em termos de performance para fins de avaliação em massa, os métodos tradicionais tendem a ser superados por outros mais modernos baseados em algoritmos de aprendizagem de máquinas (*machine learning*), conforme observado por diversos pesquisadores (Antipov e Pkryshevskaya, 2010; Carranza et al., 2018; Oliveira, 2020; Zilli e Bastos, 2024; Oliveira et al., 2024).

Apesar das vantagens de performance, esses algoritmos apresentam desafios relacionados a interpretabilidade e transparência, sendo frequentemente considerados caixas pretas. Segundo Iban (2022), não basta que o algoritmo seja capaz de fazer predições precisas, mas deve possibilitar a atribuição da importância de cada variável de entrada no processo de predição, por questão de transparência e justiça.

Nesse sentido, o crescimento do uso e dos estudos de algoritmos baseados em árvore de decisão para avaliação em massa fomenta a aplicação de técnicas que visem a explicá-los. Esse conjunto de técnicas é conhecido como *eXplainable Artificial Intelligence* (XAI), que possibilita a análise global e local de um modelo. A análise global identifica as *features* mais relevantes do modelo como um todo. Uma das técnicas utilizadas para essa seleção é a *Permutation Feature Importance* (PFI). A partir dela, podem-se omitir *features* com pouca importância e retreinar o modelo para melhorar sua performance. Já a análise local identifica o peso de cada *feature* na predição de um valor específico, o que pode ser feito com a técnica SHAP.

Este trabalho tem como objetivo geral a aplicação de técnicas XAI na elaboração e análise de um modelo de avaliação em massa para apartamentos no município de Fortaleza/CE.

Já os objetivos específicos são i) comparar a performance de modelos de avaliação em massa de algoritmos baseados em árvores com o tradicional de preços hedônicos; ii) utilizar a técnica para seleção das *features* mais relevantes, a fim de melhorar a performance dos modelos de *machine learning*; iii) aplicar técnica SHAP para análise global do modelo elaborado com algoritmo baseado em árvores e comparar os resultados com o modelo tradicional; iv) aplicar técnica SHAP para análise local de resultados do modelo elaborado com algoritmo baseado em árvores; v) discutir a possibilidade de melhorias, a partir de técnicas *XAI*, dos processos de órgãos públicos que utilizam algoritmos baseados em árvore para avaliação de imóveis.

Este trabalho está estruturado em cinco seções, sendo esta Introdução a primeira delas. Na seção dois, Referencial Teórico, é apresentada a revisão de literatura a respeito das técnicas de avaliação em massa e interpretação de seus resultados.

Na seção três, Material e Métodos, são caracterizados o município de Fortaleza e a amostra de apartamentos que servirá como base para treinamento e teste dos modelos. Também são apresentadas as variáveis, ou *features*, utilizadas, bem como o procedimento que estruturou esta pesquisa.

Na seção quatro, Resultados e Discussão, são apresentados os resultados de performance obtidas pelas diferentes abordagens e predição. Também são discutidas as interpretações global e local dos resultados dos algoritmos de aprendizagem de máquinas para diferentes apartamentos. Por fim, na seção cinco, Conclusão, são feitas as considerações finais e sugestões para trabalhos futuros.

2. REVISÃO DE LITERATURA

2.1 O mercado de imóveis

Nasser Júnior (2019) explica que o mercado imobiliário possui comportamento muito distinto dos mercados de outros bens, por existirem inúmeras fontes de desigualdades entre os imóveis, além de alterações no entorno também serem capazes de modificar seus valores. Além das características físicas dos próprios imóveis, o autor cita outros fatores que influenciam esse mercado, como a oferta de crédito, a inflação, a condução da economia, as políticas fiscais, o crescimento demográfico e a confiança no governo.

Ainda na década de 1980, Lucena (1985) estudou o mercado de habitações no Brasil, tendo obtido achados relevantes. O autor observou que i) o aumento da demanda por moradia gerado por aumentos de renda é fortemente correlacionado com a elevação dos custos de construção, o que, por sua vez, limita a expansão da oferta de novas unidades e impede que o volume de transações cresça na mesma proporção. Assim, o autor caracterizou o comportamento da produção de novas habitações como anticíclico; ii) com base em modelos de preços hedônicos, concluiu que variações na renda impactam as classes sociais de modo heterogêneo. Enquanto classes de maior renda passam a demandar por mais acessibilidade (localização), as camadas de menor renda priorizam maior área construída; iii) os preços das habitações podem ser determinados pelos preços de seus componentes. O autor chama atenção para a factibilidade de dividir a habitação em um número limitado de serviços e determinar valores específicos de acordo com a demanda e a oferta de cada um; e iv) o governo, ao descentralizar investimentos em infraestrutura básica, pode reduzir os custos de dessas amenidades, aumentando o excedente o consumidor e gerando benefícios à sociedade.

No contexto microeconômico, Nasser Júnior (2019) afirma que o mercado imobiliário fere suposições do mercado de concorrência perfeita, haja vista a heterogeneidade espacial e de características dos produtos e a desigualdade ou falta de informações entre os diversos agentes. A consequência disso é que o preço praticado não coincide, necessariamente, com o valor, existindo, assim, uma faixa de preços razoáveis, dentro da qual está o valor de mercado (valor mais provável ou esperado) para o bem.

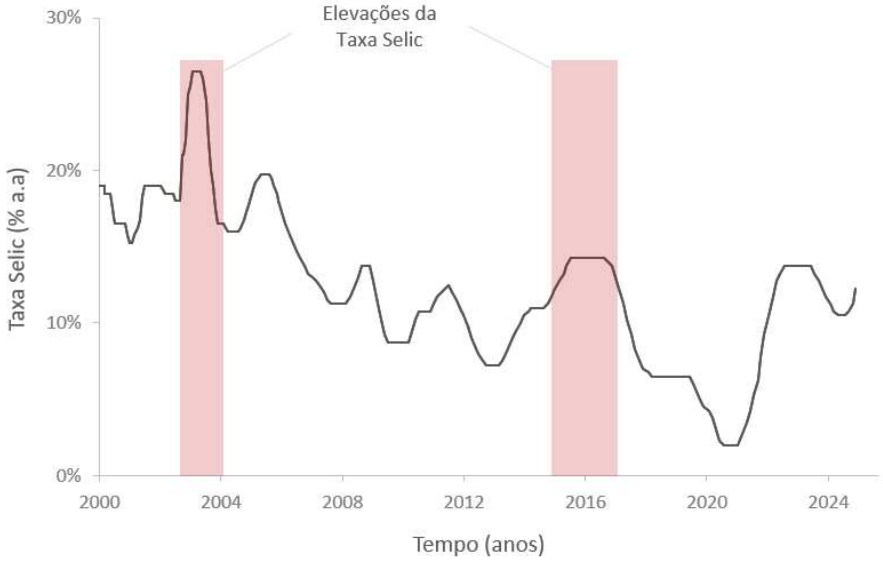
Já no contexto macroeconômico, Silva et al. (2012) evidenciam a relação entre o mercado imobiliário e a taxa básica de juros. Segundo os autores, devido aos longos períodos inerentes às operações de empréstimo destinadas aos financiamentos do setor, patamares altos

da taxa básica de juros reduzem os investimentos. Com isso, no Brasil, o setor se tornou mais atraente a partir de 2003, quando a taxa Selic se estabilizou abaixo de 20% ao ano.

Guedes, Iachan e Sant'Anna (2022), ao investigarem a oferta de moradias formais e informais no Brasil, concluíram, em consonância com outros pesquisadores, que o mercado imobiliário nacional apresenta elevado grau de inelasticidade, haja vista a rigidez da oferta. Nesse contexto, choques de demanda — provocados, por exemplo, por reduções na taxa de juros — tendem a se refletir principalmente em elevação de preços, uma vez que o mercado reage lentamente no ajuste da quantidade ofertada. Essa rigidez é compreensível, dado que a produção habitacional envolve prazos longos de execução, elevados custos fixos e restrições urbanísticas e regulatórias que limitam respostas rápidas da oferta.

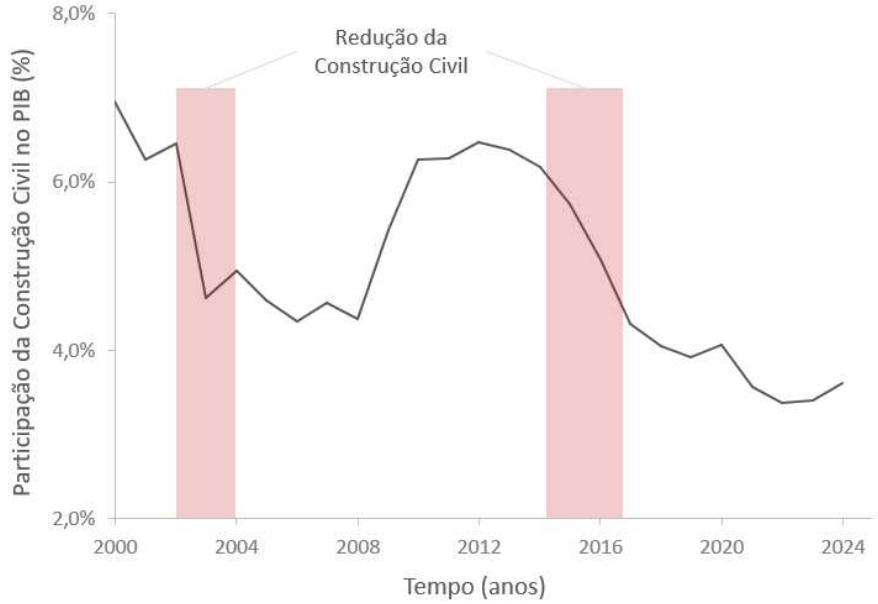
Nesse contexto, a Figura 1 mostra a série temporal da taxa básica de juros (Selic) no Brasil, com dados do Banco Central (BACEN, 2025). Ao relacioná-lo com a série de participação da Construção Civil no PIB do Brasil do mesmo período, com dados do IBGE (IBGE, 2025) (Figura 2), pode-se observar a relação entre a taxa de juros e o mercado da construção. A análise gráfica sugere influência negativa da política monetária sobre o setor, o que é esperado de acordo com a literatura já citada.

Figura 1 - Série temporal da taxa Selic (2000-2024)



Fonte: Elaboração própria, com dados do Banco Central do Brasil (BACEN, 2025).

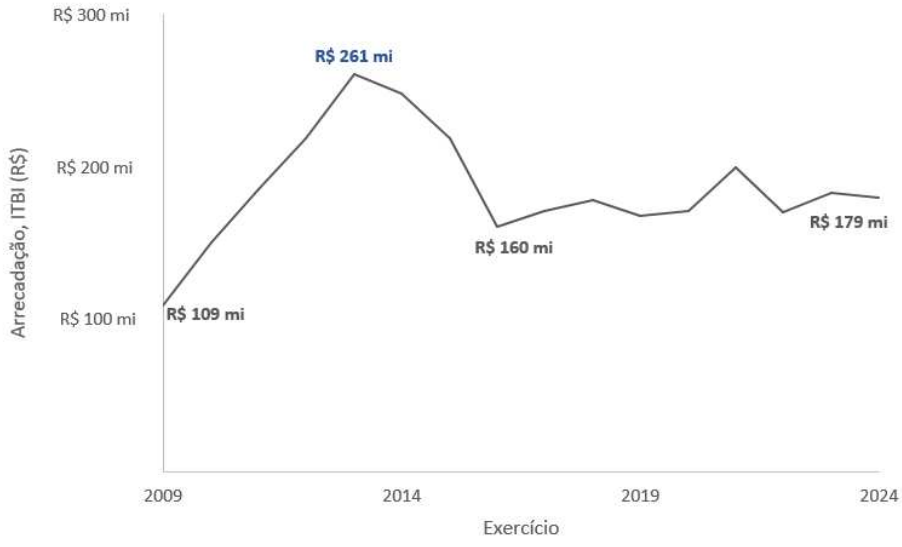
Figura 2 - Série histórica da participação da Construção Civil no PIB (2000-2024)



Fonte: Elaboração própria, com dados do IBGE (IBGE, 2025).

Essa oscilação da taxa básica de juros e do setor de construção civil, invariavelmente, afetam o mercado imobiliário, o qual, por sua vez, impacta a arrecadação dos tributos imobiliários. De acordo com informações dos Relatórios Resumidos de Execução Orçamentária dos anos de 2009 a 2024 (Fortaleza, 2025) (Figura 3), considerando os valores corrigidos para 2024, o período com maior arrecadação coincide com períodos de grande participação da construção civil no PIB e taxa básica de juros estável, em 2013.

Figura 3 - Série histórica da arrecadação de ITBI do Município de Fortaleza (2009-2024). Valores corrigidos para 2024.



Fonte: Elaboração própria, com dados do Secretaria Municipal das Finanças de Fortaleza (Sefin, 2025).

2.2 Avaliação em massa e métricas de performance

Segundo a *International Association of Assessing Officers* (IAAO, 2017), "avaliação em massa é o processo de valorar um grupo de propriedades em uma determinada data, utilizando dados comuns, métodos padronizados e testes estatísticos".

Essa atividade tem finalidades e interessados diversos, como órgãos fiscais, instituições financeiras, investidores do mercado imobiliário, construtoras e incorporadoras, seguradoras e outros. Nesse sentido, a avaliação em massa de propriedades tem ganhado importância devido à grande participação do mercado imobiliário nas medidas econômicas, tornando-se um dos indicadores de desenvolvimento em vários países (Yilmazer et al., 2020).

No contexto fiscal, essas avaliações desempenham um papel muito útil na determinação da base de cálculo dos impostos dentro da jurisdição dos municípios, como o Imposto Predial e Territorial Urbano (IPTU) no Brasil (Zilli et al., 2024). No município de Fortaleza, por exemplo, Gimenes (2020) analisou o impacto que a atualização dos valores venais dos imóveis, para fins de IPTU, traria no município, considerando a defasagem da planta de valores do município à época. Mantendo a política tributária vigente até então, a atualização poderia resultar em um incremento na arrecadação do IPTU superior a 350%.

Grover (2016) destacou que a elaboração desses modelos requer procedimentos padronizados e um expressivo conjunto de dados confiáveis. No entanto, esse cenário pode ser difícil de se configurar em países em desenvolvimento, onde os mercados imobiliários tendem a não ser tão abertos quanto necessário.

Por outro lado, Oliveira et al. (2024) apontaram que as técnicas de avaliação em massa estão em evolução. Os autores citam a maior disponibilidade e facilidade de obtenção de características e preços de propriedades a partir de portais *online*, bem como a manutenção de observatórios do mercado imobiliário por instituições públicas e privadas como ferramentas que impulsionam a atividade.

No Brasil, a norma que padroniza a atividade de avaliação de imóveis não trata de avaliação em massa. Já a norte-americana IAAO (2017), embora não defina a técnica a ser aplicada, elenca métricas que permitem a aferição da qualidade e comparação entre diferentes modelos, na norma *Standard on Ratio Studies – A criteria for measuring fairness, quality equity and accuracy*. Essas métricas são utilizadas na maior parte dos trabalhos nos quais se desenvolvem modelos para avaliação em massa (Kochulem et al., 2018; Carranza et al., 2018; Bandeira, 2019; Oliveira, 2020; Iban, 2022; Deppner et al., 2023; Zilli e Bastos, 2024 e Tarasov, 2024).

Essas métricas, em uma tradução livre, são o coeficiente de determinação (R^2), o erro percentual absoluto médio (do inglês, *mean absolute percentage error*, MAPE), a raiz quadrada do erro quadrático médio (*root mean square error*, RMSE), o coeficiente de dispersão (*coefficient of dispersion*, COD) e o diferencial relativo de preços (*price-related differencial*), as quais são apresentadas nas equações 1 a 5, abaixo.

$$R^2 = \left(\frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}\right) \quad (01)$$

$$MAPE = \frac{100\%}{n} \sum_{1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i|} \quad (02)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2} \quad (03)$$

$$COD = \frac{\frac{100}{n}\sum_{1}^{n}\left|R_i - \tilde{R}\right|}{\tilde{R}} \quad (04)$$

$$PRD = \frac{\bar{R}}{W_t\bar{R}} \quad (05)$$

Nas equações 1, 2 e 3, n é o número de dados, y_i é o valor observado, $\hat{y}$ é o valor predito e $\bar{y}$ é a média dos valores observados. Na equação 4, R_i é a razão entre predito e observado, e $\tilde{R}$ é a mediana desses valores. Na equação 5, $\bar{R}$ é a média dos R_i, e $W_t\bar{R}$ é a razão entre a soma dos valores preditos e observados.

Enquanto R^2, MAPE e RMSE tratam do desvio das predições em relação às observações, o COD mede a dispersão das razões entre predito e observado em relação à mediana, ou seja, o desvio percentual médio da razão predito e observado à mediana dessas razões.

Já o PRD é um indicador de equidade vertical, avaliando se os desvios se relacionam com os preços observados. Essa métrica é importante sobretudo quando a avaliação em massa é utilizada para fins fiscais, para evitar injustiça fiscal. Quando os imóveis de baixo valor são superavaliados e os de alto valor são subestimados, diz-se que a avaliação é regressiva (PRD > 1,0). Já na situação inversa, a avaliação é considerada progressiva (PRD < 1,0). De acordo com a IAAO (2017), para fins de cobrança de tributos, os modelos não devem ser regressivos nem progressivos, ou seja, devem ter PRD próximos de 1,0.

2.3 Abordagens tradicionais

De acordo com a teoria da demanda por características de Lancaster (1966), a satisfação do consumidor não se dá diretamente pelo bem, mas pelas características particulares que o compõem. Rosen (1974) aplica essa ideia ao mercado imobiliário formulando a teoria dos preços hedônicos, segundo a qual os atributos do imóvel explicam seu valor de mercado.

Esses modelos, baseados em regressão linear múltipla, têm sido usados por um longo período. Em mercados urbanos, pelo menos desde 1970, segundo Oliveira et al. (2024). Codes (2018) afirma que essas regressões têm sido amplamente utilizadas nas investigações das relações entre os preços dos imóveis e as suas respectivas características, nas últimas décadas.

Dantas, Magalhães e Vergolino (2007) afirmaram que, por meio da econometria tradicional, faz-se uma regressão dos preços dos imóveis sobre suas características estruturais, locacionais e econômicas. Os autores citam como características de um bem área privativa, número de cômodos, vagas na garagem, idade, conservação, padrão construtivo, bairro, distância a polos de influência, forma de pagamento, época da transação e natureza do evento. Além disso, frisaram que o modelo tem como premissa a independência das observações entre si, o que poderia ferir a natural dependência espacial do ramo imobiliário.

González e Formoso (1995) aplicaram essa metodologia para estimar modelos para locação de imóveis residenciais em Porto Alegre, com dados de 504 imóveis ofertados em junho de 1992. Esses dados foram divididos em quatro subgrupos em função da quantidade de quartos e da distância ao centro, a fim de verificar a existência de um modelo único que regesse a formação dos preços de aluguel. Os modelos atingiram mais de 80% de explicação da variabilidade dos valores, sendo atestadas as importâncias de variáveis locacionais, como distâncias aos *shopping centers* e existência de favelas nas proximidades. No entanto, a hipótese da existência de um modelo único não foi confirmada, de modo que houve diferenças nas significâncias e relevâncias das variáveis para os diferentes subgrupos da amostra.

