CENTRO UNIVERSITÁRIO SERRA DOS ÓRGÃOS - UNIFESO
DIREÇÃO ACADÊMICA DE CIÊNCIAS HUMANAS E TECNOLÓGICAS - DACHT
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL SILVA MEDINA

APLICAÇÃO DE MACHINE LEARNING E MODELOS DE PREÇOS HÊDONICOS NA PRECIFICAÇÃO DE IMÓVEIS RESIDENCIAIS EM SÃO PAULO

TERESÓPOLIS
2025

CENTRO UNIVERSITÁRIO SERRA DOS ÓRGÃOS - UNIFESO
DIREÇÃO ACADÊMICA DE CIÊNCIAS HUMANAS E TECNOLÓGICAS - DACHT
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL SILVA MEDINA

APLICAÇÃO DE MACHINE LEARNING E MODELOS DE PREÇOS HÊDONICOS NA PRECIFICAÇÃO DE IMÓVEIS RESIDENCIAIS EM SÃO PAULO

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Serra dos Órgãos como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Antônio de Paula Pedrosa

TERESÓPOLIS
2025

Dados Internacionais de Catalogação na Publicação
Centro Universitário Serra dos Órgãos
SIB-Unifeso

Ficha catalográfica gerada automaticamente com os dados fornecidos pelo(a) autor(a).

M443 Medina, Gabriel.
APLICAÇÃO DE MACHINE LEARNING E MODELOS DE PREÇOS HÊDONICOS NA PRECIFICAÇÃO DE IMÓVEIS RESIDENCIAIS EM SÃO PAULO / Gabriel Medina; orientador Antonio Pedrosa. -- Teresópolis, 2025.
49 f. : il. color.

Trabalho de Conclusão de Curso (GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO), Centro Universitário Serra dos Órgãos(Unifeso), Teresópolis, 2025.

1. Precificação de Imóveis. 2. Modelo de Preços Hedônicos. 3. Machine Learning. 4. AutoML. 5. Mercado Imobiliário. I. Pedrosa, Antonio, orient. II. Título.

CDD 004

CENTRO UNIVERSITÁRIO SERRA DOS ÓRGÃOS
DIREÇÃO ACADÊMICA DE CIÊNCIAS HUMANAS E TECNOLÓGICAS – DACHT
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA EM SÉRIES TEMPORAIS PARA A PREDIÇÃO DE PREÇOS DE IMÓVEIS: UM ESTUDO DE CASO DA CIDADE DE SÃO PAULO

GABRIEL SILVA MEDINA

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Serra dos Órgãos como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação

Prof. Antonio de Paula Pedrosa MSc.
Orientador

Prof. Gabriel Resende Machado MSc.
PUCRJ

Prof. Vagner Zeizer Carvalho Paes MSc.
PUCRJ

TERESÓPOLIS
2025

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.
- Lauro César em abnTeX2

AGRADECIMENTOS

A realização deste Trabalho de Conclusão de Curso não teria sido possível sem o apoio, a orientação e o incentivo de muitas pessoas especiais a quem desejo expressar minha mais profunda gratidão.

Em primeiro lugar, agradeço de todo coração aos meus pais, Sandra e Adevalter, por serem meu alicerce. Obrigado por todo o apoio incondicional, paciência e por acreditarem em mim durante todos os anos de estudo. Esta conquista é, sem dúvida, nossa.

Dedico este trabalho, em memória, à minha querida irmã, Evylane. Ela sempre foi um exemplo acadêmico para mim, e sua paixão pela educação, sua área de atuação, foi uma inspiração constante. Sua memória me deu forças para continuar.

No âmbito acadêmico, agradeço imensamente ao meu orientador, Professor Antonio de Paula Pedrosa. Sua presença, orientação segura e confiança neste projeto foram fundamentais para que eu chegasse até aqui.

Minha gratidão especial se estende ao Professor Gabriel Resende Machado e ao Doutor Vagner Zeizer Carvalho Paes. Ambos foram fundamentais desde a concepção deste estudo, ajudando-me a idealizar a proposta original do projeto. Além disso, obrigado por acompanharem de perto meu progresso, pelas dicas valiosas e pelas contribuições que enriqueceram imensamente esta pesquisa.

Agradeço à UNIFESO por me proporcionar um ambiente de aprendizado fértil, pelos recursos disponibilizados e pelo corpo docente que contribuiu para minha formação.

Aos meus amigos, minha mais profunda gratidão. Vocês estiveram diariamente ao meu lado durante todo este processo, tornando a jornada universitária não apenas mais leve, mas possível. O apoio mútuo, as conversas, o companheirismo e a paciência foram pilares essenciais para superar os desafios da graduação.

Um agradecimento especial dedico ao meu amigo Rodrigo Matos. Sua influência e seu incentivo foram decisivos na minha escolha por esta área tão incrível que é a computação. Se hoje concluo este ciclo, muito se deve a ajuda e motivação que tive durante todo o percurso.

RESUMO

O mercado imobiliário, caracterizado por sua heterogeneidade e assimetria de informações, apresenta desafios significativos para a avaliação precisa de imóveis. Este trabalho propõe o desenvolvimento de um modelo de preços hedônicos baseado em técnicas de *Machine Learning* para a precificação de imóveis residenciais na cidade de São Paulo, utilizando como base de dados os registros públicos do ITBI de 2006 a 2024. A metodologia abrange uma pipeline de ciência de dados, incluindo a validação e consolidação de mais de 2,3 milhões de registros, engenharia de *features* com ajuste de valores pela inflação (IPCA), e a criação de uma *feature* de macrorregião por meio de clusterização geoespacial com K-Means. A partir da Análise Exploratória de Dados, que revelou dinâmicas de preço distintas, a modelagem foi segmentada entre imóveis verticais e horizontais. Foram comparados modelos interpretáveis, como Regressão Linear e Árvore de Decisão, com uma abordagem avançada utilizando H2O AutoML e Validação Cruzada. Os resultados demonstram a ineficácia dos modelos lineares e a superioridade dos algoritmos não-lineares. O modelo final, um *Stacked Ensemble* para o segmento vertical e um *Gradient Boosting Machine* (GBM) para o horizontal, alcançou a melhor performance, reduzindo significativamente o erro médio de previsão (RMSE) em 17,6% e 11%, respectivamente, em comparação com os modelos de base. O trabalho conclui sobre a viabilidade e a robustez da aplicação de *Machine Learning* para a precificação de imóveis, estabelecendo um *baseline* de performance e uma metodologia reprodutível para futuras pesquisas na área.

Palavras-chave: Precificação de Imóveis. Modelo de Preços Hedônicos. Machine Learning. AutoML. Mercado Imobiliário.

ABSTRACT

The real estate market, characterized by its heterogeneity and information asymmetry, presents significant challenges for accurate property valuation. This work proposes the development of a hedonic pricing model based on Machine Learning techniques for the pricing of residential properties in the city of São Paulo, using public records from the ITBI database from 2006 to 2024. The methodology encompasses a data science pipeline, including the validation and consolidation of over 2.3 million records, feature engineering with inflation adjustment (IPCA), and the creation of a macro-region feature through K-Means geospatial clustering. Based on Exploratory Data Analysis, which revealed distinct price dynamics, the modeling was segmented between vertical and horizontal properties. Interpretable models, such as Linear Regression and Decision Tree, were compared with an advanced approach using H2O AutoML and Cross-Validation. The final model, a Stacked Ensemble for the vertical segment and a Gradient Boosting Machine (GBM) for the horizontal one, achieved the best performance, significantly reducing the Root Mean Squared Error (RMSE) by 17.6% and 11%, respectively, compared to the baseline models. The study concludes on the viability and robustness of applying Machine Learning for real estate pricing, establishing a performance baseline and a reproducible methodology for future research in the field.

Keywords: Real Estate Pricing. Hedonic Price Model. Machine Learning. AutoML. Housing Market.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AED	Análise Exploratória de Dados
AutoML	Aprendizado de Máquina Automatizado (do inglês, Automated Machine Learning)
CRISP-DM	Processo Padrão Inter-Indústrias para Mineração de Dados (do inglês, Cross-Industry Standard Process for Data Mining)
CV	Validação Cruzada (do inglês, Cross-Validation)
GBM	Máquina de Aumento de Gradiente (do inglês, Gradient Boosting Machine)
IPCA	Índice Nacional de Preços ao Consumidor Amplo
IPTU	Imposto Predial e Territorial Urbano
ITBI	Imposto sobre a Transmissão de Bens Imóveis
K-Means	K-Médias (algoritmo de clusterização)
R^2	Coeficiente de Determinação
RMSE	Raiz do Erro Quadrático Médio (do inglês, Root Mean Squared Error)
TCC	Trabalho de Conclusão de Curso

SUMÁRIO

1 INTRODUÇÃO

O mercado imobiliário é um dos setores mais relevantes da economia. Ele impulsiona o desenvolvimento urbano, representa um dos principais veículos de investimento da população e, acima de tudo, envolve um bem essencial, a moradia. Devido à sua complexidade, a precificação de um imóvel é uma tarefa desafiadora, pois depende de uma ampla gama de atributos. A literatura sobre o tema, em uma classificação consolidada e apresentada em trabalhos como o de Chin e Chau (2003), costuma agrupar esses atributos em três categorias principais: estruturais, locacionais e de vizinhança. Em grandes centros urbanos como São Paulo, a heterogeneidade das propriedades e a dinâmica acelerada do mercado tornam a avaliação de um ativo um processo intrincado e, muitas vezes, subjetivo (MATTA, 2007).

Essa subjetividade, associada à assimetria de informações, representa um problema recorrente no setor. Em muitas negociações, uma das partes detém mais conhecimento sobre o valor real do bem, o que gera desequilíbrios e ineficiências. A crítica aos modelos de avaliação tradicionais podem ser vistas também em estudos como os de Matta (2007) e Silva (2016), que apontam que esses métodos frequentemente resultam em precificações desalinhadas com a realidade do mercado. Isso ocorre porque eles não capturam a complexidade das interações entre os diversos atributos de um imóvel. Como consequência, investidores enfrentam maior insegurança e famílias têm mais dificuldade em tomar decisões bem-informadas, perpetuando um ciclo de ineficiência e falta de transparência.

Diante dessas limitações, a abordagem por meio de modelos de preços hedônicos surge como uma alternativa promissora. Essa teoria propõe que o preço de um bem heterogêneo pode ser explicado pela soma dos valores implícitos de seus atributos individuais (CHIN; CHAU, 2003). No contexto deste projeto, isso significa que o valor de um apartamento ou casa não é visto como um todo, mas sim como um "pacote"de atributos, como sua área construída, sua localização (seja pela coordenada exata ou pelo cluster regional) e sua idade, onde o modelo de *Machine Learning* aprende a calcular o "preço"implícito de cada um desses componentes. A aplicação de técnicas de *Machine Learning* (Aprendizagem de Máquina) representa uma evolução natural para a implementação desses modelos, permitindo superar limitações da regressão linear clássica e estimar com maior precisão o impacto de cada característica no valor final de um imóvel.

Nesse contexto, o objetivo principal deste trabalho é desenvolver um modelo hedônico de precificação de imóveis para a cidade de São Paulo, utilizando técnicas de *Machine Learning*. Para isso, será utilizada a base pública do ITBI da prefeitura de São Paulo, garantindo o uso de valores de transações reais. Em seguida, serão aplicadas técnicas de tratamento e análise exploratória para preparar os dados. Diferentes algoritmos de regressão serão treinados e avaliados para selecionar o modelo com o melhor desempenho preditivo. Espera-se, ao final, oferecer uma ferramenta capaz de reduzir a assimetria de informações e aumentar a transparência

no mercado imobiliário paulistano.[1]

1.1 PROBLEMA DE PESQUISA

O desafio central deste trabalho está em estabelecer, de forma objetiva e precisa, o valor de um imóvel em um mercado tão complexo e dinâmico quanto o de São Paulo. Esse cenário é marcado por uma acentuada assimetria de informações, em que uma das partes envolvidas na transação detém mais conhecimento sobre os fatores que determinam o preço do ativo, gerando desequilíbrios e ineficiências (SIEBRA, 2024). Essa lacuna informacional dificulta a tomada de decisão de todos os agentes, desde o cidadão comum até grandes investidores e o poder público.

As causas dessa assimetria são diversas. Cada imóvel é um bem único, com características construtivas, locacionais e de vizinhança próprias, o que torna a comparação direta entre propriedades uma tarefa difícil. Além disso, muitos métodos tradicionais de avaliação são insuficientes para lidar com essa complexidade. Estudos como o de Matta (2007) mostram que modelos simplistas, como a "planta de valores"usada para fins fiscais, frequentemente ignoram atributos essenciais para a formação de preço. Outro problema é a dependência de dados de anúncios online, que refletem o preço de oferta, e não o valor real de fechamento das transações, introduzindo um viés que distorce a percepção de mercado.