No que diz respeito à forma funcional do modelo de formação dos preços, Tarasov e Śliwiński (2024) reconhecem a importância dos modelos econométricos, úteis para predição e facilmente interpretáveis. No entanto, citam como pontos negativos algumas limitações apontadas por outros pesquisadores, como a imposição de linearidade e fixação dos parâmetros (Osland, 2010), que nem sempre correspondem à realidade. Essa limitação pode ter impacto reduzido em modelos para predições individuais, mas, no caso de avaliação em massa, a generalização dos efeitos pode ocasionar desvios relevantes. Isso se deve ao fato de os modelos

para avaliação em massa serem mais genéricos do que aqueles elaborados para avaliações únicas.

Corroborando com a ideia, Pelli Neto e Morais (2006) dão ênfase ao fato de que a aplicação da regressão linear requer o atendimento aos seus pressupostos básicos, o que é difícil de se observar no contexto imobiliário. Especialmente nesse ramo, a abordagem encontra dificuldades em duas questões de grande importância, que são a autocorrelação espacial e o desconhecimento da forma funcional para o modelo a ser adotado.

2.4 Modelos de aprendizagem de máquinas

Aprendizado de máquinas é uma subárea da ciência de dados, que por sua vez faz parte da inteligência artificial (Oliveira, 2020). Esses modelos podem ser divididos em algoritmos de aprendizado supervisionado e não supervisionado. No aprendizado supervisionado, o modelo é treinado com uma amostra de dados rotulados, ou seja, com a resposta esperada, enquanto no aprendizado não supervisionado o modelo busca identificar padrões sem a necessidade de rótulos.

Na avaliação em massa, são aplicados métodos de regressão por aprendizagem supervisionada, entre os quais podem-se citar *KNN*, regressão linear, redes neurais artificiais, métodos baseados em árvores e outros. Os modelos baseados em árvores, objetos de estudo deste trabalho, se baseiam em árvores de decisão, algoritmo que particiona conjuntos de dados em função de regras que maximizam o ganho de informação a cada ramificação.

2.4.1 Árvore de decisão

O algoritmo de árvores de decisão é um dos mais simples entre os modelos baseados em árvore. Ele pode ser utilizado tanto para classificação, nos casos de variáveis categóricas, como para regressão, quando a variável alvo assume valores contínuos.

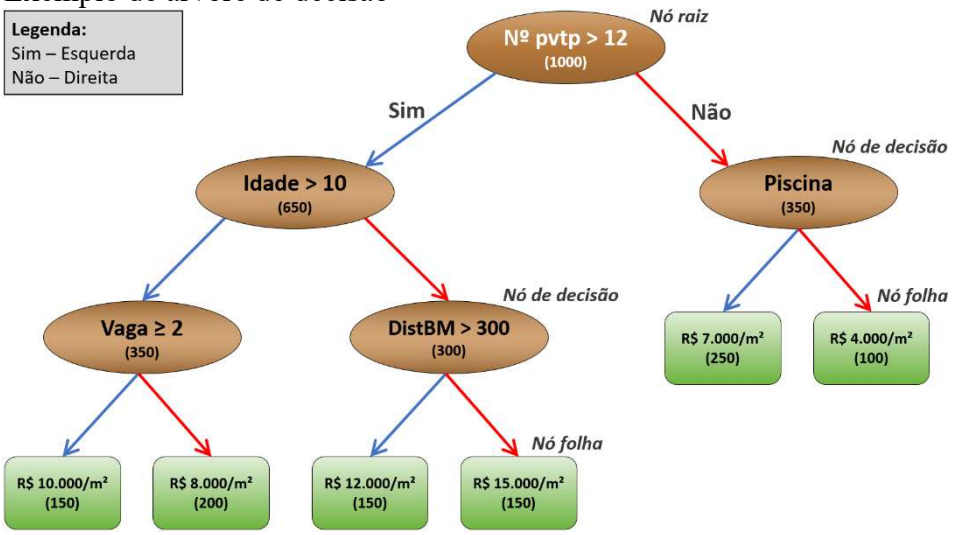
Molnar (2025) explica que o método consiste em particionar iterativamente uma amostra em duas partes, em função de uma variável de entrada específica por vez. A primeira divisão se dá no nó raiz, que contém toda a amostra de treino— isto é, o conjunto de dados usado para ajustar o modelo —, e a *feature* utilizada na divisão deve ser a que permita a melhor separação dos dados, considerando a variável alvo.

Após a primeira divisão, a amostra inicial que foi separada em duas partes poderá ser novamente particionada, em um novo nó, denominado nó de decisão. O modo como a nova

divisão se dá é o mesmo do nó raiz, ou seja, separando a subamostra em função de uma *feature* de tal sorte que os dois grupos resultantes da divisão sejam o mais homogêneo possível internamente e, consequentemente, distintos entre si.

A estrutura final da árvore de decisão parte do nó raiz e passa por nós de decisão, nos quais são feitas novas separações, até atingir os "nós folhas", quando se alcança um nível de pureza satisfatório, a partir da qual não é feita mais nenhuma separação. Com os valores dos nós finais, podem ser realizadas previsões de novos dados a partir das *features* de entrada. A Figura 4 mostra um exemplo de árvore de decisão pequena que retorna o valor unitário de um imóvel.

Figura 4 - Exemplo de árvore de decisão



Fonte: Elaborado pelo autor.

O critério para escolha de uma *feature* de referência para particionar a amostra depende do grau de impureza das duas subamostras que resultarão dessa decisão. O procedimento consiste em buscar um ponto de separação que gere grupos com o mínimo de impureza possível, ou seja, separe os dados em grupos o mais homogêneo possível e diferentes entre si. Comumente, a medida de impureza se dá por entropia ou coeficiente de Gini.

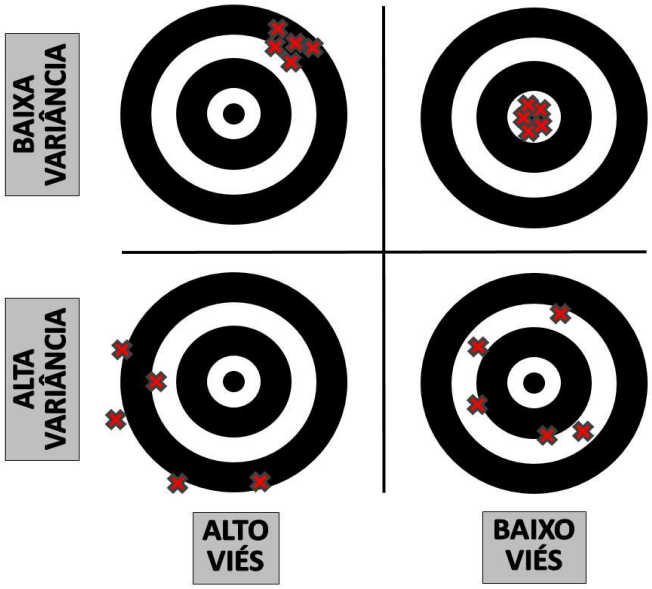
Outra característica importante da árvore de decisão é o conjunto de parâmetros que definem seu crescimento. Esses parâmetros podem ser relativos à estrutura da árvore, como profundidade máxima, número máximo de nós ou número mínimo de dados por nó. Também podem ser relacionados ao grau de impureza ou ganho de informação, como ganho de informação mínimo e redução mínima de impureza.

As definições desses parâmetros impactam no tamanho da árvore e, consequentemente, no seu grau de complexidade. Quanto maior a complexidade, maior é a

tendência de sobreajustamento do modelo aos dados de treino, o que pode dar origem ao problema de *overfitting*. Nesse cenário, o modelo se ajusta excessivamente aos dados de treino e falha nos demais, caso em que a variância das predições se torna alta.

Por outro lado, nos casos em que os parâmetros resultam em árvores pequenas, o modelo resultante pode ser muito simples. Nessa situação, haverá *features* importantes não consideradas pelo modelo, causando um problema de alto viés, ou generalização. Esse problema também é conhecido como *underfitting*. A Figura 5 exemplifica o problema de viés e variância e como esses dois fenômenos afetam as predições.

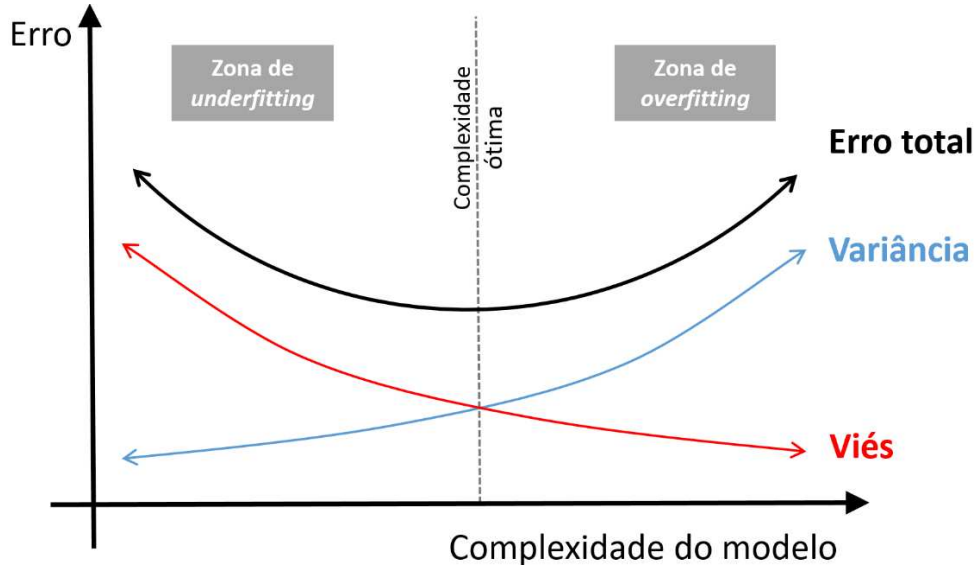
Figura 5 - Ilustração do efeito de viés e variância



Fonte: Elaborado pelo autor, baseado em Molnar (2025).

Nos modelos de *machine learning*, o dilema entre viés e variância obriga o analista a escolher por modelos com maior variância ou maior viés, uma vez que a diminuição de um importa o aumento do outro. Nessas situações, o ponto ótimo de complexidade ou tamanho do modelo se dá quando ocorre a minimização do erro total. A Figura 6 mostra os comportamentos do erro das predições, do viés e da variância em função da complexidade do modelo.

Figura 6 - Gráfico do dilema viés variância



Fonte: Elaborado pelo autor, a partir de Goodfellow, Bengio e Couville (2016).

As árvores de decisão têm como pontos positivos o fato de serem um modelo de fácil interpretação, permitindo a identificação das *features* mais importantes. Além disso, lida bem com *outliers* e dados faltantes. Por outro lado, o algoritmo pode apresentar tendência a *overfitting* e grande sensibilidade a mudanças nos dados de treino. Relações lineares simples também podem ser de difícil captura, pois os nós dividem os dados em retas horizontais ou verticais por padrão.

2.4.2 Métodos Ensemble

Métodos *ensemble* são combinações de modelos simples que, quando utilizados em conjunto, são capazes de reduzir o viés ou a variância, dependendo da arquitetura utilizada, melhorando a performance dos modelos individuais. Para o caso de métodos baseados em árvores, são usados agrupamentos de árvores de decisão para alcançar um resultado melhor.

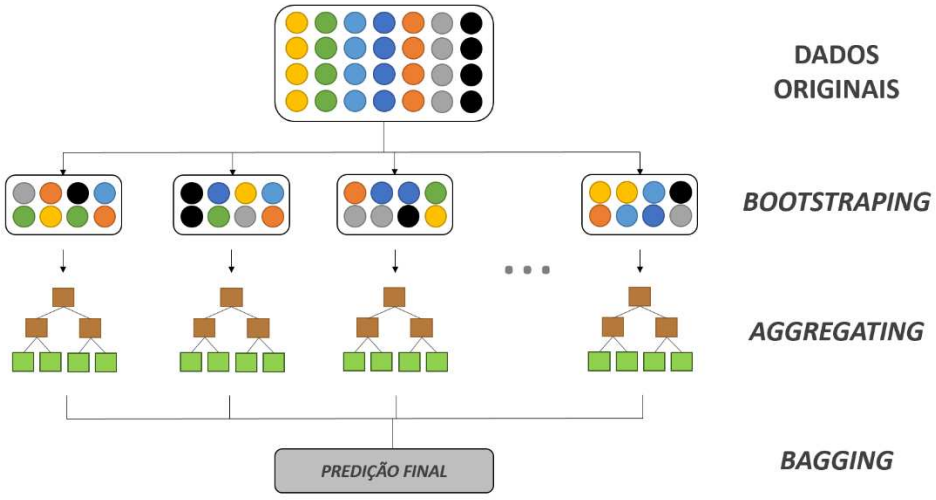
Iban (2022) resume as duas arquiteturas de modelos *ensemble*: 1) os chamados *bagging* (*bootstrap* + *aggregation*), que produzem diferentes preditores a partir de subconjuntos da amostra de treino selecionados aleatoriamente; e 2) os chamados *boosting*, que empregam modelos em sequência a partir da mesma amostra de treinamento, a fim de melhorar a performance final. O modelo *bagging* tem como um dos principais exemplos o algoritmo *Random Forest*, já para o modelo *boosting* podem ser citados os algoritmos *Gradient Boosting*, *Light Gradient Boosting* e *Extreme Gradient Boosting*.

2.4.2.1 Bagging

O algoritmo *Random Forest* foi desenvolvido por Breiman (2001), e seu objetivo é reduzir a variância ao combinar diferentes árvores para a predição final (*aggregation*), seja pela média, em problemas de regressão, ou pela maioria, em problemas de classificação; e controlar o viés a partir de regras diferentes daquelas usadas tradicionalmente para a geração de árvores de decisão. A primeira dessas regras é que uma árvore não é criada a partir de toda a amostra de treino, mas de uma parcela selecionada aleatoriamente (*bootstrap*). Além disso, os nós (raiz e de decisão) são formados a partir da melhor variável entre um subconjunto aleatório das variáveis disponíveis, e não necessariamente da melhor de todas.

Essa aleatoriedade imposta na criação das árvores faz com que elas sejam diferentes entre si e não se ajustem perfeitamente aos dados de treino, mesmo que cresçam até a profundidade máxima, de modo que a diversidade introduzida por meio da aleatorização evite o *overfitting*. Após gerada uma quantidade considerável de árvores, o resultado é agregado e combinado para uma predição final. O princípio da sabedoria das multidões ajuda a explicar o aumento da precisão ao se combinarem respostas independentes acerca de um problema. A Figura 7 mostra o funcionamento esquemático do algoritmo *Random Forest*.

Figura 7 - Esquema de funcionamento de modelos do tipo *bagging*



Fonte: Elaborado pelo autor, baseado em Harjeet-Blue (2025).

Na primeira etapa, denominada *bootstrapping*, são gerados subconjuntos de dados a partir da amostra de treino por meio da seleção aleatória com reposição. Em seguida, é construída uma árvore de decisão para cada subconjunto, sendo que o nó de divisão é escolhido com base na melhor *feature* entre um subconjunto aleatório das variáveis. Por fim, na etapa de *aggregation*, as predições individuais são reunidas para a predição final.

O *Random Forest* reduz o problema de *overfitting*, comum em algoritmos de árvores de decisão individuais, melhora a precisão, pode ser aplicado tanto a problemas de classificação como de regressão, lida bem com dados faltantes e não requer normalização dos dados. No entanto, apresenta alguns pontos negativos, como o custo computacional mais elevado, sobretudo nos casos com grandes conjuntos de dados e variáveis, e menor interpretabilidade dos resultados, em comparação com modelos mais simples, como regressão linear ou árvore de decisão simples.

2.4.2.2 Boosting

Os algoritmos de *boosting* seguem a mesma ideia de combinar predições a partir de um conjunto de preditores "fracos" para criar um "forte". A diferença conceitual entre eles e os do tipo *bagging* é que a criação desses preditores se dá sequencialmente, de modo dependente, e não em paralelo, de modo independente. Além disso, embora a predição final também seja a partir da combinação dos votos dos preditores fracos, nesses modelos podem ser atribuídos pesos para cada um dos votos, de modo que alguns são mais importantes do que outros, tendo mais representatividade na eleição do valor final.

Considerando o dilema de viés e variância das de árvores de decisão, os modelos de *boosting* buscam reduzir o viés, aumentando a precisão ao tratarem sequencialmente os resíduos das predições anteriores. Mesmo assim, a ideia é que a variância se mantenha controlada, evitando o *overfitting*. Isso é possível pela manutenção intencional de árvores rasas, ou de baixa profundidade, mantendo a simplicidade dos preditores individuais e evitando a fotografia dos dados de treino.

Freund e Schapire (1996) propuseram o algoritmo *Adaboost*, um dos mais básicos entre os do tipo *boosting*. Seu funcionamento envolve a criação de árvores de decisão com apenas um nível de profundidade, também denominada *stumps*, ou "tocos", cujas predições são avaliadas posteriormente com base na própria amostra de treino. A partir dessa avaliação, atribui-se uma importância com base na sua taxa de erro. Em seguida, ocorre um novo ciclo de criação de *stumps* e avaliação de performance, com a diferença de que os dados cujas predições foram incorretas na árvore anterior terão uma atenção maior na criação da árvore seguinte. O processo se dá continuamente, até atingir o número de árvores definido na criação do modelo, e a predição final é a combinação das predições dos *stumps* considerando seu respectivo peso.

O *Gradient Boosting*, por sua vez, proposto por Friedman (2001), se diferencia porque suas árvores tentam encontrar os resíduos da predição de um modelo anterior, e não o

valor da variável alvo propriamente dito. Seu funcionamento se inicia com uma predição igual para todos os dados, usualmente a média. Em seguida, são calculados os resíduos, com os quais será criada uma árvore de decisão, que tenta modelar os resíduos a partir das variáveis de entrada. A criação das árvores não é aleatorizada, mas determinística, por meio da minimização de uma função de perda, como erro quadrático médio, para problemas de regressão. Por isso, essas árvores devem ser pouco profundas, para evitar o sobreajustamento. Os resíduos previstos pela primeira árvore são multiplicados por uma espécie de redutor, denominado *learning rate*, ou taxa de aprendizado, e adicionados ao valor médio arbitrado inicialmente. Calculam-se, então, os novos resíduos obtidos desse somatório e se criam árvores, sucessivamente, até uma quantidade pré-definida. Por fim, a previsão final do modelo se dá pelo valor arbitrado inicialmente (média) adicionado do somatório das previsões de resíduos de todas as árvores.

Já o *XGBoost* (*eXtreme Gradient Boosting*) pode ser considerado uma evolução do *Gradient Boosting*. A principal diferença é a introdução de regularização na criação das árvores, a fim de evitar *overfitting*. Isso é feito por meio da utilização de parâmetros que penalizam a complexidade do modelo na função de perda. Além disso, Chen e Guestrin (2016), os criadores do algoritmo, explicam que o *XGBoost* permite redução da utilização de recursos computacionais, por meio de mudanças em relação ao acesso à memória cache, compressão de dados e técnicas de particionamento.

No contexto de avaliação imobiliária, esse algoritmo é capaz de tratar a dependência espacial dos resíduos, uma vez que nesses casos os erros são relacionados com as variáveis geográficas. Para isso, o modelo identifica a eventual existência dessa correlação e busca corrigi-la, tornando os resíduos aleatórios no espaço, como devem ser. Conforme Harrison (2020, p. 126), o algoritmo tenta capturar e tratar qualquer padrão nos erros, até que se tornem aparentemente aleatórios. Nos modelos tradicionais, o tratamento desse fenômeno, que quebra um dos pressupostos da regressão linear clássica, é complexo, assim como a predição para novos dados fora da amostra. Oliveira et al. (2024) demonstraram a maior capacidade do algoritmo *XGBoost* em corrigir a autocorrelação espacial dos resíduos de um modelo de avaliação em massa, quando comparado com o método de regressão linear espacial.