As consequências de uma precificação imprecisa são amplas. Para compradores e vendedores, a incerteza sobre o valor justo de um patrimônio gera insegurança e pode causar perdas financeiras. No setor financeiro, avaliações falhas comprometem o cálculo de risco e dificultam a concessão de crédito imobiliário. Em uma escala mais ampla, a falta de transparência prejudica a formulação de políticas públicas e pode levar a injustiças tributárias, afetando a arrecadação e o planejamento urbano Gazola (2002).

Com base nessas considerações, o problema de pesquisa que orienta este trabalho pode ser formulado da seguinte forma: **De que maneira a aplicação de modelos de *Machine Learning*, utilizando uma base de dados de transações reais, pode gerar um sistema de precificação de imóveis mais acurado e transparente para o mercado residencial de São Paulo, contribuindo para a redução da assimetria de informações?**

1.2 HIPÓTESE

Parte-se da hipótese de que a aplicação de técnicas de *Machine Learning* sobre uma base de dados do ITBI permite o desenvolvimento de um modelo de preços hedônicos com acurácia preditiva superior aos métodos convencionais. Esse modelo poderia atuar como uma ferramenta eficaz na redução da assimetria informacional no mercado imobiliário paulistano.

[1] O código-fonte completo da pipeline de dados, bem como os notebooks de modelagem e análise desenvolvidos para este trabalho, estão disponíveis publicamente no repositório do projeto em: <https://github.com/GabrielSMedina/TCC_modelo_hedonico.git>

Essa premissa se apoia na teoria de preços hedônicos, apresentada por Chin e Chau (2003), que permite decompor o valor de um bem em seus atributos constituintes. Além disso, estudos comparativos, como o de Borges, Salviato e Goes (2024), indicam que algoritmos de *Machine Learning*, como Redes Neurais e *XGBoost*, superam a regressão linear clássica em tarefas de avaliação imobiliária, pois modelam de forma mais eficiente as relações não-lineares que caracterizam esse mercado (LIMSOMBUNCHAI; GAN; LEE, 2004; ZULKIFLEY et al., 2020).

1.3 OBJETIVOS

1.3.1 OBJETIVO PRINCIPAL

Desenvolver um modelo preditivo de preços hedônicos, baseado em técnicas de *Machine Learning*, para o mercado de imóveis residenciais da cidade de São Paulo, oferecendo uma ferramenta robusta para reduzir a assimetria de informações e aumentar a transparência nas negociações imobiliárias.

1.3.2 OBJETIVOS ESPECÍFICOS

1. Tratar a base de dados do ITBI da prefeitura de São Paulo, estruturando um conjunto de dados consistente e adequado à modelagem;
2. Conduzir uma análise exploratória para identificar as principais características dos imóveis, suas distribuições e as correlações entre os atributos e o preço final;
3. Treinar, testar e comparar o desempenho de diferentes algoritmos de regressão de *Machine Learning*, como Regressão Linear, *Random Forest* e *XGBoost*, para a tarefa de predição de preços;
4. Selecionar o modelo com melhor desempenho preditivo, com base em métricas de erro estatísticas, e validar sua capacidade de generalização;
5. Analisar a importância das variáveis no modelo final, identificando os atributos com maior influência no valor dos imóveis em São Paulo.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos que sustentam o desenvolvimento do presente trabalho. A abordagem inicia-se com uma contextualização do mercado imobiliário, destacando suas particularidades e relevância econômica. Em seguida, é detalhado o Modelo de Preços Hedônicos, que constitui a base conceitual para a metodologia de avaliação de imóveis adotada nesta pesquisa. Subsequentemente, são explorados os princípios de *Machine Learning*, com ênfase nas técnicas de regressão que foram aplicadas na prática para ilustrar os conceitos, incluindo a Regressão Linear, a Regressão de Ridge e as Árvores de Decisão, além de uma breve exposição sobre o algoritmo de clusterização K-Means. Por fim, são discutidos os conceitos de Análise Exploratória de Dados (AED) e as métricas de erro utilizadas para avaliar a performance dos modelos.

2.1 MERCADO IMOBILIÁRIO

O mercado imobiliário é amplamente reconhecido como um dos pilares da economia, exercendo um papel crucial no crescimento do Produto Interno Bruto (PIB) e no estímulo a toda a cadeia produtiva da construção civil (SIEBRA, 2024). Sua relevância transcende a esfera econômica, pois está diretamente associado a questões fundamentais como moradia, qualidade de vida, desenvolvimento urbano e formação de patrimônio para a população (TEODORO; KAPPEL, 2020). Dada a sua importância, este setor recebe atenção especial tanto em economias desenvolvidas quanto naquelas em desenvolvimento, sobretudo em períodos de instabilidade econômica (SIEBRA, 2024).

Diferentemente de outros mercados, o imobiliário possui características singulares que o tornam particularmente complexo. Conforme aponta Matta (2007), os imóveis são bens distintos por sua vida útil elevada, singularidade e localização fixa. Essas três características, durabilidade, heterogeneidade e imobilidade espacial fazem com que cada propriedade seja um produto economicamente único, sem substitutos perfeitos (GAZOLA, 2002; CHIN; CHAU, 2003). Como consequência, não há no mercado imobiliário um imóvel exatamente igual a outro, o que o configura como um ambiente de concorrência imperfeita, onde a formação de preços não segue a mesma lógica de bens manufaturados (MATTA, 2007).

Essa heterogeneidade intrínseca torna a determinação do valor de um imóvel uma tarefa inerentemente complexa. Nesse contexto, é fundamental distinguir os conceitos de "valor"e "preço". O valor de mercado é definido como a quantia mais provável pela qual um bem seria negociado voluntariamente em um mercado livre e competitivo, sendo, portanto, uma estimativa ou uma projeção (MATTA, 2007). O preço, por sua vez, é a quantia monetária efetivamente paga em uma transação específica. A discrepância entre esses dois conceitos é frequente e justifica a necessidade de metodologias científicas que busquem estimar o valor de mercado de forma mais precisa e isenta de subjetividade, aproximando-o ao máximo do preço que seria praticado em

uma negociação justa e transparente (GAZOLA, 2002).

2.2 MODELOS HEDÔNICOS

A precificação de bens heterogêneos, como os imóveis, exige uma abordagem que vá além da simples comparação de preços. Nesse contexto, a teoria de preços hedônicos, ou modelo hedônico, fornece o principal arcabouço teórico para a avaliação imobiliária moderna. A premissa central desta teoria é que um bem não é visto como uma unidade única, mas sim como um pacote de atributos ou características individuais, e o seu valor de mercado pode ser decomposto na soma dos preços implícitos de cada um desses atributos (CHIN; CHAU, 2003; SILVA, 2016).

O desenvolvimento teórico do modelo hedônico é amplamente atribuído aos trabalhos seminais de Lancaster (1966) e, posteriormente, à formalização de Rosen (1974). A teoria postula que os consumidores derivam utilidade não do bem em si, mas das características que ele possui. No mercado imobiliário, isso significa que um comprador não adquire simplesmente uma "casa"ou um "apartamento", mas sim um conjunto de atributos como a área construída, o número de quartos, a qualidade do acabamento, a segurança da vizinhança e a acessibilidade a serviços. O preço que o mercado atribui a um imóvel, portanto, reflete a valoração agregada que os consumidores dão a esse conjunto específico de características (CHIN; CHAU, 2003).

Para estimar os preços implícitos de cada atributo, o modelo hedônico utiliza, em sua forma mais tradicional, a Regressão Linear Múltipla. A relação é formalizada por meio de uma equação onde o preço do imóvel é a variável dependente, e seus diversos atributos são as variáveis independentes, conforme a equação genérica:

$$P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \tag{2.1}$$

onde P é o preço do imóvel, $X_1, X_2, \ldots, X_k$ representam os diferentes atributos (como área, número de quartos, etc.), os coeficientes $\beta_1, \beta_2, \ldots, \beta_k$ são os preços implícitos marginais de cada atributo, β_0 é o intercepto e ε representa o termo de erro do modelo (TEODORO; KAPPEL, 2020). A literatura especializada costuma agrupar os atributos imobiliários em três categorias principais:

- **Atributos Estruturais:** Referem-se às características físicas da edificação, como área construída, número de quartos, número de banheiros, vagas de garagem e idade do imóvel (CHIN; CHAU, 2003; ZULKIFLEY et al., 2020).
- **Atributos Locacionais:** Descrevem a localização da propriedade e sua acessibilidade a pontos de interesse, como a distância ao centro da cidade, a proximidade de estações de transporte público, escolas, hospitais e centros comerciais (CHIN; CHAU, 2003; ZULKIFLEY et al., 2020).

- **Atributos de Vizinhança:** Englobam características da área circundante que afetam a qualidade de vida, como os níveis de segurança ou criminalidade, a qualidade das escolas locais e a presença de áreas verdes e de lazer (CHIN; CHAU, 2003).

Ao estimar os coeficientes β, o modelo hedônico permite quantificar o impacto de cada um desses atributos no preço final de um imóvel, tornando-se uma ferramenta poderosa para a avaliação imobiliária objetiva e baseada em dados.

2.3 MACHINE LEARNING

Enquanto a teoria de preços hedônicos fornece o arcabouço conceitual para a avaliação imobiliária, a área de *Machine Learning* (Aprendizagem de Máquina) oferece o conjunto de ferramentas computacionais necessárias para construir, treinar e validar esses modelos em larga escala e com alta complexidade. A aprendizagem de máquina é um campo da inteligência artificial que se dedica a desenvolver algoritmos capazes de "aprender"padrões e realizar predições a partir de dados, sem serem explicitamente programados para cada tarefa específica (GÉRON, 2019).

No contexto da precificação de imóveis, os algoritmos de *Machine Learning* são particularmente poderosos. Diferentemente da regressão estatística clássica, muitas dessas técnicas são capazes de modelar relações não-lineares complexas, uma característica comum em mercados heterogêneos como o imobiliário, onde o impacto de um atributo no preço pode variar dependendo de outras características (ZULKIFLEY et al., 2020; LIMSOMBUNCHAI; GAN; LEE, 2004). Como demonstram diversos estudos, a aplicação desses modelos frequentemente resulta em uma maior acurácia preditiva quando comparada a abordagens puramente lineares (SILVA, 2019; BORGES; SALVIATO; GOES, 2024).

As subseções a seguir apresentarão os fundamentos dos principais algoritmos de regressão e clusterização aplicados neste trabalho. O objetivo é detalhar o funcionamento teórico de cada técnica, que será posteriormente ilustrado com os resultados práticos obtidos nos experimentos de modelagem.

2.3.1 MEDIDAS DE ERRO

Para avaliar quantitativamente a performance de um modelo de regressão e comparar diferentes algoritmos, é indispensável o uso de métricas de erro. Essas métricas calculam a magnitude da diferença entre os valores preditos pelo modelo e os valores reais observados. Embora existam diversas métricas, cada uma possui características específicas, sendo mais ou menos sensível a determinados tipos de erro (SILVA, 2019). As métricas mais comuns, utilizadas nos experimentos deste trabalho, são detalhadas a seguir.

- **Mean Absolute Error (MAE):** O Erro Absoluto Médio calcula a média das diferenças absolutas entre os valores preditos e os reais. Por não elevar os erros ao quadrado, o MAE

é menos sensível a *outliers* (valores discrepantes) e é facilmente interpretável, pois sua unidade é a mesma da variável alvo (BORGES; SALVIATO; GOES, 2024). Sua fórmula é:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.2}$$

onde n é o número de observações, y_i é o valor real e $\hat{y}_i$ é o valor predito.

- **Mean Squared Error (MSE):** O Erro Quadrático Médio é uma das métricas mais utilizadas, sendo a função de custo que muitos modelos, como a Regressão Linear, buscam minimizar. Ele calcula a média dos quadrados das diferenças entre predições e valores reais. Ao elevar o erro ao quadrado, o MSE penaliza fortemente erros grandes, tornando-se muito sensível a *outliers* (GÉRON, 2019). Sua fórmula é:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.3}$$

- **Root Mean Squared Error (RMSE):** O Erro Quadrático Médio da Raiz é simplesmente a raiz quadrada do MSE. Sua principal vantagem em relação ao MSE é que o resultado retorna à unidade original da variável alvo, facilitando a interpretação da magnitude do erro (LIMSOMBUNCHAI; GAN; LEE, 2004). Assim como o MSE, o RMSE também é sensível a *outliers*. Esta foi uma das principais métricas utilizadas para avaliar os modelos nos experimentos práticos deste trabalho.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.4}$$

- **Coeficiente de Determinação (R^2):** Diferente das métricas de erro, o R^2 (ou R-quadrado) mede o poder de explicação do modelo. Ele representa a proporção da variância da variável dependente que é explicada pelas variáveis independentes (GAZOLA, 2002). Seus valores variam tipicamente entre 0 e 1, onde 1 indica um ajuste perfeito. Contudo, como observado nos experimentos deste trabalho, valores negativos podem ocorrer quando o modelo se ajusta aos dados de forma pior do que uma simples média dos valores observados.