2.4.3 Aplicações de machine learning para avaliação em massa

Algoritmos de *machine learning* baseados em árvore têm sido aplicados com sucesso em avaliações em massa com propósitos tributários, de acordo com Zilli, Bastos e da Silva (2024). Entre os modelos disponíveis, os autores priorizaram os baseados em árvores para

mapear o estado da arte, por suas predições apresentarem maior precisão quando comparadas com outros métodos. Eles abordaram as principais técnicas utilizadas nesse contexto, desde as mais simples, como árvores de decisão, até os métodos avançados de modelos combinados, conhecidos como ensemble, incluindo *Random Forest*, *AdaBoost*, *CatBoost*, *Gradient Boosting*, *XGBoost* e *LightBoost*.

Como resultado, os autores identificaram o modelo *Random Forest* como sendo o mais estudado, seguido por *Gradient Boosting* e *XGBoost*. Além disso, o *Random Forest* apresentou a melhor performance em metade dos trabalhos analisados. Já em relação às métricas de desempenho, o RMSE foi a mais adotada pelos pesquisadores. Os autores destacaram que a pesquisa na área ainda é incipiente e sugeriram novos estudos futuros, como a avaliação dos modelos para diferentes regiões e tipologias e a exploração e o teste da eficácia de técnicas de *feature engineering* para a seleção de variáveis relevantes.

Antipov e Pokryshevskaya (2010) elaboraram modelos de avaliação em massa de apartamentos na cidade de São Petersburgo, na Rússia. A amostra era formada por 2.848 dados de transações de apartamentos. No estudo, o algoritmo de *Random Forest* teve a melhor performance, medida pelo MAPE (14,86%), entre vários outros, como regressão múltipla, KNN, redes neurais e outros.

Nas considerações finais desse trabalho, os autores destacaram a capacidade do modelo de lidar com dados faltantes, variáveis categóricas múltiplas e *outliers*. Além disso, observaram que todos os algoritmos testados apresentaram melhor desempenho ao prever o preço por metro quadrado em vez do preço total do imóvel. No que diz respeito à importância das variáveis explicativas, ou *features*, as duas principais foram locacionais: distrito e tempo até o centro da cidade por metrô.

Carranza et al. (2018) realizaram estudo de avaliação em massa de terrenos na cidade de Río Cuarto, na província de Córdoba, Argentina. Com amostra de 283 dados, os autores aplicaram o algoritmo *Random Forest* para predição de valores da terra, combinado com técnica de krigagem no tratamento dos resíduos. A média do erro percentual absoluto (MAPE) foi de 19%, enquanto a mediana desses erros foi de 13%, indicando um bom desempenho do modelo, segundo os pesquisadores. Entre os benefícios destacados da aplicação da técnica, foram destacadas a agilidade na atualização dos valores imobiliários e a melhoria na equidade do sistema tributário.

Oliveira (2020) utilizou uma amostra robusta com 8.209 dados de terrenos coletados de um período de cinco anos do Município de Fortaleza, no Brasil, para comparar os modelos *Random Forest* e *XGBoost*, tendo como referência o de regressão clássica. Por meio

de gráficos de dependência parcial, mostrou que os algoritmos foram capazes de capturar e descrever bem o comportamento da variável dependente (preço unitário) em relação às explicativas, como área do terreno e distância à via principal. Com os resíduos das predições, aplicou o teste de Wilcoxon para confirmar a hipótese de que os modelos são estatisticamente todos diferentes entre si.

Como resultado desse estudo, a performance dos algoritmos de aprendizagem de máquinas foi confirmada como superior à regressão clássica, tendo obtido MAPE de 24,40% e 28,64%, respectivamente, para *XGBoost* e *Random Forest*, contra 34,09% da regressão. O *XGBoost* também superou os demais modelos em métricas como COD, RMSE, MAE e apresentou o menor resíduo mediano.

Zilli e Bastos (2024) também compararam performances de modelos, tendo como amostra 1.572 dados de ofertas de apartamentos da região central de Florianópolis, Brasil, coletados via *webscrapping* e tratados posteriormente. As modelagens incluíram regressão linear múltipla, *Random Forest* e *Gradient Boosting*. Nestas duas últimas, foi utilizado o preço total linear como resposta, enquanto na regressão linear se utilizou a transformada em logaritmo natural como variável dependente.

Na comparação, o algoritmo *Gradient Boosting* obteve os melhores resultados em todas as métricas: RMSE, MAE, MAPE, COD, R^2, com predições até 30% mais precisas que as dos demais. Além disso, foi o único que alcançou PRD dentro da faixa estabelecida pelo IAAO (0,98-1,03), parâmetro importante para avaliações em massa com fins tributários, enquanto os outros dois modelos apresentaram tendência de subavaliar os imóveis mais caros. Os autores destacaram a capacidade dos algoritmos de aprendizagem de máquinas de capturar as complexidades do mercado imobiliário, como a dependência espacial. Por fim, sugeriram a aplicação do método em outras regiões e com outros algoritmos.

Oliveira et al. (2024) realizaram um estudo para comparar nove algoritmos de aprendizagem de máquinas para avaliação em massa com a abordagem estatística clássica. O objetivo era observar especialmente a interpretabilidade, considerando a finalidade tributária, e a capacidade de tratar a autocorrelação espacial dos resíduos. Para isso, dispuseram de uma amostra com 43.585 dados de apartamentos do Município de Fortaleza, entre ofertas e transações efetivadas, capturados dos anos de 2017 a 2021. Além das métricas de desempenho comumente aferidas, como MAPE, R^2 e RMSE, os pesquisadores também calcularam o índice de Moran I para medir a autocorrelação espacial.

Como resultado desse estudo, foram encontradas performances semelhantes entre os algoritmos de aprendizagem de máquinas, com uma leve superioridade do *XGBoost*. Nos

modelos clássicos, foi constatada presença de autocorrelação espacial e regressividade. Percebeu-se também que há uma dicotomia, uma espécie de *trade off*, entre melhorar a performance e fornecer interpretação do modelo. Isso porque os clássicos não foram capazes de resolver o problema da autocorrelação espacial, mas são de fácil interpretação, enquanto o algoritmo *XGBoost*, por exemplo, cujos resultados são mais difíceis de interpretar, sanou o problema. Por fim, como futuros trabalhos, os autores recomendaram a aplicação de técnicas de interpretação de modelos de Inteligência Artificial para simplificar a complexidade desses resultados.

2.5 Técnicas de interpretação

Segundo Das e Ras (2020), *Explainable Artificial Intelligence* (XAI) é o campo da Inteligência Artificial que promove uma série de ferramentas e técnicas que permitem uma interpretação intuitiva e compreensível das decisões dos algoritmos. Os autores dividem essas técnicas em dois escopos: global e local.

Na parte global, busca-se entender o modelo como um todo, hierarquizando as variáveis de entrada em termos de relevância para a predição final. Já localmente, as técnicas tentam traduzir como cada variável influenciou a predição de um único dado. A interpretação local é fundamental não apenas para a compreensão do resultado, mas também para identificação de eventuais erros e comportamentos inesperados. Além disso, pode revelar padrões ou relações de influência até então não documentados na literatura, que tenham sido detectados pelo algoritmo.

No contexto de avaliações em massa, de acordo com Iban (2022), os estudos de algoritmos de aprendizagem de máquinas baseados em árvores têm focado na performance dos algoritmos, e não tanto nas técnicas de interpretação dos resultados.

2.5.1 Permutation Feature Importance

Permutation Feature Importance (PFI) mede o aumento do erro na predição de um modelo, após os valores de uma variável serem embaralhados (Molnar, 2025). O embaralhamento dos valores das variáveis mais relevantes tende a causar erros maiores, o que se reflete no resultado do PFI. Dessa forma, é possível hierarquizar as *features* em função de sua relevância.

O procedimento, introduzido por Breiman (2001), consiste em treinar um modelo e calcular uma medida de performance, como o coeficiente de determinação R^2. Em seguida, com uma amostra separada, embaralham-se os valores de uma variável explicativa e são geradas predições com o modelo treinado. Com essas predições, calcula-se o novo R^2 e se determina a queda do desempenho em relação ao valor original. Repete-se esse processo (com diferentes embaralhamentos) para a mesma variável e se calcula a média das quedas de desempenho. O procedimento é aplicado a todas as variáveis explicativas. As variáveis mais importantes são aquelas cujas permutações provocaram as maiores perdas de desempenho.

Iban (2022) defende a utilização de PFI para a seleção das variáveis mais relevantes antes da criação de um modelo definitivo. Essa filtragem permite a criação de um modelo com menos *features*, as mais relevantes, de modo a reduzir o custo computacional e podendo até melhorar seu desempenho.

Molnar (2025) cita como principais benefícios o fato de a técnica propiciar uma visão global a respeito do comportamento do modelo e de não exigir retreinamento do modelo. Por outro lado, aponta a limitação de não ser possível conhecer o sentido do efeito da variável e a fragilidade para casos em que há correlação entre duas ou mais variáveis.

Essa fragilidade para variáveis correlacionadas pode ocorrer por dois motivos: i) não se conseguir isolar o efeito de apenas uma variável e, ao ser aplicada a técnica, o efeito da variável é subestimado, pois outras que são correlacionadas já estariam demonstrando seu efeito; ii) ao ter os valores de uma das variáveis correlacionadas permutados, são criadas situações que não seriam possíveis no mundo real.

2.5.2 Gráfico de dependência parcial

Friedman (2001), ao apresentar o algoritmo *Gradient Boosting Machine*, propôs também uma técnica para analisar a influência marginal de uma variável explicativa sobre a variável alvo denominada *partial dependence plot*, ou gráfico de dependência parcial. Segundo o autor, o gráfico do valor da variável alvo como função de seus argumentos fornece um resumo abrangente de sua dependência em relação aos valores conjuntos das variáveis de entrada.

Com o gráfico, é possível observar como a variável de interesse se comporta com a variação uma ou duas *features*. Além da relevância de cada *feature*, a técnica permite compreender a natureza da relação entre as variáveis, se linear, monotônica ou mais complexa (Molnar, 2025).

A metodologia para criação do gráfico é simples e intuitiva: cada ponto representa a média das predições quando se alteram, virtualmente, todos valores de determinada variável a um valor específico. Por exemplo, para construir o gráfico de dependência parcial do valor unitário de um terreno em função de sua área, a partir de um modelo já treinado, basta calcular a média das predições para uma amostra alterando todos os valores da área para, por exemplo, 100m², e plotar no gráfico. Em seguida, o processo é repetido para outros valores, como 200m², 300m² e assim por diante.

Oliveira (2020) analisou justamente essa relação (preço unitário do terreno em função de sua área) para o município de Fortaleza e observou três comportamentos distintos do efeito do acréscimo de área no valor unitário de terrenos: i) até 750m², o valor unitário decresce com o aumento da área, explicado pelo princípio da utilidade marginal decrescente; ii) a partir daí, até 10.000m², o valor unitário aumenta com o acréscimo de área, o que é explicado pela maior possibilidade de incorporação imobiliária, segundo o autor e iii) para terrenos maiores que 10.000m², o acréscimo de área volta a reduzir o valor unitário devido à utilidade marginal decrescente.

Molnar (2025) reconhece a praticidade da técnica, por ser de fácil implementação, interpretação e compreensão. No entanto, aponta limitações semelhantes às da PFI, sendo uma das principais a suposição de que as variáveis explicativas são independentes. Essa premissa pode ocasionar cenários que não existem na vida real, devido à interdependência das variáveis explicativas, na prática. Além disso, diz que efeitos diferentes podem ser omitidos, pelo fato de o gráfico mostrar a média dos efeitos marginais, o que poderia ser sanado com a plotagem das curvas individuais, e não da média.

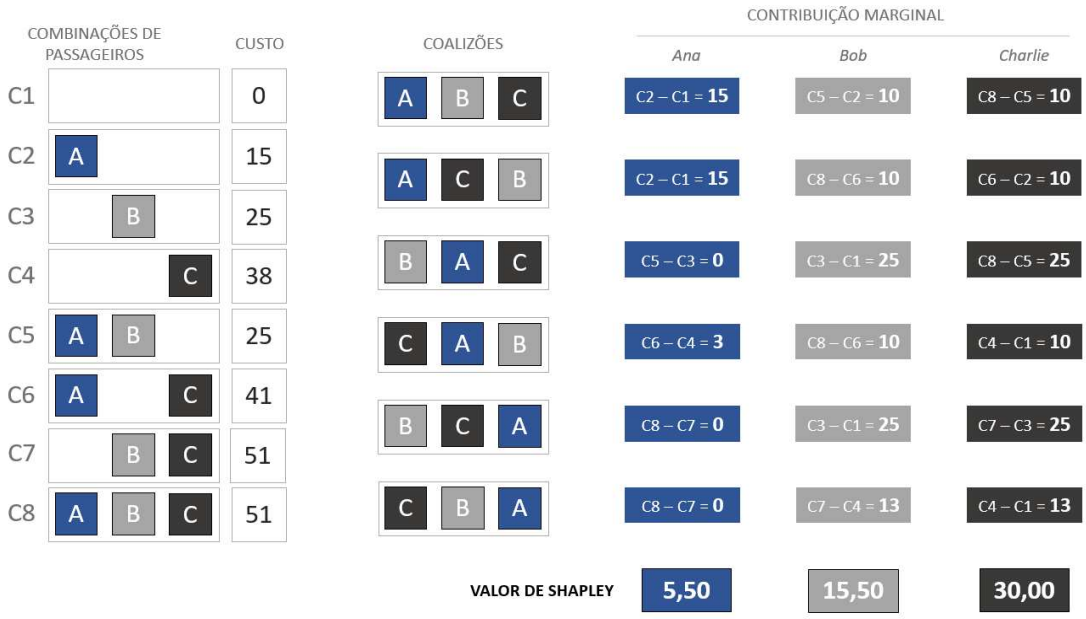
2.5.3 SHAP

Shapley (1952), no contexto da teoria dos jogos, fundamentou o que ficou conhecido como valor de Shapley: um método justo de dividir determinado valor entre os participantes de uma coalizão. Baseado nos axiomas da eficiência, simetria, nulidade e aditividade, o método reparte o ganho de um jogo cooperativo entre os participantes em função da contribuição marginal de cada um deles.

Ao explicar o valor de Shapley, Molnar (2022) utiliza um exemplo prático: a divisão do custo de uma corrida de táxi entre três pessoas. Nesse exemplo, assume-se que são conhecidos os custos das corridas para todas as possíveis combinações de passageiros, ou seja, o custo quando cada indivíduo viaja sozinho, quando dois viajam juntos e quando os três

compartilham o táxi. Para calcular o valor de Shapley de cada participante, considera-se sua contribuição marginal em todas as coalizões possíveis das quais poderia fazer parte. Essa contribuição é obtida subtraindo o custo da corrida sem o indivíduo do custo da corrida com o indivíduo. A média dessas contribuições para cada indivíduo é o seu valor de Shapley, e a soma desses valores de cada participante será igual ao custo total. A Figura 8 esquematiza o problema e calcula o valor de Shapley para os indivíduos A (Ana), B (Bob) e C (Charlie).

Figura 8 - Exemplo de cálculo de valor de Shapley.



Fonte: Elaborado pelo autor, baseado em Molnar (2025).

Na imagem, as possíveis combinações de passageiros estão à esquerda, C1 a C8, com seus respectivos custos. Há seis coalizões possíveis e, para cada uma delas, é calculada a contribuição marginal da inclusão do indivíduo de interesse, começando por Ana, com o valor final de 5,50. Ao repetir o processo para os indivíduos B e C, obtêm-se, respectivamente, 15,50 e 30,00. A soma dos três valores resulta em 51,00, o valor total da corrida. Matematicamente, o valor de Shapley é dado pela equação 6.

$$\Phi_j = \sum_{S \subseteq N\{j\}} \frac{|S|!\,(N - |S| - 1)!}{N!} [(v(S \cup \{j\}) - v(S)] \qquad (06)$$

em que:

N é o conjunto total de jogadores;

S é o subconjunto de jogadores que não inclui o jogador j;

$|S|$ é o número de jogadores em S;

$v(S)$ é o valor da função característica para a coalizão S, ou seja, o valor gerado por esse grupo;

$(v(S \cup \{j\}) - v(S)$ é a contribuição marginal do jogador j ao entrar na coalizão S;
$|S|!\,(N - |S| - 1)!$ é o peso da contribuição marginal, que representa a probabilidade de j se juntar à coalizão S em uma ordem aleatória de jogos.

Vale destacar que o último termo, o peso da contribuição marginal, é utilizado para considerar os casos em que há repetição dos participantes que antecedem o jogador j, em ordens diversas, e que, portanto, devem ser contadas em peso maior. No exemplo do táxi, o peso maior seria para representar a repetição das contribuições marginais 0 e 15, no caso do indivíduo A, que são formados pelas mesmas coalizões, mudando apenas a ordem dos participantes.

No contexto de *machine learning*, Lundberg e Lee (2017) introduziram o SHAP (*Shapley Additive exPlanations*), técnica que tem como objetivo atribuir para cada variável a contribuição marginal correspondente. Essa atribuição é feita individualmente para cada predição, ou seja, as *features* podem contribuir de modo diferente para cada nova observação. Conceitualmente, o cálculo é análogo ao do valor de Shapley. Na prática, existem diferentes métodos, sendo o *TreeSHAP* o mais adequado para algoritmos baseados em árvores.

O somatório dos valores SHAP explica a diferença entre a médias das previsões de um conjunto de dados e a previsão para uma observação específica. Para isso, uma vez que o modelo está treinado e as árvores estruturadas, de modo análogo ao valor de Shapley tradicional, calcula-se a contribuição marginal média da inclusão de uma variável na estimativa de uma predição. No método *TreeSHAP*, o caminho na árvore se dá de modo ponderado nos ramos em que a divisão se dá em função de uma *feature* que está ausente na coalizão, proporcionalmente à quantidade de dados no treinamento do modelo. Já quando a *feature* está presente na coalizão, o percurso na árvore se dá normalmente. Assim, é possível avaliar o impacto da inclusão de cada variável em diferentes coalizões e calcular a média dessas contribuições.

Lundberg e Lee (2017) afirmam que a técnica permite o aumento de credibilidade dos modelos, uma vez que explica em parte as suas considerações. Além disso, provê meios para melhorar seu funcionamento, bem como o entendimento a respeito do funcionamento do processo modelado.

Por outro lado, Molnar (2025), embora reconheça a grande contribuição e utilidade da técnica, cita a dificuldade de atribuição de valores SHAP ao lidar com variáveis correlacionadas e a criação de combinações impossíveis de variáveis, no caso do método *KernelSHAP*. Além disso, um ponto importante diz respeito à dificuldade da interpretação de variáveis que atuam em conjunto, como no caso das coordenadas geográficas (latitude e

longitude) que são importantes em diversos contextos e suas interpretações são necessariamente feitas em conjunto.

Em um modelo qualquer, os valores SHAP para cada *feature* são interpretados como a contribuição individual da respectiva variável para a previsão de uma observação em específico. Uma forma de visualização dessas contribuições é o gráfico de forças ou *force plot*. A Figura 9 mostra um exemplo desse gráfico aplicado à previsão de uma observação presente em um conjunto de dados de preços de casas na Califórnia, com valores em centenas de milhares de dólares.

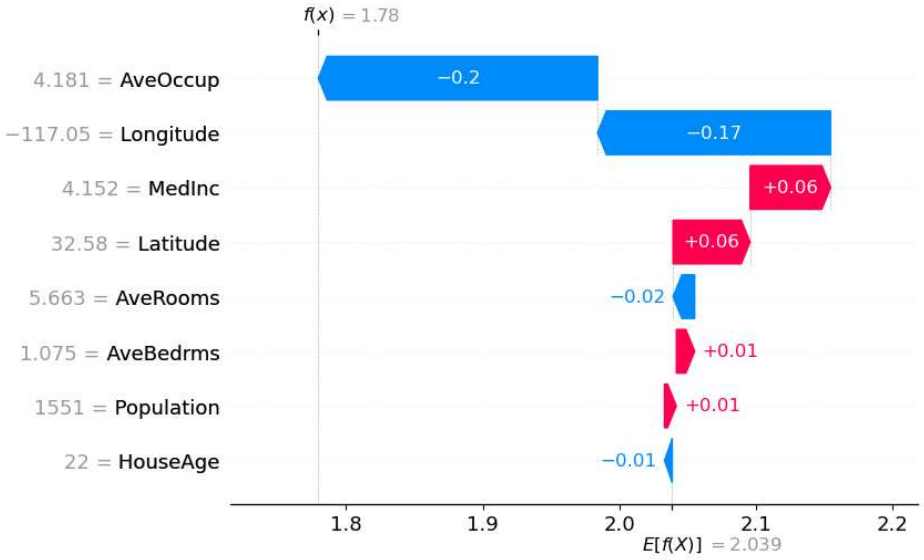
Figura 9 - Exemplo de *force plot*

Fonte: Elaborado pelo autor.

Na figura, são mostradas as contribuições de cada variável, de modo que o total explica a diferença entre o valor base, próximo a 2,04, para o valor da predição, 1,78. As *features* que contribuíram para o aumento do valor são mostradas em vermelho, enquanto as que contribuíram para a redução, em azul. Nesse caso, a renda média (*MedInc*), cujo valor foi de 4,152, foi a que mais contribuiu para o aumento da predição. Já do outro lado, a média de ocupantes por residência, 4,181 para esta observação, foi a variável que mais contribuiu para a redução da predição.

Essas mesmas informações podem ser visualizadas em um gráfico denominado cascata, ou *waterfall plot*, que tem como vantagem a hierarquização das variáveis em função da importância, para aquela observação. Ao mesmo tempo, o gráfico em cascata também exibe os valores dos atributos e as respectivas contribuições, conforme a Figura 10.

Figura 10 - Gráfico de valores SHAP de cascata



Fonte: Elaborado pelo autor.

Existem gráficos que agregam os valores SHAP de todas as observações e variáveis. Com isso, é possível realizar análises globais a respeito do comportamento de cada *feature* e do modelo como um todo, conforme mostrado nas figuras abaixo.

Figura 11 - Dependência parcial com SHAP

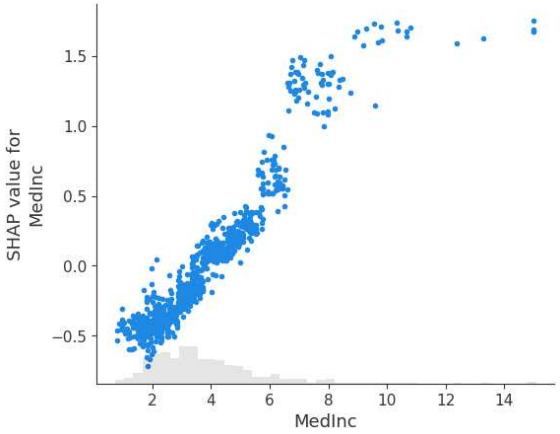


Figura 12 - *Summary plot*

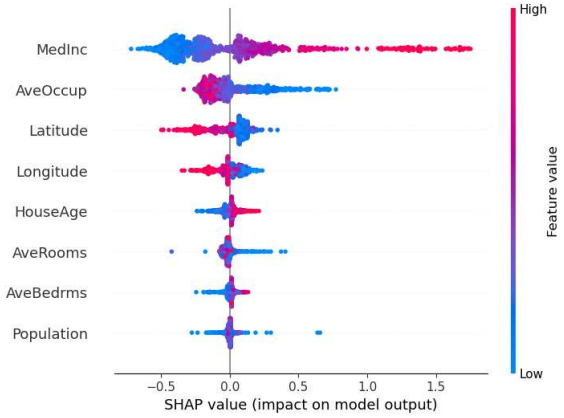


Figura 13 - *Heatmap plot*

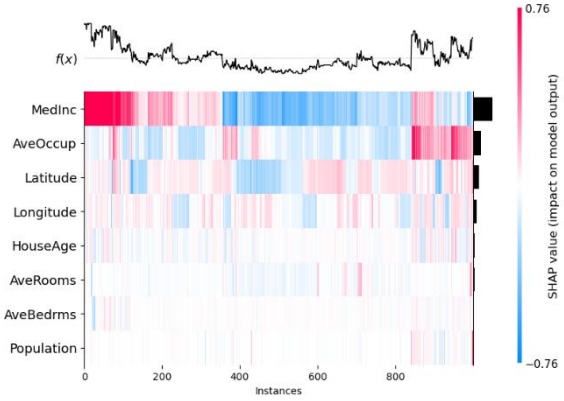
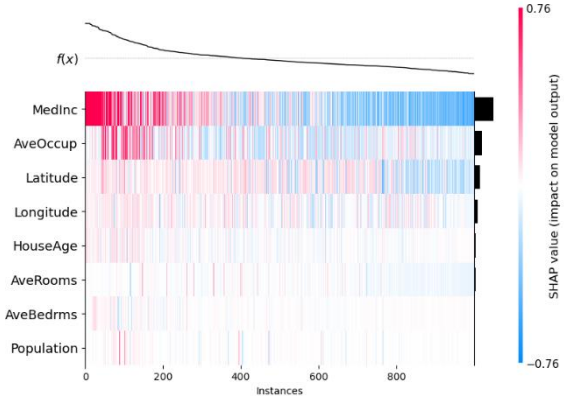


Figura 14 - *Heatmap plot ordenado*



Fonte: Elaborado pelo autor

A Figura 11 mostra um gráfico de dependência parcial, criado a partir dos valores SHAP. Também é exibido um histograma dos dados, em cinza. Nesse caso, pode-se observar uma relação positiva forte entre renda e preço. Já a Figura 12 exibe o tradicional *summary plot*, que lista as *features* em ordem de importância global e permite a análise do comportamento de cada variável. Para a variável *AveOccup*, por exemplo, os valores mais altos (em vermelho/rosa) influenciam negativamente na predição, pois seus valores SHAP são negativos, estando à esquerda da reta vertical que indica o zero, enquanto os valores mais baixos tendem a aumentar o valor da predição. Daí, pode-se interpretar que quanto maior a média de ocupantes, menor tende a ser o valor predito pelo modelo.

Já a Figura 13 e a Figura 14 exibem o chamado *heatmap graph*. Nesse gráfico, as variáveis também são listadas conforme sua importância, e as cores mostram os valores SHAP por cada observação. A princípio, os dados são ordenados por um algoritmo de clusterização hierárquica (Lundberg, 2018), como na Figura 13, mas podem ser reordenados, como na Figura 14, para a qual se utilizou da ordenação pelo valor da predição.

No contexto de avaliação em massa, essa técnica tem sido utilizada em estudos recentes. Iban (2022) discutiu os resultados obtidos com a técnica SHAP para um modelo de previsão de valor de mercado para apartamentos no distrito de Yenişehir, na Turquia, com o algoritmo *XGBoost*, escolhido após ser comparado com um modelo clássico de regressão linear e outros de *machine learning* (*Random Forest*, *LightGBM* e *Gradient Boosting*). Os modelos foram criados a partir de uma amostra com 1002 dados e 43 variáveis independentes. Com a técnica, o autor observou as relações complexas que o algoritmo foi capaz de capturar, o que resultou em uma performance superior à dos demais modelos. Destacou-se também a possibilidade de verificar a presença de viés em variáveis nos casos em que as predições são muito distantes do valor base.

Tarasov e Śliwiński (2024) elaboraram modelos com os algoritmos *Random Forest* e *XGBoost* para estimar os valores de residências em Varsóvia, na Polônia, com base em 55 variáveis independentes e 10.827 dados de transação, no ano de 2021. Com SHAP, além de analisarem predições pontuais, os pesquisadores identificaram a captura de uma relação não linear entre o preço e a idade da edificação, uma das variáveis mais relevantes globalmente. Foi observada uma redução do preço em imóveis de até 20 anos, seguida de aumento a partir dos 40 anos, o que foi explicado pela mudança de percepção ou apreço por estilos arquitetônicos mais antigos na cidade. Junto da idade, coordenadas geográficas e área foram as variáveis mais relevantes globalmente. Os autores concluíram que foi demonstrada a praticidade da aplicação de técnicas XAI no contexto de avaliação de mercado de imóveis e sugeriram prosseguimento dos estudos, com novas pesquisas em diferentes regiões e recortes temporais.

3. MATERIAL E MÉTODOS

3.1 Caracterização e análise exploratória dos dados

A base de dados desta pesquisa é composta por eventos de transações e ofertas de apartamentos, coletados nos anos de 2024 e 2025 e mantidos pela Secretaria Municipal das Finanças (Sefin), no município de Fortaleza. A amostra com a qual foram elaborados os modelos é composta por 6.660 dados. Os dados possuem, basicamente, duas origens: i) coleta de anúncio de imóveis da *internet*, placas de venda, tabelas de preços ou outros meios e ii) Declarações de Transmissão Imobiliária (DTI) de transações efetivamente realizadas.

Os dados de ofertas são coletados e mantidos pelo Observatório do Mercado Imobiliário (OMI) da Sefin. O OMI é responsável por monitorar o mercado de imóveis do município, com coletas contínuas e georreferenciadas de anúncios de imóveis de diferentes. Já as informações oriundas de DTI se devem ao fato de a Sefin ser o órgão responsável por recolher o ITBI e, por isso, todas as transações onerosas de imóveis passarem pelo processo de declaração formal como requisito para pagamento do imposto. A amostra com esses dados foi obtida por meio de um pedido formal ao órgão, por meio do Portal da Transparência, no Sistema Eletrônico do Serviço de Informação ao Cidadão.

Ressalta-se a opção pela inclusão de observações oriundas de anúncios de vendas (ofertas), e não apenas de transações efetivadas, mesmo que o objetivo dos modelos seja prever valores de mercado. Há duas razões principais para isso:

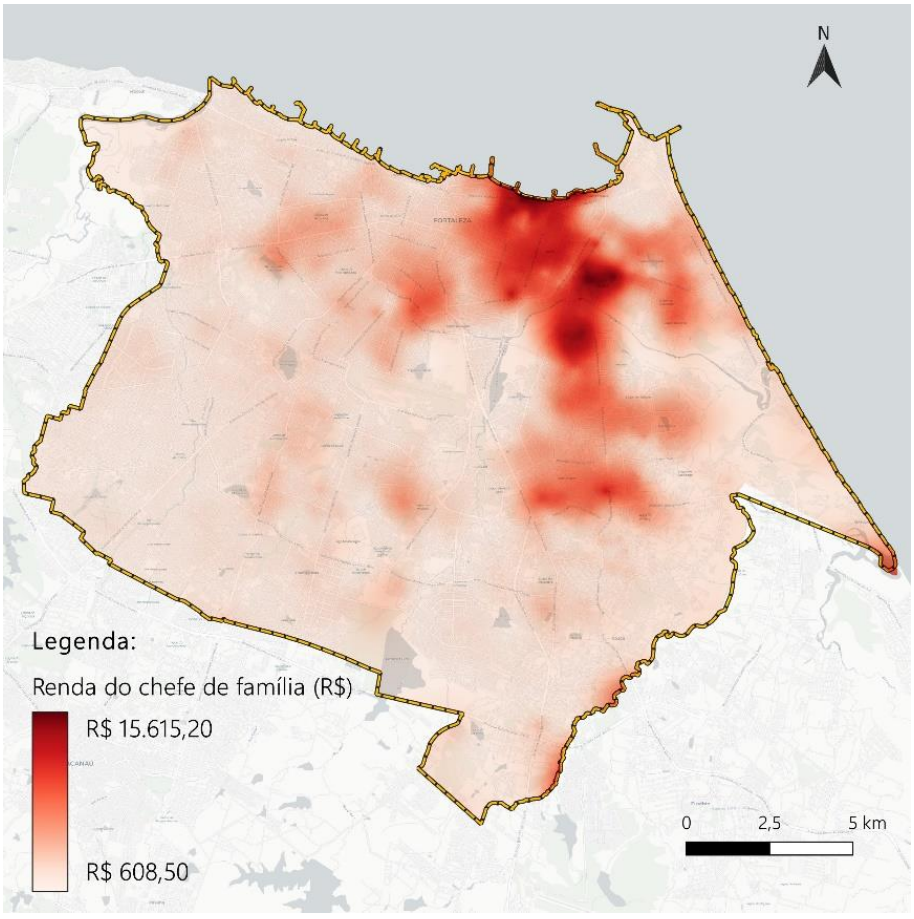
i) Conforme Çağdaş (2013), o valor de transações declarados por cidadãos para fins de cálculo de impostos são usualmente mantidos abaixo do real, a fim de reduzir o valor do tributo cobrado. Com isso, é importante para a Administração Tributária o monitoramento contínuo do mercado imobiliário, por meio de anúncios de vendas e, com isso, controlar o "fator de oferta" aceitável de diferença de preços entre anúncios e transações efetivadas, a partir do parâmetro do regressor;

ii) De acordo com Grover (2016), pode ser difícil montar uma amostra robusta o suficiente apenas com dados de transações em regiões onde o mercado imobiliário não seja aberto o suficiente. Nessa situação, em países com o Brasil, é bastante usual o emprego de ofertas na amostra para fins de avaliação imobiliária, inclusive em pesquisas acadêmicas (Carranza et al., 2018; Oliveira, 2020; Zilli e Bastos, 2024; Oliveira, 2024).

Devido à amplitude temporal das observações que se estende a um prazo superior a um ano, uma das variáveis empregadas é uma dicotômica, *a2025*, que indica se a observação é do ano de 2025 ou não, caso em que é de 2024. Com isso, espera-se que seja capturada a variação dos preços observados nos distintos anos.

Localizada no Ceará, Nordeste do Brasil, Fortaleza possui uma área de 312,35km², totalmente urbana, com uma população estimada de 2,57 milhões de pessoas, em 2024. Isso resulta em uma densidade populacional de 8.242 habitantes/km² (IBGE, 2025), sendo a capital mais densamente povoada do País. A despeito do destaque econômico entre as capitais do Nordeste (1ª) e até do Brasil (8ª), Fortaleza é uma cidade desigual economicamente. Espacialmente, essa desigualdade é observada com clareza, sendo a região ao norte e à leste do Centro aquela com maior renda. À medida que se distancia dessa área, a renda média *per capita* tende a diminuir, sendo as menores encontradas na periferia, próximo aos limites do Município. A Figura 15 mostra essa desigualdade espacial de renda a partir de dados do Censo de 2022.

Figura 15 - Mapa de renda do chefe de família, elaborado a partir dos dados do Censo do IBGE, de 2022



Fonte: Elaboração própria.

Essa desigualdade espacial evidencia a necessidade do georreferenciamento dos dados, uma vez que a valoração de um bem imóvel é altamente influenciada pelo espaço no qual está inserido, sobretudo para municípios nos quais os índices socioeconômicos são tão diversificados como os de Fortaleza.

De acordo com dados do Cadastro Imobiliário Municipal, Fortaleza possuía cerca de 855 mil imóveis cadastrados ao final de 2024, dos quais 261 mil são apartamentos, representando 30,51% do total. Em 2024, essa tipologia foi responsável por 59,3% das transações imobiliárias registradas na Sefin. O fato de os apartamentos terem maior representatividade nas transações indica uma alta demanda por imóveis dessa tipologia no mercado imobiliário atual. Essa demanda se traduz na arrecadação municipal de ITBI, cuja receita provém majoritariamente (53,7%) de transações de apartamentos.

Com dados oriundos de declarações de valores dos contribuintes ao realizarem transações, bem como de ofertas coletadas e mantidas pelo OMI da Sefin, a amostra abrange todo o território do Município e temporalmente se restringe ao período entre janeiro de 2024 e junho de 2025. A variável alvo dos modelos foi o preço unitário do imóvel, ou seja, o preço por metro quadrado de área privativa. As variáveis explicativas são mostradas na Tabela 1.

Tabela 1 - Descrição das variáveis

Variável	Descrição
renda	Representa a renda média do chefe de família, em R$, ajustada a uma superfície de tendência construída por krigagem, tendo como base os dados de renda média por setor censitário, divulgados pelo IBGE, por ocasião do Censo de 2022 (IBGE, 2025).
distpv	Distância, em metros, do centro do lote em que se localiza o imóvel ao polo de influência valorizante mais próximo.
distassp	Distância, em metros, do centro do lote em que se localiza o imóvel ao assentamento precário mais próximo.
distsh	Distância, em metros, do centro do lote em que se localiza o imóvel ao *shopping center* mais próximo.
distbm	Distância, em metros, do centro do lote em que se localiza o imóvel à Avenida Beira-Mar.
test	Representa, em metros, o comprimento da testada principal do lote em que se situa o dado observado.
iamaxeq	Representa o índice de aproveitamento máximo equivalente, segundo Lei de uso e ocupação do solo de Fortaleza (LC 236/2017).
dv	Variável que representa a densidade de verticalização, calculada a partir da concentração de condomínios verticais com elevador, obtida por interpolação *kernel* com função quadrática de raio 200m.
dorm	Variável discreta que representa a quantidade de dormitórios existentes no apartamento.
elev	Variável dicotômica que indica a presença ou não de elevador.
pisc	Variável dicotômica que indica a presença ou não de piscina.
pvtp	Variável discreta que representa a quantidade de pavimentos tipos do edifício a que pertence o apartamento observado.

vg	Variável discreta que representa a quantidade de vagas de garagens disponíveis à unidade do apartamento observado.
areapriv	Área privativa do imóvel, em metros quadrados.
andar	Variável discreta que representa o andar em que se localiza o apartamento observado. Para apartamentos em edifícios sem elevador, a ordem de contagem do *andar* foi invertida.
idade	Variável contínua que representa, em anos, a diferença de tempo entre a época da coleta do dado e a finalização da construção do edifício. Imóvel em construção têm *idade* negativa.
a2025	Variável dicotômica que indica se o dado foi coletado no ano de 2025 ou não, caso em que o dado foi coletado em 2024.
of	Variável dicotômica que indica se a origem do dado é oferta ou não, caso em que a origem é uma transação.
x	Representa a longitude do centroide do lote, em coordenadas UTM.
y	Representa a latitude do centroide do lote, em coordenadas UTM.
punit	Preço unitário do imóvel, em R$/m², obtido pela relação entre o preço dividido pela área privativa do imóvel (*areapriv*).

Fonte: Elaborado pelo autor

A estatística descritiva e os histogramas das variáveis utilizadas para elaboração do modelo estão dispostos, respectivamente, na Tabela 2 e no Apêndice A. Além disso, a Figura 16 mostra a localização dos dados. Inicialmente, a amostra era composta por mais observações, distribuídas entre 2.331 diferentes condomínios. No entanto, como alguns poucos edifícios concentravam uma quantidade desproporcional de dados, optou-se por limitar a no máximo cinco registros selecionados aleatoriamente por condomínio. Após esse ajuste, a amostra utilizada na elaboração dos modelos passou a contar com os 6.660 registros definitivos.