2.3.2 REGRESSÃO LINEAR

A Regressão Linear é um dos algoritmos mais fundamentais e amplamente utilizados no campo da aprendizagem de máquina supervisionada. Seu objetivo é modelar a relação entre uma variável de saída dependente (o alvo) e uma ou mais variáveis de entrada independentes (as características ou *features*), ajustando uma equação linear aos dados observados (GÉRON, 2019). Apesar de sua simplicidade, ela serve como um excelente ponto de partida e uma base de comparação (*baseline*) para modelos mais complexos. A premissa central é que o valor da variável dependente pode ser previsto como uma soma ponderada dos valores das características, acrescida de uma constante chamada de intercepto ou *bias*.

Matematicamente, o modelo de Regressão Linear Múltipla é expresso pela seguinte equação:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{2.5}$$

onde:

- $\hat{y}$ é o valor predito.
- n é o número de características (*features*).
- x_i é o valor da i-ésima característica.
- θ_0 é o intercepto (termo de *bias*).
- θ_j (para $j = 1, 2, \ldots, n$) é o peso ou coeficiente associado à j-ésima característica.

O processo de "treinamento"do modelo consiste em encontrar o conjunto de parâmetros θ (o intercepto θ_0 e os pesos das características θ_j) que melhor se ajusta aos dados de treinamento. Para determinar o "melhor ajuste", é necessário definir uma medida de desempenho ou, mais comumente, uma função de custo. A função de custo mais utilizada para a Regressão Linear é o Erro Quadrático Médio (*Mean Squared Error* - MSE), que calcula a média dos quadrados das diferenças entre os valores preditos e os valores reais, buscando encontrar os parâmetros θ que minimizam essa métrica Géron (2019). O MSE é uma de várias métricas de avaliação de desempenho, que serão exploradas em detalhe na Seção 2.3.1.

Para ilustrar o comportamento prático da Regressão Linear, um modelo foi treinado utilizando-se apenas as coordenadas geográficas (latitude e longitude) como variáveis de entrada para prever o preço dos imóveis no conjunto de dados deste trabalho. Os resultados, extraídos dos logs de execução, revelaram uma performance insatisfatória. No conjunto de teste para imóveis verticais, o modelo obteve um coeficiente de determinação (R^2) de -3.38 e um Erro Quadrático Médio da Raiz (RMSE) de R$ 775.984,49. Um valor de R^2 negativo indica que o modelo se ajustou aos dados de forma pior do que um simples modelo que previsse o valor médio para todos os imóveis. Este resultado evidencia a principal limitação da Regressão Linear: sua incapacidade de capturar relações complexas e não-lineares entre as variáveis, como a que existe entre a localização geográfica e o preço dos imóveis, justificando a necessidade de algoritmos mais avançados que serão explorados nas próximas seções.

2.3.3 REGRESSÃO DE RIDGE

A Regressão de Ridge é uma variação regularizada do modelo de Regressão Linear. Sua principal finalidade é evitar o sobreajuste (*overfitting*) e mitigar os problemas causados pela multicolinearidade, que ocorre quando as variáveis independentes são altamente correlacionadas

entre si (GAZOLA, 2002). Para alcançar esse objetivo, a Regressão de Ridge adiciona um termo de regularização à função de custo do modelo linear (o MSE). Esse termo, conhecido como norma L_2, corresponde à soma dos quadrados dos pesos dos coeficientes e é controlado por um hiperparâmetro α (alfa) (GÉRON, 2019). Na prática, essa adição penaliza coeficientes com valores muito grandes, forçando o modelo a encontrar um balanço entre se ajustar bem aos dados de treino e manter os pesos dos seus parâmetros pequenos, o que tende a produzir um modelo que generaliza melhor para novos dados.

No experimento prático realizado neste trabalho, um modelo de Regressão de Ridge foi treinado sob as mesmas condições do modelo linear anterior. Os resultados, no entanto, foram virtualmente idênticos aos da Regressão Linear simples, com o modelo para imóveis verticais atingindo um R^2 de -3.38 no conjunto de teste. Essa ausência de melhoria sugere que, para o conjunto de dados e as características utilizadas neste experimento específico, a multicolinearidade não era o problema dominante. A performance insatisfatória de ambos os modelos reforça a hipótese de que a principal limitação a ser superada é a natureza não-linear dos dados, algo que a Regressão de Ridge, por ser um modelo linear, também não consegue capturar.

2.3.4 ÁRVORE DE DECISÃO

As Árvores de Decisão são algoritmos de *Machine Learning* poderosos e versáteis, capazes de realizar tanto tarefas de classificação quanto de regressão. Sua popularidade deriva de sua estrutura intuitiva e de sua capacidade de modelar relações não-lineares complexas nos dados (GÉRON, 2019). O modelo funciona de maneira análoga a um fluxograma, onde cada nó interno representa um "teste"em uma característica (*feature*), cada ramo representa o resultado do teste, e cada nó folha (um nó terminal) contém a predição final. Para fazer uma previsão para uma nova instância, basta percorrer a árvore desde o nó raiz até um nó folha, respondendo às perguntas em cada nó.

Para tarefas de regressão, como a predição de preços de imóveis, o funcionamento é específico. Uma vez que uma instância percorre a árvore e alcança um nó folha, a predição do modelo não é uma classe, mas sim um valor numérico contínuo. Esse valor é calculado como a média dos valores da variável alvo (neste caso, o preço do imóvel) de todas as instâncias de treinamento que caíram naquele mesmo nó folha durante a fase de treinamento (GÉRON, 2019). Assim, a árvore segmenta o espaço de características em diferentes regiões (as folhas) e atribui um valor de predição constante para cada uma dessas regiões.

O processo de construção da árvore é realizado majoritariamente pelo algoritmo CART (*Classification and Regression Tree*). O algoritmo funciona de forma recursiva e "gulosa"(*greedy*). A partir do conjunto de treinamento completo no nó raiz, ele procura a melhor divisão possível dos dados. Uma divisão consiste em uma característica k e um valor de limiar t_k. O algoritmo testa todas as características e todos os limiares possíveis, buscando a combinação que divide o

conjunto de dados em dois subconjuntos que sejam os mais "puros"possíveis. Para a regressão, a "pureza"é medida pela redução do Erro Quadrático Médio (MSE). A divisão escolhida é aquela que minimiza o MSE ponderado dos subconjuntos resultantes. Esse processo é então repetido para cada um dos novos subconjuntos, criando novos nós e ramos, até que um critério de parada seja atingido (GÉRON, 2019).