Tabela 2 – Estatística descritiva da amostra

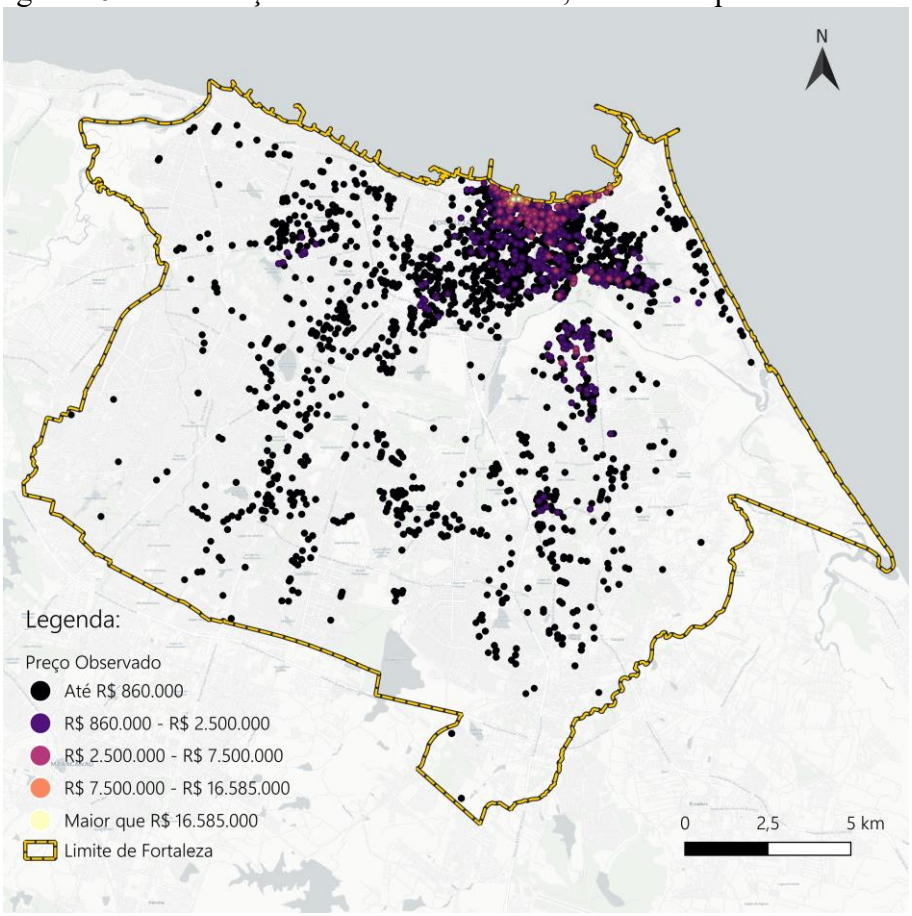
Variável	Média	Desvio padrão	0%	25%	50%	75%	100%
renda	7.477,01	3.837,99	1.240,51	3.714,68	7.587,39	10.866,94	15.595,48
distpv	201,21	175,08	0,00	72,58	175,86	282,49	2.074,36
distassp	278,25	249,67	0,00	89,20	219,82	413,77	2.537,17
distsh	1.294,78	977,33	0,00	568,09	1.065,20	1.788,40	7.876,12
distbm	4.306,67	3.774,84	0,36	1.307,60	2.906,23	6.530,72	18.022,82
test	58,96	42,10	5,50	31,85	47,00	75,00	384,00
iamaxeq	2,38	0,57	0,47	2,00	2,50	3,00	3,00
dv	0,29	0,28	0,00	0,03	0,18	0,51	1,00
dorm	2,78	0,68	1,00	2,00	3,00	3,00	6,00
elev	0,76	0,42	0,00	1,00	1,00	1,00	1,00
pisc	0,44	0,50	0,00	0,00	0,00	1,00	1,00
pvtp	12,89	7,82	1,00	5,00	12,00	20,00	51,00
vg	1,86	1,05	0,00	1,00	2,00	2,00	15,00
areapriv	106,17	76,46	30,00	60,24	83,69	125,62	1.217,54
andar	6,86	5,68	0,00	3,00	5,00	10,00	44,00

idade	19,50	13,43	-5,10	9,05	18,52	29,70	63,36
a2025	0,33	0,47	0,00	0,00	0,00	1,00	1,00
of	0,28	0,45	0,00	0,00	0,00	1,00	1,00
x	554.272,39	3.439,50	541.637,87	552.088,71	555.154,38	556.652,52	562.125,38
y	9.584.770,05	3.586,81	9.570.298,69	9.583.117,06	9.586.070,13	9.587.383,10	9.590.580,34
punit	5.932,47	3.435,66	1.082,70	3.462,00	4.995,60	7.457,33	38.051,21

Fonte: Elaborado pelo autor.

A Tabela 2, com a estatística descritiva da amostra, evidencia a heterogeneidade do Município ao exibir extremos tão destoantes da média, mesmo considerando os desvios padrão. Essa situação representa um desafio para a elaboração de modelos de predição de valores de mercado, sobretudo para o modelo clássico de regressão, em que os parâmetros tendem a capturar a influência média das variáveis.

Figura 16 - Localização dos dados utilizados, com os respectivos valores



Fonte: Elaboração própria.

Observa-se uma alta concentração de dados na região próxima à Praia de Iracema, nos bairros Meireles e Aldeota, onde se concentra a maior quantidade de edifícios residenciais. Essa também é a região mais valorizada, o que é observado pela cor dos dados, indicando o alto valor dos imóveis.

A matriz de correlação (Figura 17) indica, par a par, a correlação entre as variáveis do modelo. A análise desses valores é importante não só para identificar previamente as variáveis que tendem a ser mais importantes para explicar a variável explicada. Ela serve também para identificar correlação entre variáveis explicativas, o que pode, posteriormente, dificultar as interpretações dos resultados dos modelos.

Figura 17 - Matriz de correlação

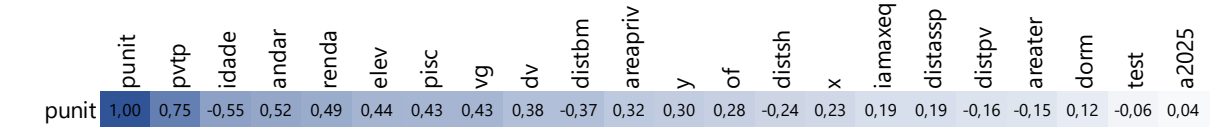
	test	areater	renda	dv	iamaxeq	distpv	distassp	distsh	distbm	elev	pisc	pvtp	dorm	vg	areapriv	andar	idade	a2025	of	x	y	punit
test	1,00	0,70	-0,31	-0,36	-0,34	0,08	-0,17	0,22	0,42	-0,21	0,24	-0,09	-0,11	-0,14	-0,16	-0,08	-0,20	0,02	-0,07	-0,16	-0,38	-0,06
areater	0,70	1,00	-0,45	-0,44	-0,43	0,09	-0,22	0,32	0,54	-0,31	0,19	-0,23	-0,23	-0,24	-0,25	-0,18	-0,17	0,02	-0,09	-0,27	-0,47	-0,15
renda	-0,31	-0,45	1,00	0,80	0,44	-0,17	0,41	-0,54	-0,72	0,52	0,01	0,56	0,39	0,53	0,51	0,38	0,09	-0,01	0,12	0,54	0,54	0,49
dv	-0,36	-0,44	0,80	1,00	0,37	-0,20	0,20	-0,53	-0,75	0,42	-0,10	0,45	0,31	0,44	0,48	0,31	0,17	-0,01	0,13	0,42	0,60	0,38
iamaxeq	-0,34	-0,43	0,44	0,37	1,00	-0,30	0,16	-0,45	-0,67	0,35	-0,12	0,33	0,27	0,23	0,24	0,23	0,24	-0,01	0,03	0,17	0,67	0,19
distpv	0,08	0,09	-0,17	-0,20	-0,30	1,00	0,03	0,26	0,29	-0,16	-0,01	-0,20	-0,10	-0,10	-0,13	-0,13	-0,09	0,00	-0,01	-0,02	-0,31	-0,16
distassp	-0,17	-0,22	0,41	0,20	0,16	0,03	1,00	-0,09	-0,18	0,19	0,02	0,19	0,16	0,20	0,17	0,16	-0,03	0,00	0,06	0,09	0,11	0,19
distsh	0,22	0,32	-0,54	-0,53	-0,45	0,26	-0,09	1,00	0,58	-0,38	0,05	-0,36	-0,30	-0,31	-0,30	-0,26	-0,14	0,00	-0,08	-0,32	-0,55	-0,24
distbm	0,42	0,54	-0,72	-0,75	-0,67	0,29	-0,18	0,58	1,00	-0,50	0,11	-0,51	-0,35	-0,42	-0,45	-0,36	-0,27	0,00	-0,10	-0,46	-0,92	-0,37
elev	-0,21	-0,31	0,52	0,42	0,35	-0,16	0,19	-0,38	-0,50	1,00	0,27	0,66	0,24	0,37	0,29	0,43	-0,25	0,01	0,11	0,33	0,42	0,44
pisc	0,24	0,19	0,01	-0,10	-0,12	-0,01	0,02	0,05	0,11	0,27	1,00	0,36	0,01	0,19	0,04	0,20	-0,58	-0,01	0,08	0,01	-0,11	0,43
pvtp	-0,09	-0,23	0,56	0,45	0,33	-0,20	0,19	-0,36	-0,51	0,66	0,36	1,00	0,26	0,51	0,39	0,63	-0,43	0,00	0,16	0,30	0,43	0,75
dorm	-0,11	-0,23	0,39	0,31	0,27	-0,10	0,16	-0,30	-0,35	0,24	0,01	0,26	1,00	0,62	0,67	0,19	0,21	-0,03	0,10	0,20	0,29	0,12
vg	-0,14	-0,24	0,53	0,44	0,23	-0,10	0,20	-0,31	-0,42	0,37	0,19	0,51	0,62	1,00	0,78	0,36	-0,06	-0,02	0,13	0,29	0,32	0,43
areapriv	-0,16	-0,25	0,51	0,48	0,24	-0,13	0,17	-0,30	-0,45	0,29	0,04	0,39	0,67	0,78	1,00	0,28	0,13	-0,02	0,18	0,27	0,37	0,32
andar	-0,08	-0,18	0,38	0,31	0,23	-0,13	0,16	-0,26	-0,36	0,43	0,20	0,63	0,19	0,36	0,28	1,00	-0,23	-0,02	0,09	0,20	0,30	0,52
idade	-0,20	-0,17	0,09	0,17	0,24	-0,09	-0,03	-0,14	-0,27	-0,25	-0,58	-0,43	0,21	-0,06	0,13	-0,23	1,00	0,02	-0,13	0,04	0,27	-0,55
a2025	0,02	0,02	-0,01	-0,01	-0,01	0,00	0,00	0,00	0,00	0,01	-0,01	0,00	-0,03	-0,02	-0,02	-0,02	0,02	1,00	0,09	-0,01	0,00	0,04
of	-0,07	-0,09	0,12	0,13	0,03	-0,01	0,06	-0,08	-0,10	0,11	0,08	0,16	0,10	0,13	0,18	0,09	-0,13	0,09	1,00	0,08	0,07	0,28
x	-0,16	-0,27	0,54	0,42	0,17	-0,02	0,09	-0,32	-0,46	0,33	0,01	0,30	0,20	0,29	0,27	0,20	0,04	-0,01	0,08	1,00	0,17	0,23
y	-0,38	-0,47	0,54	0,60	0,67	-0,31	0,11	-0,55	-0,92	0,42	-0,11	0,43	0,29	0,32	0,37	0,30	0,27	0,00	0,07	0,17	1,00	0,30
punit	-0,06	-0,15	0,49	0,38	0,19	-0,16	0,19	-0,24	-0,37	0,44	0,43	0,75	0,12	0,43	0,32	0,52	-0,55	0,04	0,28	0,23	0,30	1,00

Fonte: Elaborado pelo autor

Mesmo antes da elaboração de qualquer modelo, a análise da matriz de correlação fornece indícios a respeito de quais são as *features* mais relevantes para a explicação da variável alvo. A análise da última linha, que é mostrada com filtro e ordenação na Figura 18, permite

identificar que as variáveis *pvtp*, *idade*, *andar* e *renda* são as que possuem maior correlação com *punit*.

Figura 18 - Correlações com a variável alvo



	punit	pvtp	idade	andar	renda	elev	pisc	vg	dv	distbm	areapriv	y	of	distsh	x	iamaxeq	distassp	distpv	areater	dorm	test	a2025
punit	1,00	0,75	-0,55	0,52	0,49	0,44	0,43	0,43	0,38	-0,37	0,32	0,30	0,28	-0,24	0,23	0,19	0,19	-0,16	-0,15	0,12	-0,06	0,04

Fonte: Elaborado pelo autor

Além disso, as variáveis correlacionadas entre si são facilmente identificadas: *distbm* e *y*, o que é natural, pois a direção da Av. Beira-Mar é leste-oeste e se situa no norte da cidade, de modo que altos valores de *y* tendem a ser mais próximos da avenida, ou seja, têm menor distância. Uma outra correlação observada é entre *renda* e *dv*, o que também é, intuitivamente, compreensível, haja vista que em regiões mais ricas da cidade há maior concentração de edifícios verticais.

Essa análise é importante porque, no contexto dos modelos clássicos, a multicolinearidade pode afetar a sensibilidade dos parâmetros do modelo ou dificultar suas interpretações. Nesses casos, pode-se esperar o aumento do erro padrão, tornando-as estatisticamente não significativa, caso em que se deve avaliar a manutenção ou não da variável no modelo. Além disso, os resultados de PFI e SHAP podem ser influenciados pela correlação entre as variáveis, tornando mais difícil isolar o efeito de uma das variáveis da correlação.

3.2 Modelo de regressão linear

O modelo de regressão linear foi elaborado com *Python*, utilizando o pacote *statsmodels*. Além da forma natural, foi permitida transformação em logaritmo natural para as variáveis independentes. A opção pela restrição das duas transformações (linear e log) se dá pela facilidade da interpretação dos parâmetros. Além disso, para a variável explicada, foram testados modelos considerando o preço total e o preço unitário, ou seja, o preço dividido pela área privativa do imóvel. Para o modelo final, foi escolhido aquele que resultasse no maior coeficiente de determinação. Também foi analisado se os sinais dos parâmetros respeitavam o comportamento natural de mercado para cada uma das variáveis, bem como a significância de cada regressor.

Para tornar mais justa a comparação com outros modelos, para os quais a divisão da amostra entre treino e teste é importante, o desempenho do modelo foi avaliado em uma amostra de teste igual à dos demais, que representa 20% do total das observações.

Por fim, após a elaboração do modelo, foram calculadas as métricas de desempenho. Vale destacar que, no caso do modelo elaborado com preço unitário, o cálculo é feito comparando os valores preditos e observados dos preços totais, ou seja, após a multiplicação do preço unitário pela área privativa. Também foi aplicada a técnica PFI para a análise de importância global das variáveis, para que essas importâncias fossem comparadas, posteriormente, com as dos outros modelos de aprendizagem de máquinas.

3.3 Random Forest e XGBoost

Para a elaboração desses modelos, todas as variáveis, tanto a alvo como as explicativas, foram mantidas na forma linear, sem transformação. A amostra foi dividida duas, sendo uma parte para treino e a outra para teste, na proporção de 80% e 20%, respectivamente. A variável a ser explicada foi o preço unitário, ou seja, o preço do imóvel dividido por sua área privativa.

Os modelos foram elaborados em *Python*, com a biblioteca *scikit-learn*. Para a escolha dos parâmetros a serem utilizados na criação das árvores de decisão de ambos os algoritmos, foram testadas diferentes combinações. Para cada combinação de parâmetros, foi feita a avaliação por meio de validação cruzada (*cross validation*), com divisão da amostra em três partes, sendo cada uma dessas três partes utilizada como teste para verificação do desempenho. Cada combinação é avaliada pela média do desempenho na predição da amostra de teste. Para isso, foi utilizada a função *GridSearchCV*, uma ferramenta para seleção de modelos da biblioteca *scikit-learn*. Os parâmetros selecionados são mostrados abaixo.

- *Random Forest*: *max_depth*: *None*, *max_features*: *sqrt*, *min_samples_leaf*: 2, *min_samples_split*: 2, *n_estimators*: 300.
- *XGBoost*: *learning_rate*: 0,05, *max_depth*: 8, *colsample_bytree*: 0,8, *n_estimators*: 240, *subsample*: 0,8.

Com o modelo treinado e as métricas calculadas, foram verificadas as possibilidades de melhoria de desempenho com a retirada das variáveis menos importantes para a explicação dos valores. Finalmente, os modelos foram treinados, e as métricas para comparação de desempenho foram calculados.

4. RESULTADOS E DISCUSSÃO

4.1 Performance

O modelo clássico de regressão linear, que serve como *baseline* para avaliação de performance dos demais algoritmos, tem seus parâmetros, com as respectivas significâncias estatísticas, mostrados na Tabela 3.

Tabela 3 - Resultados do modelo de regressão linear

	coef	*Erro padrão*	*t*	*p-value*
Intercepto	3,816e+05	1,010e+05	3,762	0,0%
elev	-1.336,6473	74,212	-18,011	0,0%
dv	-1.487,7230	157,304	-9,458	0,0%
renda	0,2502	0,013	19,251	0,0%
distpv	-0,7383	0,137	-5,403	0,0%
ln_distassp	46,7659	15,875	2,946	0,3%
ln_distsh	-101,4184	26,173	-3,875	0,0%
ln_distbm	-1050,8910	30,789	-34,164	0,0%
pisc	836,8354	56,725	14,752	0,0%
pvtp	123,3023	5,727	21,531	0,0%
ln_vg	664,6122	123,683	5,374	0,0%
ln_areapriv	-600,7553	81,304	-7,389	0,0%
andar	51,2799	4,981	10,295	0,0%
idade	-105,2907	2,802	-37,577	0,0%
2025	330,5176	47,227	6,998	0,0%
of	1.058,7478	51,381	20,606	0,0%
x	-0,0521	0,008	-6,314	0,0%
y	-0,0352	0,010	-3,368	0,1%
Notas		R^2: 0,785		
Variável dependente: *punt_priv*		R^2 ajustado: 0,784		
N: 5.328		Estatística F: 1140		

Fonte: Elaborado pelo autor

No modelo final, as variáveis *elev* e *dv* apresentaram sinais contrários aos esperados inicialmente. Esse resultado pode ser explicado pela alta correlação que ambas mantêm com outras variáveis independentes. A variável *elev*, que indica a existência de elevador, apresenta correlação relativamente elevada com *pvtp*, que representa a quantidade de pavimentos tipo do edifício. Essa correlação é natural, pois edifícios mais altos tendem a possuir elevador. Assim, o sinal negativo de *elev* pode ser interpretado como um efeito colinear, e não como causa da redução de valor. Como o efeito da verticalização já é captado por *pvtp*, a variável *elev* passa a apenas ajustar a predição condicionalmente à altura do prédio, como se o modelo atribuísse à variável uma compensação devido à sobreposição de informação entre as duas variáveis.

A mesma situação ocorre entre *dv*, que indica a densidade de verticalização da região, e *renda* e pode explicar o sinal contrário do parâmetro de *dv*. No mais, todas as outras

variáveis do modelo final apresentaram comportamentos coerentes com o que se espera, de acordo com a literatura. Destaca-se o sinal negativo da variável *areapriv*, o que se explica pela lei da utilidade marginal decrescente. Uma vez que a variável de interesse é o valor unitário, ou seja, o preço de 1m² de área privativa, esse valor tende a ser menos valorizado à medida que mais unidades de área privativa são "consumidas".

O modelo de regressão permite a interpretação de seus parâmetros nas formas linear e logarítmica com facilidade. No caso lin-log, como em *areapriv*, o parâmetro indica que a variação percentual de 1% na área privativa está associada a uma variação média de -6,01 unidades no preço unitário, mantendo todas as demais variáveis constantes. Já *pvtp* está na forma linear, então seu coeficiente indica que o aumento de uma unidade de pavimento tipo gera um aumento de cerca de R$ 123,30 no preço unitário, *ceteris paribus*. Apesar disso, não é possível indicar quais variáveis são mais relevantes para a formação dos preços apenas com a análise dos coeficientes, pois as variáveis estão em escalas diferentes.

Uma vez definido o modelo clássico, foram treinados os outros dois, com os algoritmos *Random Forest* e *XGBoost*. Um deles foi escolhido para discussão mais aprofundada, a partir das métricas de desempenho, as quais são mostradas na Tabela 4. Os modelos foram avaliados em uma amostra de teste que contém 20% das observações.

Embora os modelos tenham tido como objetivo prever o valor unitário, a análise das métricas de desempenho se deu com o preço "cheio" dos imóveis, ou seja, após multiplicar o preço unitário predito pela área privativa. Essa opção permite observar eventuais vieses relacionados ao preço total, como subavaliação de imóveis mais caros e superavaliação de imóveis mais baratos, o que é medido pelo PRD.

Tabela 4 – Métricas de performance

	R^2	MAPE	RMSE	COD	PRD
Clássico	92,52%	24,59%	R$ 297.720,58	23,78%	1,0190
RF	95,91%	14,34%	R$ 220.149,86	14,03%	1,0141
XGBoost	96,40%	13,81%	R$ 206.601,09	13,62%	1,0128

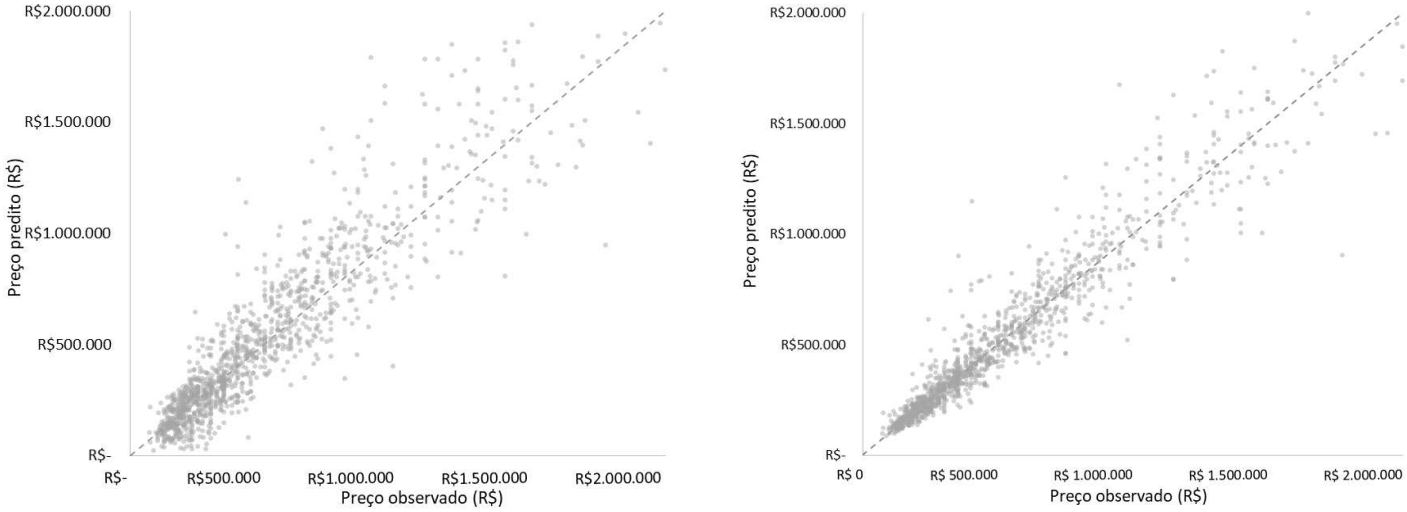
Fonte: Elaborado pelo autor

O modelo treinado com o algoritmo *XGBoost* se mostrou superior aos demais em todas as métricas adotadas para avaliação de performance. Além de possuir maior poder de predição, com maior coeficiente de determinação, apresenta menores erros (MAPE e RMSE) e menor dispersão (COD). Além disso, também obteve o melhor desempenho no que diz respeito à equidade vertical, medida pelo PRD, métrica muito importante para avaliação em massa,

sobretudo para fins tributários. Nesse quesito, o modelo clássico apresentou o pior desempenho, mas ainda dentro da faixa definida pela IAAO, entre 0,98 e 1,03.

A Figura 19 mostra os gráficos de valores preditos (eixo y) *vs* valores observados (eixo x), limitados os eixos a R$ 2 milhões. A partir dos gráficos de dispersão, torna-se evidente a menor dispersão das predições do modelo baseado em árvores, comparado ao modelo de regressão.

Figura 19 - Dispersão: valores preditos (eixo y) *vs* valores observados (eixo x).
Modelo MQO (à esquerda) e *XGBoost* (à direita).

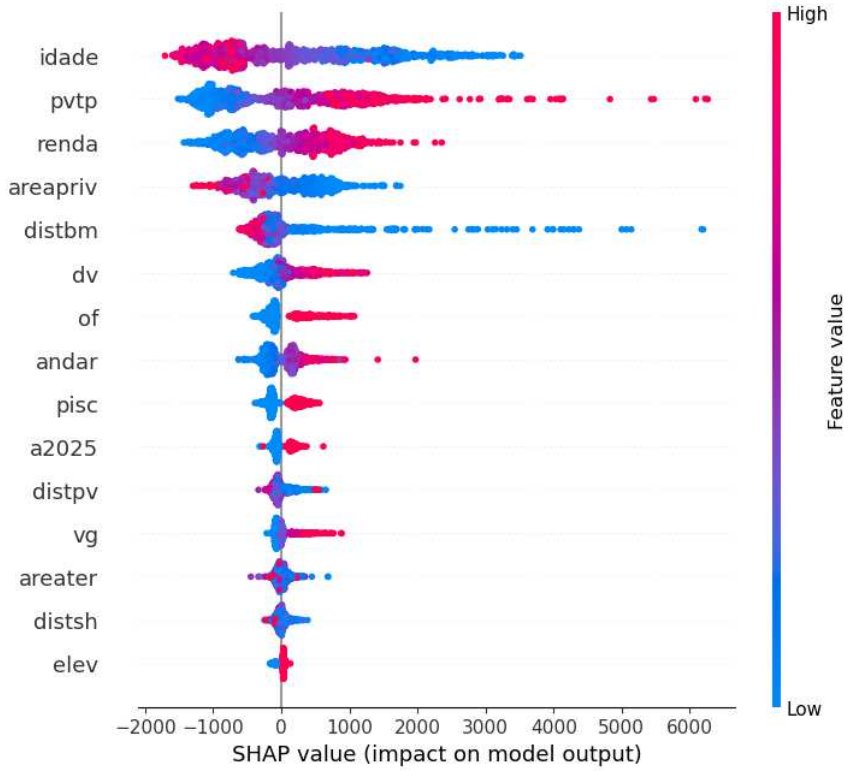


Fonte: Elaborado pelo autor.

4.2 Análise global com SHAP

Inicialmente, tratando da avaliação global do modelo *XGBoost*, a análise do *summary plot* permite entender quais as variáveis foram mais relevantes, bem como qual é a relação da variável de interesse com cada variável explicativa. Esse gráfico reúne e plota os valores SHAP de todas as observações para todas as *features*. A Figura 20 mostra o gráfico para o modelo elaborado.

Figura 20 - *Summary plot* (*XGBoost*)



Fonte: Elaborado pelo autor

A primeira informação possível de se extrair é a ordem de relevância das *features* para a predição dos valores de mercado. As cinco principais variáveis na formação do modelo foram: *idade*, *pvtp*, *renda*, *arapriv* e *distbm*. Em uma ordem diferente, são também as mais relevantes, pela técnica de PFI para o modelo clássico.

Além disso, também é possível observar a relação de cada *feature* com a variável alvo. Como uma analogia com o modelo de regressão, seria como analisar o sinal de cada parâmetro da equação. Para isso, deve-se observar quais cores estão com valores SHAP positivo e negativo. No caso em tela, cores mais próximas do azul significam valores baixos para a *feature*, e cores mais próximas do vermelho indicam valores altos, conforme a legenda à direita do gráfico.

Para *idade*, por exemplo, observam-se pontos azuis posicionados mais à direita, com valores SHAP positivo, indicando que idades baixas (imóveis novos) contribuem positivamente na formação da predição, ou seja, tendem a aumentar o valor de mercado. Por outro lado, valores altos de idade, ou seja, imóveis mais velhos, possuem valores SHAP negativos, reduzindo o valor da predição. Essa relação é equivalente ao sinal negativo da variável no modelo clássico, conforme os resultados mostrados na Tabela 3, e é a relação esperada para o fenômeno estudado.

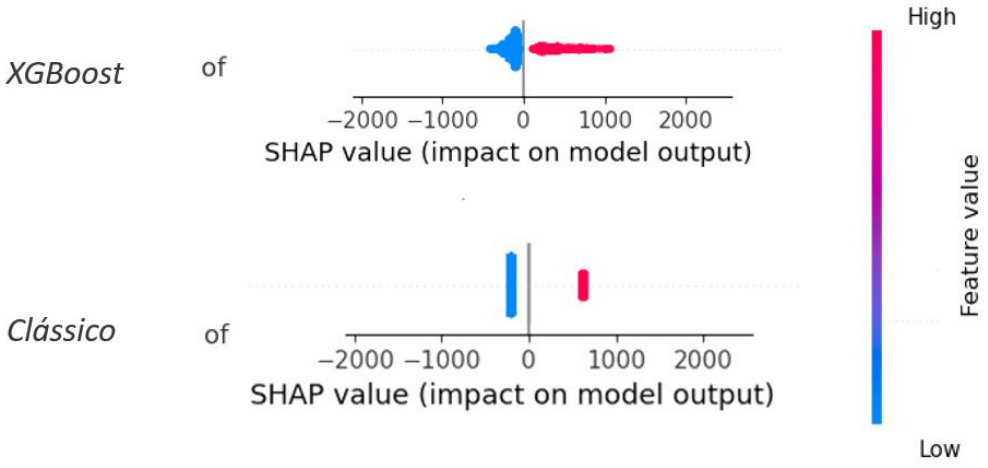
Já para *renda*, por exemplo, a relação é positiva com a variável alvo. Observações com valores altos para a variável (em vermelho) têm valores SHAP mais altos, posicionados mais à direita no gráfico. Já as observações com valores baixos de renda (em azul) têm valores SHAP menores. Essa mesma relação também foi encontrada no modelo clássico, pelo sinal positivo do parâmetro para a variável e também é a relação esperada, de acordo com a literatura.

Um comportamento interessante do modelo *XGBoost* é o fato de que duas observações com o mesmo valor para determinada variável podem sofrer efeitos diversos, em função dos demais atributos da observação. Por outro lado, no modelo clássico, valores iguais para determinada variável produzem sempre os mesmos efeitos na predição. Essa diferença pode ser observada pelo *summary plot*. Por exemplo, para variáveis dicotômicas, em que há dois valores possíveis, 0 ou 1, o modelo criado com *XGBoost* produz diversos efeitos diferentes para diferentes observações, o que pode ser observado também no *summary plot* pela diversidade de valores SHAP. Já no modelo clássico, a mesma variável produz efeitos semelhantes.

Esse fenômeno pode ser observado para a variável *of*, que indica se o dado observado tem origem em uma oferta ou não. É natural que exista uma relação positiva do valor predito com essa variável, porque, ao anunciar a venda de um imóvel, o ofertante tende a apresentar um valor um pouco superior ao valor de mercado, diferença essa que, geralmente, é reduzida durante negociações, até que a transação se efetive por um preço menor do que o anunciado. Esse comportamento foi observado em ambos os modelos: sinal positivo para o coeficiente, no clássico (Tabela 3) e relação positiva, com valores preenchidos com 1, em vermelho, tendo valores SHAP positivos (Figura 20). No modelo clássico, obteve-se um parâmetro de +1.058,75, indicando que as ofertas de imóveis possuem valores unitários superiores aos de transação em R$ 1.058,75/m², em média.

Ocorre que, uma vez elaborado o modelo clássico, esse mesmo efeito será aplicado para todos os dados a serem preditos, enquanto que, no *XGBoost*, esse efeito de oferta é diferente para cada dado, em função de seus outros atributos. Esse comportamento de ambos os modelos é percebido pela comparação do *summary plot* da variável *of* de ambos os modelos, como mostrado na Figura 21.

Figura 21 - *SHAP values* de ambos os modelos para a variável *of*



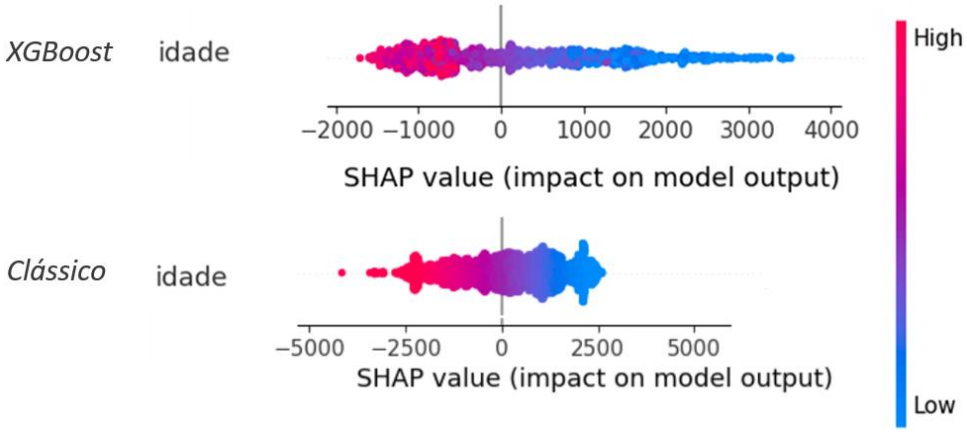
Fonte: Elaborado pelo autor

A figura mostra os valores SHAP para a variável oferta em ambos os modelos. Com os gráficos, a explicação da diferença de comportamento se torna bem mais fácil: enquanto no modelo clássico os efeitos são sempre os mesmos, haja vista os mesmos valores SHAP para dados com mesma origem, no *XGBoost*, a intensidade dos efeitos é distinta para cada observação, mesmo que todos os dados oriundos de oferta tenham efeitos positivos e os de transação tenham efeitos negativos.

Esse mesmo fenômeno também ocorre com variáveis contínuas. Enquanto no modelo estimado com regressão linear os efeitos de uma variável são estritamente proporcionais ao seu valor — produzindo sempre os mesmos efeitos para entradas iguais — no *XGBoost* isso não acontece.

Observando o *summary plot* do atributo idade no modelo clássico (Figura 22), nota-se que os valores SHAP são perfeitamente escaláveis com o valor do atributo (a cor, ou o valor do atributo, varia de modo uniforme com o valor SHAP, o efeito), de modo que idades iguais resultam sempre no mesmo impacto sobre a predição. No *XGBoost*, entretanto, há observações com a mesma idade e com efeitos distintos sobre a previsão.

Figura 22 – SHAP *values* de ambos os modelos para a variável *idade*



Fonte: Elaborado pelo autor

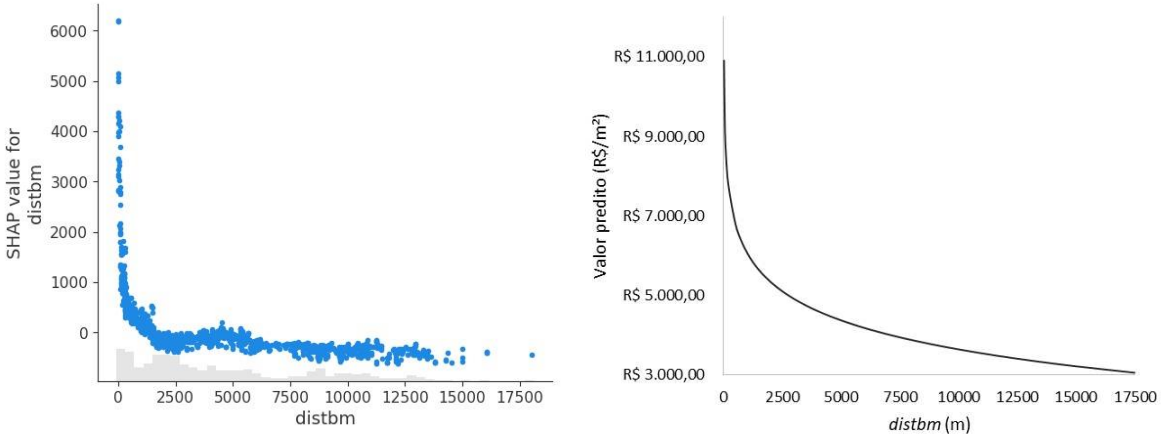
Essa diversidade de efeitos para entradas idênticas pode ser explicada pelo fato de que variáveis como *idade* não influenciam os imóveis da mesma forma, em diferentes contextos. Por exemplo, pode-se supor que imóveis em regiões de maior renda tendem a perder valor mais rapidamente com o aumento da *idade*, dado que há maior oferta de imóveis novos. Já em regiões de menor renda, onde a oferta de imóveis novos é mais restrita, o envelhecimento impacta o valor de maneira mais lenta. Essa relação será analisada adiante de modo mais detalhado.

Modelos baseados em árvore conseguem capturar essas relações complexas porque o efeito de uma variável não é aprendido de forma isolada: ele depende do "caminho" formado pelas divisões anteriores na árvore. Assim, o modelo ajusta o impacto de cada variável de maneira específica para cada cenário. Os valores SHAP, por sua vez, permitem visualizar e analisar essas variações de efeito entre diferentes contextos.

Para aprofundar o estudo da relação das *features* com a variável alvo, foram analisados os gráficos de dependência parcial, ou *partial dependence plot*. Nesse gráfico, os valores SHAP de uma variável e de todas as observações são plotados em um gráfico de dispersão contra a própria variável. Com isso, pôde-se verificar se o modelo foi capaz de capturar o comportamento esperado do fenômeno.

Esse mesmo gráfico pode ser produzido para os modelos tradicionais de regressão, variando os valores para a variável de interesse e mantendo as demais constantes, em um valor mediano, por exemplo. A comparação dos gráficos permite uma melhor compreensão a respeito do funcionamento de ambos os modelos. A Figura 23 mostra esses gráficos para a variável *distbm*.