A principal desvantagem das Árvores de Decisão é sua forte tendência ao sobreajuste (*overfitting*). Se não for restringida, a árvore continuará a se dividir até que cada nó folha contenha o menor número possível de instâncias, muitas vezes apenas uma. Um modelo como este se ajustará perfeitamente aos dados de treinamento, mas não terá capacidade de generalizar para novos dados. Para combater isso, são aplicadas técnicas de regularização, que consistem em restringir a liberdade do modelo. Isso é feito através de hiperparâmetros, como max_depth (que limita a profundidade máxima da árvore) ou min_samples_leaf (que define o número mínimo de amostras que um nó folha deve conter para ser criado), forçando o modelo a ser mais simples e, consequentemente, a generalizar melhor (GÉRON, 2019).

A superioridade da Árvore de Decisão em capturar padrões não-lineares é claramente demonstrada nos experimentos práticos deste trabalho. Ao contrário dos modelos lineares, que produziram um R^2 negativo, o modelo de Árvore de Decisão treinado para imóveis verticais (utilizando apenas latitude e longitude como dados de geolocalização) alcançou um R^2 de 0.6346 no conjunto de teste. Este resultado positivo e substancialmente superior demonstra que o modelo foi capaz de aprender e mapear as complexas relações entre a localização geográfica e o preço dos imóveis, algo que os modelos lineares, por sua natureza, não conseguiram fazer. Essa capacidade de modelar interações complexas torna as Árvores de Decisão e suas variações, que serão vistas a seguir, ferramentas fundamentais para a precificação hedônica.

2.3.5 VARIAÇÕES DA ÁRVORE DE DECISÃO

Apesar de sua capacidade de modelar relações complexas, uma única Árvore de Decisão sofre de algumas limitações, como a alta variância e uma forte tendência ao sobreajuste (*overfitting*) (GÉRON, 2019). Para superar essas desvantagens, foram desenvolvidas variações que utilizam a árvore como um bloco de construção para modelos mais robustos. A abordagem mais proeminente é a de *ensemble methods* (métodos de conjunto), onde múltiplos preditores são combinados para obter uma previsão de maior qualidade. A principal variação nesse quesito é o *Random Forest* (Floresta Aleatória).

Um *Random Forest* é, essencialmente, um conjunto de várias Árvores de Decisão. Ele funciona através da técnica de *Bootstrap Aggregating* (*Bagging*), onde cada árvore do conjunto é treinada em uma subamostra aleatória e com reposição dos dados de treinamento. Além disso, ao construir cada árvore, o algoritmo de divisão dos nós considera apenas um subconjunto aleatório de características, em vez de todas elas. Ao final, a predição do modelo para uma nova instância é obtida pela média das predições de todas as árvores individuais (GÉRON, 2019;

SILVA, 2019). Essa dupla aleatorização (nas amostras e nas características) produz árvores diversificadas, e a agregação de suas previsões resulta em um modelo com variância muito menor e maior capacidade de generalização do que uma única árvore. A eficácia dessa abordagem foi comprovada nos experimentos deste trabalho, onde o modelo *Random Forest* para imóveis verticais alcançou um R^2 de **0.7087** no conjunto de teste, um desempenho superior ao obtido pela Árvore de Decisão individual.

2.3.6 K-MEANS

Diferentemente dos algoritmos de regressão supervisionada abordados anteriormente, o K-Means é um dos mais populares algoritmos de aprendizagem não-supervisionada. Seu objetivo não é prever um valor numérico a partir de características rotuladas, mas sim identificar agrupamentos (ou *clusters*) em um conjunto de dados, particionando as observações em k grupos distintos com base em sua similaridade (GÉRON, 2019). A "similaridade"é medida pela distância de cada ponto de dados a um ponto central do *cluster*, conhecido como centróide.

O algoritmo funciona de maneira iterativa. Primeiramente, k centróides são inicializados aleatoriamente no espaço das características. Em seguida, o processo se repete em duas etapas: primeiro, cada instância de dados é atribuída ao *cluster* de seu centróide mais próximo; segundo, a posição de cada centróide é recalculada para ser a média de todas as instâncias atribuídas a ele. Essas duas etapas são repetidas até que os centróides parem de se mover significativamente, indicando que os *clusters* estão estáveis e o algoritmo convergiu (GÉRON, 2019).

Embora seja uma ferramenta de análise exploratória, o K-Means também é amplamente utilizado como uma técnica de engenharia de características. Em vez de usar variáveis contínuas diretamente, pode-se primeiro agrupar os dados e usar o rótulo do *cluster* como uma nova característica categórica para alimentar um modelo supervisionado.

2.4 ANÁLISE EXPLORATÓRIA DE DADOS

Antes da construção de qualquer modelo preditivo, uma etapa fundamental e indispensável é a Análise Exploratória de Dados (AED), também conhecida pelo seu acrônimo em inglês, EDA (*Exploratory Data Analysis*). O principal objetivo desta fase é investigar o conjunto de dados a fim de resumir suas principais características, descobrir padrões, identificar anomalias e testar hipóteses iniciais, frequentemente com o auxílio de métodos visuais e estatísticas descritivas (SILVA, 2016; GAZOLA, 2002). A AED não se trata de um conjunto rígido de regras, mas de uma filosofia de investigação que funciona como um guia na compreensão profunda dos dados antes de submetê-los a técnicas de modelagem mais formais.

O processo de AED geralmente envolve um conjunto de técnicas para investigar as variáveis e suas relações. Conforme demonstrado em trabalhos como o de Silva (2016), as principais atividades incluem:

- **Análise Univariada:** Focada em entender cada variável isoladamente. Para variáveis quantitativas, calculam-se medidas de tendência central (média, mediana) e de dispersão (desvio padrão, quartis), e utilizam-se gráficos como histogramas e *box plots* para visualizar a distribuição e a presença de *outliers*. Para variáveis qualitativas, analisam-se as frequências de cada categoria.

- **Análise Bivariada e Multivariada:** Investiga a relação entre duas ou mais variáveis. Gráficos de dispersão (*scatter plots*) e matrizes de correlação são ferramentas essenciais para visualizar e quantificar a força da relação linear entre pares de variáveis, ajudando a identificar possíveis preditores para o modelo e a detectar problemas como a multicolinearidade (SILVA, 2016).