Figura 23 - Gráficos de dependência parcial para *distbm XGboost* (à esquerda) e clássico (à direita)

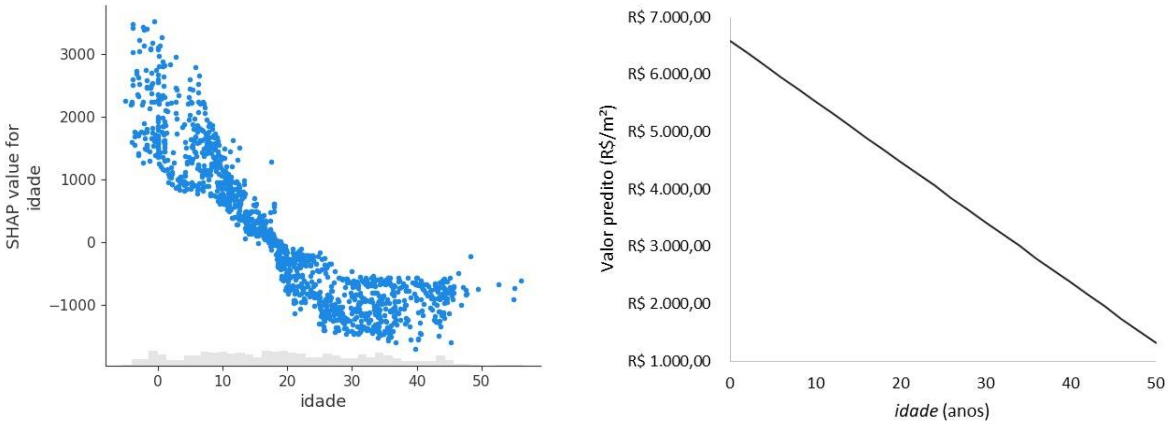


Fonte: Elaborado pelo autor.

Para ambos os modelos, há uma valorização acentuada para os imóveis localizados na região imediatamente próxima à Av. Beira Mar. Para o modelo baseado em árvores, o efeito positivo de estar nessa região se dá até pouco mais de 1.000m da avenida, a partir do que há certa neutralidade em relação à variação dessa distância, ou seja, imóveis situados a mais de 2,5km da Av. Beira Mar não sofrem tanta influência da avenida. O modelo clássico capturou essa relevância por meio da transformação da variável em logaritmo natural, de modo a retratar esse efeito nas distâncias mais imediatamente próximas à avenida. Mesmo assim, apresenta uma curva mais suave, com efeito relevante da variável para distâncias maiores. Em relação à magnitude dos valores, os modelos parecem concordar: há uma variação na ordem de R$ 6000/m² entre a maior e a menor influência da variável.

Já a Figura 24 mostra os mesmos gráficos para a variável *idade*.

Figura 24 - Gráficos de dependência parcial para *idade*.

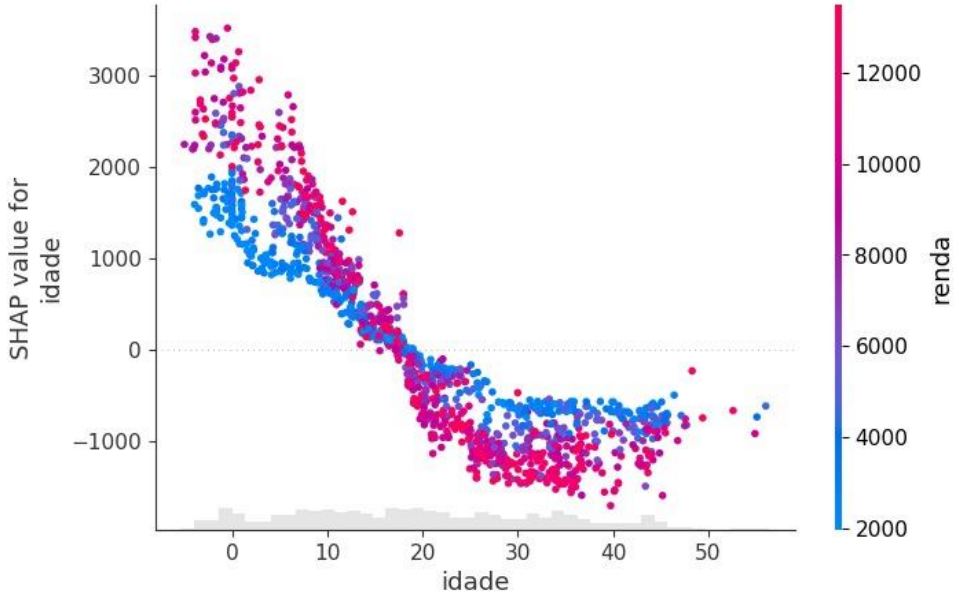


Fonte: Elaborado pelo autor

Nesse caso, foi observado comportamento linear para o modelo clássico, visto que a variável não foi transformada, enquanto no *XGBoost* o gráfico indica um comportamento não linear. Para o *XGBoost*, imóveis novos sofrem um efeito positivo na ordem de até R$ 3.000/m², e os velhos um efeito negativo de até R$ 1.000/m², aproximadamente, uma variação de R$ 4.000/m², valor similar à variação observada no gráfico do modelo clássico. Nesse caso, a principal diferença entre os modelos se dá pela captura de efeitos mais complexos no caso do *XGBoost*, o que é claramente observado no gráfico.

Nesse contexto, uma funcionalidade muito interessante da biblioteca SHAP é a possibilidade de, dentro do *partial dependence plot*, introduzir a informação do valor de outra variável, a partir do que se podem analisar relações mais complexas entre as *features*. A Figura 25 mostra o mesmo *pdp* da *idade*, mas acrescenta cores nos pontos, que, no caso, representam a variável *renda*. Os pontos em azul são aqueles localizados em regiões de menor renda, enquanto os de cor de rosa estão em regiões de renda mais alta. Essa relação foi discutida anteriormente, com a suposição de que a *idade* poderia afetar de modo diferente em função de outros atributos do imóvel, como a *renda* da região em que ele se localiza.

Figura 25 - Gráfico de dependência parcial da *idade*, relacionado com *renda*



Fonte: Elaborado pelo autor

A partir do gráfico, pode-se concluir que a *idade* afeta o valor dos imóveis mais intensamente, para mais ou para menos, nas regiões de alta *renda*. Isso porque os pontos em rosa apresentam os valores SHAP mais extremos, tanto positivos (para imóveis novos) como negativos (para imóveis mais velhos). Já os pontos em azul, mesmo que também apresentem influência da *idade*, têm valores SHAP menores, em módulo, ou seja, mais próximos de zero.

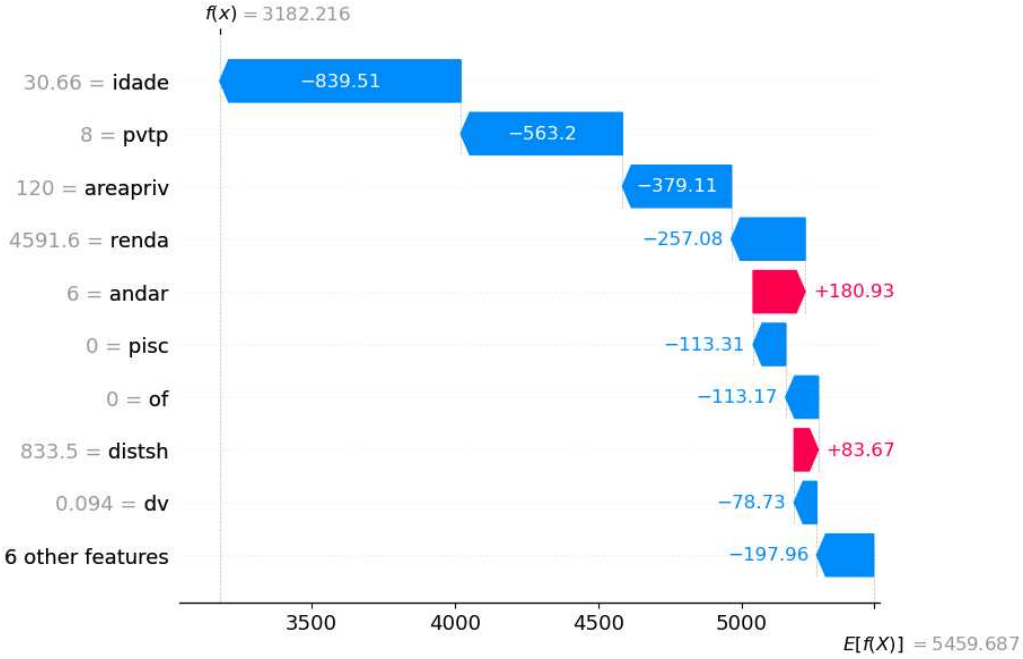
No geral, os resultados obtidos convergem com as conclusões de Oliveira (2020), que também observou a superioridade do algoritmo *XGBoost* para fins de avaliação em massa de terrenos, em relação ao *Random Forest* e ao modelo clássico de Regressão Linear Múltipla. Ambas as pesquisas permitiram observar a capacidade do modelo de capturar relações não lineares e interações complexas do preço com as variáveis independentes.

As análises globais dos modelos em ambos os estudos focaram em hierarquizar as *features* por importância e entender o comportamento os gráficos de dependência parciais. Contudo, os trabalhos se diferenciam pela forma como esses resultados foram obtidos, com o foco exclusivo na aplicação do SHAP neste trabalho. Além disso, a diferença mais significativa reside na exploração da análise local dos resultados, também com SHAP. Enquanto Oliveira (2020) focou na validação global do modelo para a proposição de uma Planta de Valores Genéricos para terrenos, a presente pesquisa avança ao demonstrar a possibilidade de explicar o valor predito por modelos baseados em árvore de cada imóvel individualmente, utilizando o recurso da análise local do SHAP.

4.3 Análise local com SHAP

No contexto da análise local dos resultados, caso em que se analisam os valores SHAP de todas as variáveis para uma observação específica, a utilização da técnica permite a interpretação do resultado mesmo por aqueles que não têm conhecimento a respeito do algoritmo que deu origem ao modelo. A Figura 26 mostra o *waterfall plot* de um dado da amostra de teste do modelo *XGBoost*.

Figura 26 - Análise local com *waterfall plot* – GI952431



Fonte: Elaborado pelo autor

O gráfico é construído como um caminho, ou cascata, que explica a diferença entre uma medida de tendência central da amostra de treino, chamado valor base, (R$ 5.459,68/m², nesse caso, localizado na parte inferior à direita do gráfico) e a predição específica para o dado (R$ 3.182,21/m², no canto superior esquerdo). No eixo y, a variáveis são ordenadas em função da relevância de suas contribuições, positivas ou negativas. No eixo, também são apresentados os valores da observação para cada um desses atributos. As barras dizem respeito à contribuição de cada variável, ou valor SHAP, sendo azul a contribuição negativa, e vermelha a positiva.

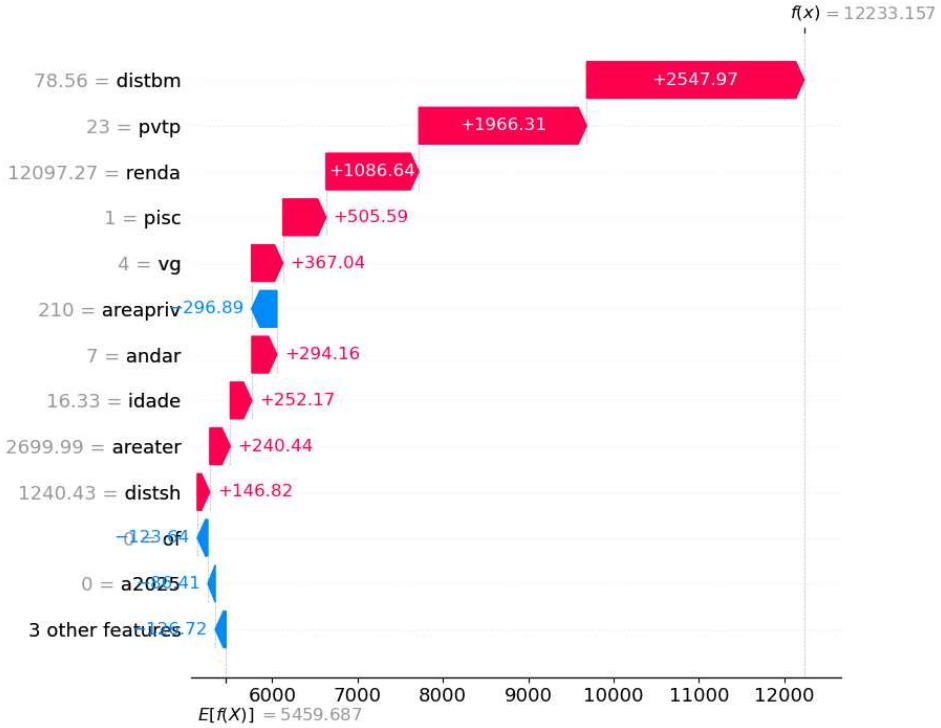
A observação em questão, GI952431, trata de uma transação efetivamente realizada, cujo valor observado foi de R$ 3.500/m², ou R$ 420.000,00, considerando o preço total. As predições foram de R$ 3.182,21/m² para o *XGBoost* (-9,07%, comparando com valor observado) e R$ 2.027,47 para o modelo clássico (-42,07%). As três variáveis mais relevantes (*idade*, *pvtp* e *areapriv*) contribuíram para a redução do valor da predição do dado em relação ao valor base. A natureza das contribuições dessas *features* é esperada, haja vista tratar-se de um imóvel relativamente velho (mais de 30 anos), baixo (oito pavimentos tipo) e de área privativa grande (120m²). Isso porque a análise global do modelo mostrou que o preço unitário tende a diminuir quanto maior a idade, menor o número de pavimentos e maior a área. Esta última se explica pela lei econômica da utilidade marginal decrescente.

Vale ressaltar que, para o modelo baseado em árvores, atributos com o mesmo valor podem apresentar contribuições diferentes, em função das outras características da observação. Esse fato torna mais importante a análise local das contribuições, quando esses algoritmos são

empregados. Isso porque, como discutido na interpretação da análise global, o valor da *feature* é analisado no contexto em que o dado está inserido, relacionando-o com as demais variáveis, e não de modo isolado. Um exemplo disso é o fato de que *distbm*, uma das variáveis mais relevantes globalmente, não afeta significativamente o valor do dado. Por outro lado, o mesmo *waterfall plot* para o modelo de regressão indica que essa foi a segunda variável mais relevante para a redução do valor, o que pode ser questionado, devido ao fato de que, de acordo com os gráficos de dependência parcial, a partir de certa distância, essa *feature* perde importância.

A Figura 27 mostra o *waterfall plot* para outro dado da amostra de teste, cujo valor observado foi de R$ 2.680.000,00 ou R$ 12.761,90/m². Trata-se da transação efetivamente realizada de um apartamento de grande área privativa (210m²), padrão relativamente alto e próximo à Av. Beira Mar.

Figura 27 - Análise local com *waterfall plot* - GI947836



Fonte: Elaborado pelo autor

Esse dado é um bom exemplo de como algoritmos baseados em árvore podem alcançar predições altas com mais facilidade, em contraposição ao modelo clássico. Especialmente para avaliação em massa com fins tributários, essa característica é essencial para garantir justiça fiscal. O modelo com *XGBoost* estimou o valor em R$ 2.568.962,97 (-4,14%), enquanto a estimativa com o modelo tradicional foi de R$ 2.414.743,21 (-9,90%).

Destacou-se na composição da estimativa de valor o fato de o imóvel estar a poucos metros da Av. Beira Mar, o que ressalta a grande importância da região para o mercado

imobiliário local e do emprego de variáveis locacionais para o estudo do valor de imóveis. Em seguida, o fato de o edifício possuir 23 pavimentos tipos, o máximo permitido de altura até pouco tempo atrás, no Município estudado, bem como a alta renda média da região, valoriza o imóvel de acordo com o gráfico.

Ao comparar os resultados obtidos com a literatura recente, são corroborados os achados de Iban (2022), que, após comparar performances de algoritmos baseados em árvore com o modelo clássico, explorou global e localmente as predições do algoritmo *XGBoost*. Assim como neste trabalho, Iban (2022) demonstra a capacidade do SHAP em evidenciar a natureza do impacto (positivo/negativo) das variáveis independentes, globalmente. Também adentra na interpretação local, discutindo as principais contribuições na formação de valor de mercado em observações específicas. As diferenças principais entre os trabalhos dizem respeito à tipologia, localidade e época do estudo. Essa diferença de escopo é importante para validação dos resultados de uma técnica como SHAP.

Além desse, os resultados da pesquisa também se assemelham aos da de Tarasov e Dessoulavy-Śliwiński (2024). Os autores também compararam performances dos algoritmos *Random Forest* e *XGBoost* para predição de residências, na Polônia, tendo este obtido melhor desempenho. Ademais, também analisaram o modelo, local e globalmente, com SHAP. Entre as relações capturadas pelo modelo que se puderam observar, destaca-se a da idade, cujo efeito no preço é negativo até cerca de 20 anos e, após 40 anos, se torna positivo. Para explicar esse comportamento, levantou-se a possibilidade de haver uma mudança gradual na percepção ou apreciação da arquitetura da cidade ao longo dos anos. Já neste trabalho, para os apartamentos em Fortaleza, o efeito da idade se mostrou negativo até cerca de 30 anos, após o que se manteve estável. Esses achados reforçam que a formação de valor de imóveis tem características distintas para diferentes mercados, o que foi possível de se concluir com SHAP.

No contexto público, a análise local com SHAP se mostra como uma alternativa viável para conferir mais transparência para agentes interessados na predição de valores com emprego de algoritmos baseados em árvore. No caso de avaliação em massa para fins tributários, especialmente no caso em que os imóveis de particulares são avaliados por órgãos públicos, a técnica é uma alternativa para a explicação da composição do valor final. Uma forma simples para isso seria informar aos interessados quais as variáveis foram as mais relevantes, em ordem, e qual a natureza do impacto de cada uma.

Por fim, a análise local também permite a identificação de dados inconsistentes, com potencial para melhoria dos modelos, bem como fornece *insights* sobre relações entre variáveis até então não exploradas. Deve-se destacar, contudo, que os valores SHAP são

baseadas nas contribuições que explicam o modelo, e não necessariamente a realidade. Isso significa que um modelo ruim pode gerar explicações coerentes com seus próprios padrões internos, mesmo que estes estejam baseados em vieses, dados de baixa qualidade ou relações espúrias. Portanto, as interpretações fornecidas pelo SHAP devem ser sempre analisadas em conjunto com a avaliação da performance preditiva do modelo e com conhecimento de domínio, a fim de evitar conclusões equivocadas.

5. CONCLUSÃO

Esta pesquisa teve como objetivo aplicar técnicas de interpretação de modelos de *machine learning* baseados em árvores, no contexto de avaliação em massa de apartamentos no município de Fortaleza/CE. A partir da comparação entre a técnica clássica e outras modernas, buscou-se identificar aquela que resultasse em melhor desempenho e, sobretudo, explorar o potencial das ferramentas *XAI* para interpretação dos resultados do modelo *XGBoost*, que apresentou melhor desempenho, com destaque para a técnica SHAP.

A análise global com SHAP permitiu identificar as *features* mais relevantes para o modelo, bem como entender as relações capturadas entre cada variável explicativa com a variável alvo. A técnica se mostrou capaz de apresentar de modo simples como o modelo capturou essas relações, e como elas são diferentes daquelas consideradas para o modelo clássico de regressão linear. Isso permite o melhor entendimento do modelo pelos técnicos, a compreensão de padrões do mercado imobiliário local e a identificação de eventuais erros, seja na elaboração do modelo ou na composição da amostra.

Uma das relações analisadas foi entre as variáveis *idade* e *punit*. Enquanto o modelo clássico considera o efeito linear entre elas, a técnica SHAP mostra que o *XGBoost* capturou uma relação não linear, em que o efeito da depreciação tendeu a se estabilizar após decorridos cerca de 30 anos. Além disso, foi observado o fato de que a *idade* atua de modo diverso para imóveis em função da *renda* da região em que ele se localiza: para aqueles situados em regiões de alta renda, a *idade* tem efeitos mais intensos, tanto positivos como negativos. Ou seja, nessas regiões, imóveis novos são ainda mais valorizados, e imóveis antigos são ainda mais desvalorizados do que em regiões de baixa renda. Levantou-se a hipótese de que esse fenômeno poderia ser explicado pelo fato de que, em regiões de alta *renda*, há grande oferta e demanda de imóveis novos, de modo que o aumento da *idade* reduz significativamente a demanda e, consequentemente, o valor dos imóveis mais velhos.

Além disso, a análise global dos modelos reforçou um conhecimento já consolidado na literatura de avaliação de imóveis: a importância de variáveis locacionais para a predição de valor.

Já na análise local, foram discutidos os efeitos das variáveis para duas observações da amostra. Com o *waterfall plot*, a interpretação dos resultados se mostrou intuitiva e de simples entendimento, sendo considerada uma forma simples para exposição dos resultados para agentes interessados leigos, no contexto de avaliação em massa para órgãos públicos, especialmente da Administração Tributária. Para isso, uma forma simples e prática de

comunicação desses resultados seria a informação de como cada uma das variáveis mais relevantes afetou a predição de determinada observação. A implementação de uma ferramenta XAI representa um avanço na busca por transparência na gestão pública e na introdução de técnicas mais modernas para avaliação de imóveis no contexto fiscal.

Para trabalhos futuros, sugere-se a aplicação dessas e de outras técnicas de interpretação de modelos de avaliação em massa em outros contextos de espaço e tempo, bem como da exploração de outras relações mais complexas entre as variáveis.

REFERÊNCIAS

ANTIPOV, E.; POKRYSHEVSKAYA, E. Mass appraisal of residential apartments: an application of Random Forest for valuation and a CART-based approach for model diagnostics. **Munich Personal RePEc Archive**, n. 27645, 2010. Disponível em: https://mpra.ub.uni-muenchen.de/27645/. Acesso em: 15 de abr. de 2025.

BACEN. **Taxa de juros básicas - Histórico**. Banco Central do Brasil. Disponível em: https://www.bcb.gov.br/controleinflacao/historicotaxasjuros. Acesso em: 9 out. 2025.

BANDEIRA, S. R.V. **Regressão espacial e avaliação de terrenos: um estudo de caso para a Cidade de Fortaleza/CE**. 2019. Dissertação (Mestrado em Economia do Setor Público) – Universidade Federal do Ceará, Fortaleza, 2019.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324. Acesso em: 2 de maio de 2025.

ÇAĞDAŞ, V. An application domain extension to CityGML for immovable property taxation: A Turkish case study. **International Journal of Applied Earth Observation and Geoinformation**. DOI: https://doi.org/10.1016/j.jag.2012.07.013. Acesso em: 06 out. 2025.

CARRANZA, J. P. et al. Random Forest como técnica de valuación masiva del valor del suelo urbano: una aplicación para la ciudad de Río Cuarto, Córdoba, Argentina. In: **COBRAC 2018 – Congresso Brasileiro de Cadastro Técnico Multifinalitário e Gestão Territorial**, 2018, Florianópolis.

CODES, B. N. **Avaliação dos preços de imóveis na cidade de Fortaleza, com a utilização de redes neurais artificiais, para a composição do ITBI**. 2018. Dissertação (Mestrado em Economia) – Universidade Federal do Ceará, Fortaleza, 2018.

DANTAS, R. A.; MAGALHÃES, A. M.; VERGOLINO, J. R. de O. Avaliação de imóveis: a importância dos vizinhos no caso de Recife. **Economia Aplicada**, São Paulo, v. 11, n. 2, p. 231-251, abr./jun. 2007. DOI: https://doi.org/10.1590/S1413-80502007000200004. Acesso em: 14 de abr. de 2025.

DAS, A.; RAD, P.. **Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey**. 2020. DOI: https://doi.org/10.48550/arXiv.2006.11371. Acesso em 15 de abr. de 2025.

DEPPNER, J.; AHLEFELDT-DEHN, B.; BERACHA, E.; SCHÄFERS, W.. Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach. **The Journal of Real Estate Finance and Economics**, v., p. 1-38, 22 mar. 2023. DOI: 10.1007/s11146-023-09944-1. Acesso em: 17 de abr. de 2025.

FORTALEZA. Lei nº 11.515, de 27 de dezembro de 2024. **Lei Orçamentária Anual**. Suplemento ao Diário Oficial do Município de Fortaleza 17.977 de 27 de dezembro de 2024.

FORTALEZA. Secretaria Municipal das Finanças. **Relatórios de Execução Orçamentária (RREO)**. Disponível em: https://www.sefin.fortaleza.ce.gov.br/contas-publicas/relatorios-de-execucao-rreo. Acesso em: 9 out. 2025.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. **International Conference On Machine Learning**, 1996.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, out. 2001. DOI: https://doi.org/10.1214/aos/1013203451. Acesso em: 2 de maio de 2025.

GIMENES, F. S. F. **Defasagem na planta genérica de valores imobiliários e impactos na arrecadação do Imposto Predial e Territorial Urbano no Município de Fortaleza**. 2020. Dissertação (Mestrado em Economia do Setor Público) – Universidade Federal do Ceará, Fortaleza, 2020.

GONZÁLEZ, M. A. S.; FORMOSO, C. T. Estimativa de modelos de preços hedônicos para locação residencial em Porto Alegre. **Production**, v. 5, n. 1, p. 65-77, 1995.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016, disponível em http://www.deeplearningbook.org. Acesso em: 6 de out. de 2025.

GROVER, R. Mass valuations. **Journal of Property Investment & Finance**, v. 34, n. 2, p. 191–204, 2016. DOI: https://doi.org/10.1108/JPIF-01-2016-0001. Acesso em: 15 de abr. de 2025.

HARRISON, M. **Machine learning: guia de referência rápida: trabalhando com dados estruturados em Python**. Tradução de Lúcia A. Kinoshita. Revisão gramatical de Tássia Carvalho. São Paulo: Novatec, 2020. ISBN 978-85-7522-818-0.

HARJEET-BLUE. **Ensemble Learning Bagging and Boosting**. 2025. Disponível em: https://github.com/harjeet-blue/Ensemble-Learning-Bagging-and-Boosting. Acesso em: 06 out. 2025.

IBAN, M. C. An explainable model for the mass appraisal of residences: the application of tree-based machine learning algorithms and interpretation of value determinants. **Heliyon**, 2022. DOI: https://doi.org/10.1016/j.heliyon.2022.e10340. Acesso em: 9 de maio de 2025.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Censo Brasileiro de 2022**. Rio de Janeiro: IBGE, 2025.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Sistema de Contas Nacionais: Brasil**. Disponível em: https://www.ibge.gov.br/estatisticas/economicas/servicos/9052-sistema-de-contas-nacionais-brasil.html. Acesso em: 9 out. 2025.

INTERNATIONAL ASSOCIATION OF ASSESSING OFFICERS (IAAO). **Standards on Mass Appraisal of Real Property**, 2017.

KOCHULEM, E., MWANIKI, D. and MUTTUA, F. (2023) Mass Valuation of Unimproved Land Value Case Study: Nairobi County. **Journal of Geographic Information System**, 15, 122-139. doi: 10.4236/jgis.2023.151008. Acesso em: 17 de abr. de 2025.

LANCASTER, K. J. A new approach to consumer theory. **Journal of Political Economy**, v. 74, n. 2, p. 132–157, 1966. DOI: https://doi.org/10.1086/259131. Acesso em: 14 de abr. de 2025.

LUCENA, José Mário Pereira de. **O mercado habitacional no Brasil**. 1985. Tese - FGV, Escola de Economia de São Paulo. Disponível em: https://hdl.handle.net/10438/13074. Acesso em: 8 de out. de 2025.

LUNDBERG, S. M. **SHAP Documentation**. Disponível em: https://shap.readthedocs.io/en/latest/. Acesso em: 24 de maio de 2025.

LUNDBERG, S. M.; LEE, Su-In. A unified approach to interpreting model predictions. In: **Advances in Neural Information Processing Systems**, 2017.

MOLNAR, C. (2025). **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable** (3rd ed.). Disponível em: christophm.github.io/interpretable-ml-book/. Acesso em: 13 de abr. de 2025.

OLIVEIRA, A. A. F. **Avaliação em massa com modelos de aprendizado de máquina aplicados aos terrenos urbanos do município de Fortaleza**. 2020. 80 f. Dissertação (Mestrado em Economia do Setor Público) – Universidade Federal do Ceará, Fortaleza, 2020.

OLIVEIRA, A. A. F.; REYES-BUENO, F.; GONZÁLEZ, M. A. S.; SILVA, E. Comparing traditional and machine learning techniques in apartments mass appraisal in Fortaleza, Brazil. **Aestimum**, v. 85, p. 21-38, 2024. DOI: 10.36253/aestim-15344. Acesso em: 20 de maio de 2025.

OLIVEIRA, A. A. F. de; BANDEIRA, S. R. V.; TÁVORA, V. J. Mass appraisal of urban land with homogenization factors: a spatial models-based approach. **Revista Valorem**, [S. l.], v. 3, n. 1, p. 16–32, 2024. DOI: 10.29327/2290393.3.1-2. Disponível em: https://www.revistavalorem.com/index.php/home/article/view/23. Acesso em: 17 abr. 2025.

OSLAND, L. An application of spatial econometrics in relation to hedonic house price modeling. **Journal of Real Estate Research**, v. 32, n. 3, p. 289–320, 2010. DOI: https://doi.org/10.1080/10835547.2010.12091282. Acesso em: 20 de abr. de 2025.

PELLI NETO, A.; MORAIS, G. R. Redes neurais artificiais sob dupla ótica: modelando a análise envoltória de dados (EDO DEA) para aplicação nas avaliações de imóveis urbanos. In: **XXII Congresso Pan Americano de Avaliações**, 2006.

ROSEN, S. Hedonic prices and implicit markets: product differentiation in pure competition. **Journal of Political Economy**, v. 82, n. 1, p. 34–55, 1974. DOI: https://doi.org/10.1086/260169. Acesso em: 16 de abr. de 2025.

SANT'ANNA, Marcelo; IACHAN, Felipe S.; GUEDES, Ricardo. Housing supply in the presence of informality. **Ensaios Econômicos**, 2021. Disponível em: https://hdl.handle.net/10438/5. Acesso em: 8 out. 2025.

SECRETARIA MUNICIPAL DAS FINANÇAS DE FORTALEZA (Sefin). **Infraestrutura de Dados Espaciais de Fortaleza**. Disponível em: https://ide.sefin.fortaleza.ce.gov.br/. Acesso em: 02 de jun. de 2025.

SHAPLEY, L. S. **A value for n-person games**. Santa Monica, CA: RAND Corporation, 1952. (Paper P-295). DOI: 10.7249/P0295. Acesso em: 20 de abr. de 2025.

SILVA, R. C. E. de O.; DOS SANTOS, D. F.; MAURINHO, G. A.; BUENO, P.V; CATAPAN, A. As transformações do mercado imobiliário brasileiro nos anos 2000 – uma análise do ponto de vista legal e econômico. **Revista da Ciência da Administração**.

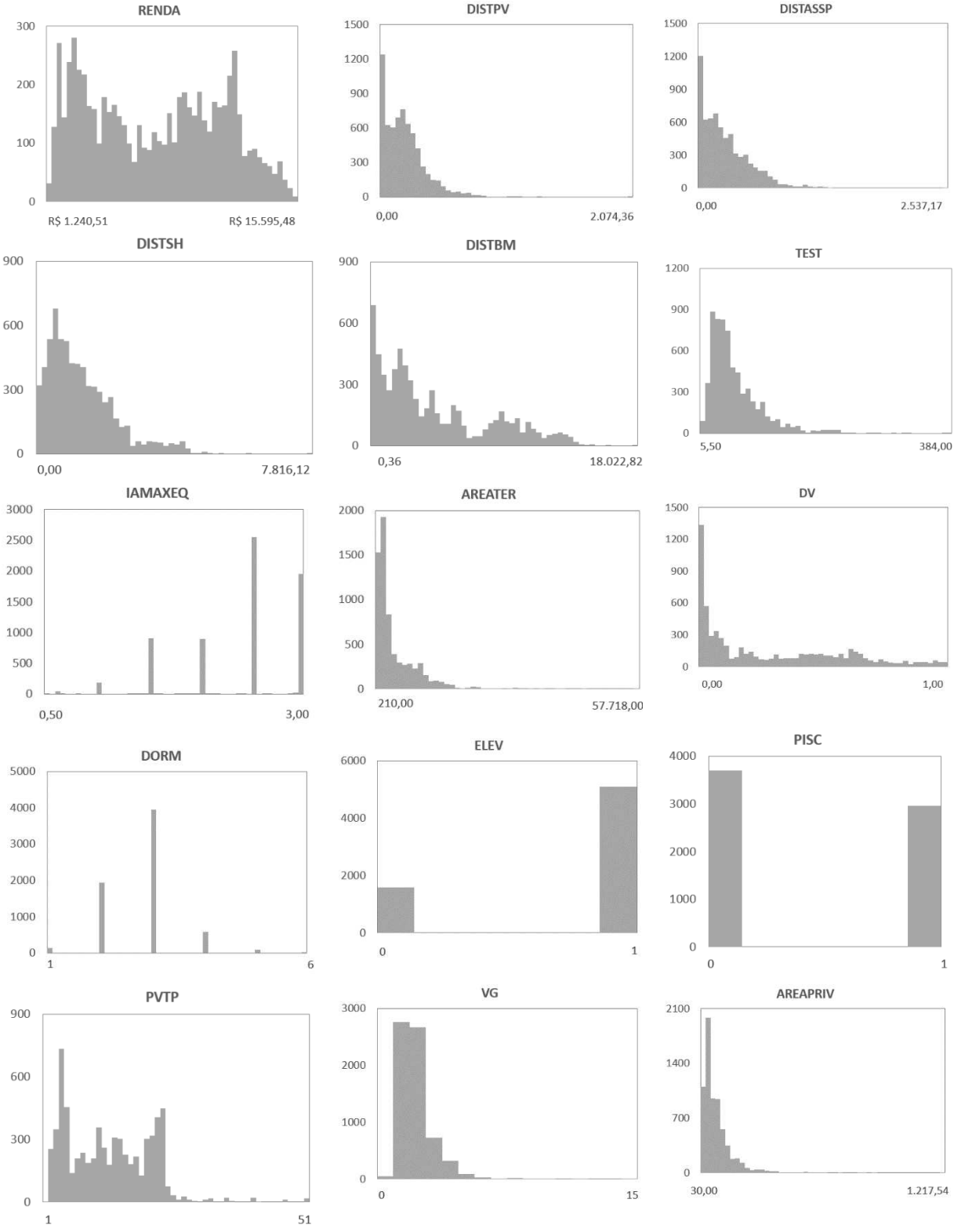
TARASOV, S.; DESSOULAVY-ŚLIWIŃSKI, B. Algorithm-driven hedonic real estate pricing: an explainable AI approach. **Real Estate Management and Valuation**, 2024. DOI: https://doi.org/10.2478/remav-2025-0003. Acesso em: 20 de mar. de 2025.

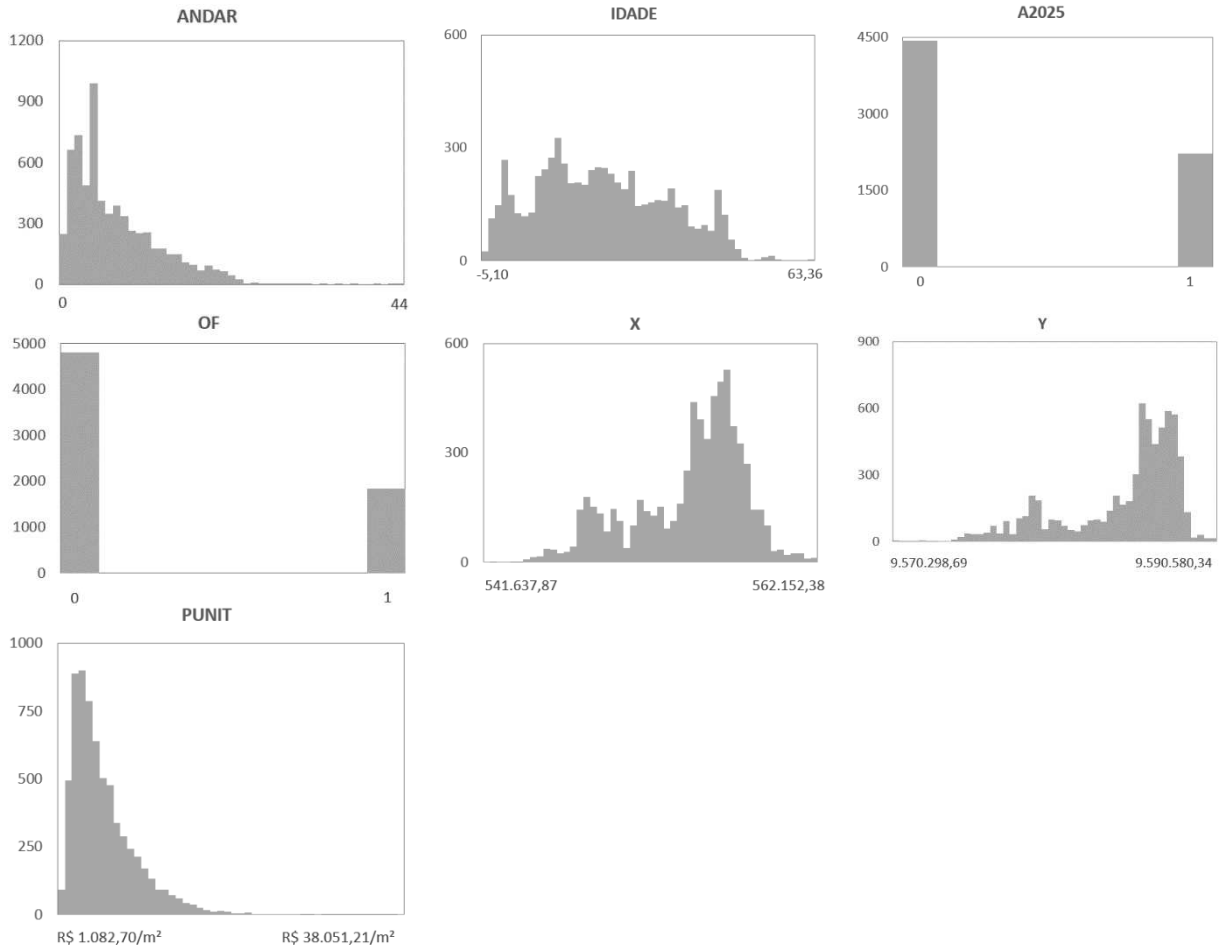
YILMAZER, S.; KOCAMAN, S. A mass appraisal assessment study using machine learning based on multiple regression and random forest. **Land Use Policy**, 2020. DOI: 10.1016/j.landusepol.2020.104889. Acesso em: 17 de abr. de 2025.

ZILLI, C. A.; BASTOS, L. C. Mass appraisal of apartments using Random Forest and Gradient Boosting algorithms: case study of Florianópolis, Brazil. **Revista do Departamento de Geografia**, Universidade de São Paulo, v. 44, 2024. DOI: 10.11606/eISSN.2236-2878.rdg.2024.212297. Acesso em: 2 de mai. de 2025.

ZILLI, C. A.; BASTOS, L. C.; SILVA, L. R. da. Machine learning models in mass appraisal for property tax purposes: a systematic mapping study. **Aestimum**, v. 84, p. 31-52, 2024. DOI: 10.36253/aestim-15792. Acesso em: 13 de maio de 2025.

APÊNDICE A – HISTOGRAMAS DAS VARIÁVEIS DA AMOSTRA





Fonte: Elaborado pelo autor.