- **Tratamento de Dados Ausentes e *Outliers*:** A AED é crucial para identificar registros com informações faltantes e valores discrepantes (*outliers*). A detecção, como a realizada por Teodoro e Kappel (2020) através da regra de intervalo interquartil (IQR), permite ao pesquisador tomar decisões informadas sobre como tratar esses pontos, seja por remoção, substituição ou outros métodos.

- **Transformação de Variáveis:** A análise das distribuições das variáveis frequentemente revela assimetrias que podem violar pressupostos de certos modelos, como a Regressão Linear. A AED ajuda a identificar a necessidade de aplicar transformações, como a logarítmica, para normalizar a distribuição dos dados e linearizar relações, como demonstrado por Silva (2019) na análise de preços de imóveis.

No contexto deste trabalho, a Análise Exploratória de Dados foi uma etapa primordial, aplicada intensamente sobre a base de dados do ITBI. Através dela, foi possível compreender a distribuição assimétrica dos preços dos imóveis em São Paulo, o que justificou a aplicação de uma transformação logarítmica na variável alvo. Além disso, a AED permitiu a identificação e o tratamento de registros inconsistentes e forneceu os primeiros *insights* sobre quais atributos imobiliários possuíam maior poder preditivo, orientando todo o processo de pré-processamento e modelagem subsequente.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta uma análise de trabalhos acadêmicos recentes que abordam a precificação de imóveis no Brasil por meio de modelos preditivos. O objetivo é contextualizar a presente pesquisa, identificar as metodologias e fontes de dados utilizadas em outros estudos e, a partir dessa comparação, evidenciar as lacunas existentes e as contribuições originais deste trabalho para a área. Foram selecionados dois estudos que, embora compartilhem o mesmo objetivo geral de avaliação imobiliária, diferem em escopo, fonte de dados e complexidade metodológica, permitindo um contraste que destaca a relevância desta pesquisa.

3.1 MODELO HEDÔNICO PARA ESTIMAÇÃO DO VALOR DE IMÓVEIS: APLICAÇÃO EM NOVA FRIBURGO-RJ

Em seu trabalho, Teodoro e Kappel (2020) propuseram o desenvolvimento de um modelo hedônico para estimar o valor de casas e apartamentos na cidade de Nova Friburgo, no Rio de Janeiro. O objetivo era identificar as características mais importantes para a avaliação de um imóvel na região e construir um modelo matemático simples baseado nelas. Para isso, os autores utilizaram a técnica de *web scraping* para extrair dados de portais de anúncios online, formando uma base de 905 imóveis. A metodologia envolveu a aplicação do algoritmo *Random Forest* para a seleção das variáveis mais relevantes, seguida pela construção do modelo preditivo utilizando a Regressão de Ridge, uma variação da regressão linear múltipla. O modelo final obteve um desvio percentual médio de aproximadamente 25% na base de testes.

Apesar da semelhança no objetivo de criar um modelo de precificação, o presente trabalho se diferencia em dois pontos fundamentais. O primeiro e mais crítico é a fonte de dados: enquanto o estudo em Nova Friburgo se baseou em preços de anúncios online, que representam o valor de oferta e estão sujeitos a vieses de negociação, esta pesquisa utiliza a base de dados do ITBI de São Paulo, que contém os valores reais de transação, oferecendo uma representação muito mais fiel e robusta da realidade do mercado. O segundo diferencial reside na robustez da modelagem: enquanto Teodoro e Kappel (2020) focaram em um único modelo de regressão linear, a abordagem deste TCC envolve a comparação de múltiplos e mais complexos algoritmos de *Machine Learning*, além da aplicação de técnicas avançadas como a clusterização geográfica com K-Means e a otimização de modelos com AutoML, buscando uma solução com maior poder preditivo e capacidade de generalização.

3.2 MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PRECIFICAÇÃO DE IMÓVEIS NA CIDADE DE FORTALEZA

De forma metodologicamente mais próxima a esta pesquisa, o trabalho de Silva (2019) propôs a aplicação e comparação de diversas técnicas de aprendizagem de máquina para a avaliação de imóveis na cidade de Fortaleza, Ceará. O estudo também teve como base de dados

as transações registradas para o cálculo do ITBI, fornecidas pela Secretaria de Finanças do município. Foram testados cinco modelos distintos: Regressão Linear, Regressão Gaussiana, *Random Forests*, *Gradient Boosting Machine* (GBM) e Redes Neurais Artificiais. Ao final, o autor concluiu que os modelos baseados em árvores de decisão (*Random Forests* e GBM) apresentaram os melhores resultados, com um erro médio percentual de 1,64% e 1,65%, respectivamente.

O trabalho de Silva (2019) serve como um excelente paradigma, pois valida a eficácia do uso de dados do ITBI e de algoritmos de *Machine Learning* para a precificação de imóveis em um grande centro urbano brasileiro. A contribuição desta pesquisa, portanto, se dá em duas frentes principais. A primeira é a aplicação e validação dessa abordagem em um mercado imobiliário ainda maior e mais complexo, o da cidade de São Paulo. A segunda, e mais importante, é a introdução de etapas metodológicas adicionais para o aprimoramento do modelo. Enquanto a pesquisa em Fortaleza focou na aplicação direta dos algoritmos, este TCC implementa uma etapa de engenharia de características, utilizando o K-Means para criar uma variável de localização mais sofisticada, e emprega uma plataforma de AutoML (H2O) para a busca sistemática e otimização do melhor modelo preditivo, representando um passo adiante na robustez e na automação do processo de modelagem.

4 METODOLOGIA

4.1 INTRODUÇÃO À METODOLOGIA

Este capítulo apresenta a metodologia empregada para o desenvolvimento deste trabalho, detalhando de forma sequencial todas as etapas que compõem a pipeline executada. O processo abrange desde a aquisição e validação dos dados brutos até a configuração e avaliação dos experimentos de modelagem preditiva.

A estrutura metodológica foi desenhada de forma a garantir a reprodutibilidade, a transparência e a robustez dos resultados, seguindo um fluxo lógico inspirado em frameworks consolidados da área de Data Science, como o Cross-Industry Standard Process for Data Mining (CRISP-DM) (CHAPMAN et al., 2000). O fluxo de trabalho foi dividido nas seguintes etapas: Obtenção e Validação dos Dados, Pré-processamento e Consolidação, Engenharia de Features, Análise Exploratória, Preparação dos Datasets para Modelagem e, por fim, o Desenho Experimental para o treinamento e avaliação dos algoritmos de Machine Learning. Cada uma dessas fases será detalhada nas seções subsequentes.

4.2 OBTENÇÃO E VALIDAÇÃO DOS DADOS BRUTOS

O ponto de partida de qualquer projeto de *data science* é a obtenção e validação dos dados. Esta etapa inicial é fundamental para garantir a integridade, a completude e a consistência estrutural dos dados brutos que servirão de alicerce para todas as fases subsequentes de análise e modelagem.

4.2.1 FONTE DE DADOS

A base de dados utilizada neste trabalho é composta pelos registros públicos do ITBI, disponibilizados pela prefeitura da cidade de São Paulo. A escolha desta fonte de dados é um diferencial crítico, pois, diferentemente de dados provenientes de portais imobiliários que refletem preços de oferta, o ITBI contém os valores reais de transação dos imóveis, oferecendo uma representação muito mais fiel e robusta da realidade do mercado.

A coleta abrangeu um período de 19 anos, iniciando em janeiro de 2006 e se estendendo até os dados mais recentes disponíveis em 2024. Inicialmente, o conjunto de dados era composto por 19 arquivos anuais em formato `.xlsx`, totalizando mais de 2,3 milhões de registros de transações imobiliárias na cidade.

4.2.2 PROCESSO DE VALIDAÇÃO ESTRUTURAL

Para garantir a qualidade e a consistência dos dados brutos, foi implementada uma pipeline de validação automatizada, documentada no log de execução. Este processo sistemático verificou os seguintes aspectos:

- **Validação de Arquivos Anuais:** O script confirmou a presença de todos os 19 arquivos anuais esperados, de 2006 a 2024, no diretório de dados brutos. Nenhum arquivo anual estava ausente.
- **Validação da Estrutura Interna:** Cada arquivo anual foi inspecionado para assegurar que continha as 12 planilhas correspondentes a cada mês do ano (de janeiro a dezembro). O log confirmou que todos os arquivos estavam completos, sem nenhuma planilha mensal faltante.
- **Consistência do Esquema:** Foi realizada uma verificação para garantir que todas as planilhas, em todos os arquivos, possuíssem o mesmo conjunto de colunas. O esquema de referência, definido a partir do primeiro arquivo, continha 28 colunas. A validação confirmou que não havia divergências no esquema de dados em todo o período analisado.

Ao final desta etapa, a integridade estrutural da base de dados foi atestada, permitindo prosseguir com segurança para as fases de pré-processamento e enriquecimento dos dados.

4.3 PRÉ-PROCESSAMENTO E CONSOLIDAÇÃO DE DADOS

Após a validação da integridade estrutural, a próxima fase consistiu em pré-processar e consolidar os múltiplos arquivos anuais em um único e coeso *dataset*. Esta etapa foi dividida em dois processos principais: o enriquecimento geográfico dos dados por meio do CEP e a construção do *dataset* principal, que envolveu a limpeza e a unificação das transações.

4.3.1 ENRIQUECIMENTO GEOGRÁFICO VIA CEP

A localização é um dos fatores mais preponderantes na determinação do valor de um imóvel. No entanto, os dados brutos continham apenas o Código de Endereçamento Postal (CEP), uma informação que, por si só, possui baixo valor preditivo direto. Para transformar essa informação em *features* geoespaciais de alto valor, foi executado um processo de enriquecimento.

Primeiramente, todos os arquivos de dados brutos foram processados para extrair uma lista completa e consolidada de todos os CEPs únicos mencionados nas transações. Em seguida, essa lista foi cruzada com um arquivo mestre de geolocalização (`sp_data.xlsx`), que continha a correspondência entre os CEPs e seus respectivos bairros e coordenadas (latitude e longitude).

O resultado deste processo foi a geração de dois artefatos de dados:

- **Tabela de Consulta Geográfica:** Um arquivo `.csv` contendo a relação `CEP -> Bairro, Longitude, Latitude` para todos os CEPs encontrados no arquivo mestre. A Tabela 1 ilustra a estrutura deste arquivo.

- **Relatório de CEPs Faltantes:** Um segundo arquivo contendo a lista de CEPs que estavam presentes nos dados transacionais mas não foram localizados no arquivo mestre, permitindo uma futura análise da qualidade dos dados.

Tabela 1 – Exemplo da Tabela de Consulta Geográfica gerada a partir do cruzamento de dados.

CEP	Bairro	Latitude	Longitude
1526010	Aclimação	-23.5634186	-46.6320658
1526020	Aclimação	-23.5684778	-46.6309113
1526030	Aclimação	-23.5643733	-46.6307993
1526040	Aclimação	-23.5665705	-46.6317736
1526050	Aclimação	-23.5676832	-46.6317707

Fonte: Autor (2025)

4.3.2 CONSTRUÇÃO DO DATASET PRINCIPAL

Com a tabela de consulta geográfica pronta, o passo seguinte foi construir o *dataset* principal do projeto. Este processo unificou os dados de múltiplas fontes em uma única tabela estruturada, seguindo as etapas abaixo:

1. **Consolidação dos Dados Transacionais:** Todos os arquivos anuais e suas respectivas planilhas mensais foram lidos e concatenados em um único DataFrame em memória.

2. **Padronização do Esquema:** Os nomes originais das colunas foram mapeados para um formato padronizado e mais legível (*snake_case*), facilitando a manipulação e a análise subsequente.

3. **Limpeza Primária e Filtragem:** Foram aplicadas transformações para garantir a qualidade dos dados. Registros duplicados foram removidos, e os tipos de dados foram convertidos para formatos adequados (numérico para valores e data para transações). Em seguida, foi aplicado um filtro de domínio para reter apenas os imóveis de interesse para esta análise, ou seja, aqueles classificados como "Residencial Vertical"ou "Residencial Horizontal".

4. **Enriquecimento com Dados Geográficos:** O DataFrame consolidado foi então unido (*merged*) com a tabela de consulta geográfica através da coluna `CEP`. Esse passo adicionou as informações de `Latitude` e `Longitude` a cada transação, enriquecendo o *dataset* com as *features* de localização exata.

5. **Armazenamento:** Finalmente, o *dataset* consolidado e enriquecido, contendo aproximadamente 1,83 milhão de registros, foi salvo em disco no formato Apache Parquet (`.parquet`). A escolha desse formato se deu por sua alta eficiência em compressão e velocidade de leitura, otimizando o desempenho nas etapas subsequentes de engenharia de *features* e modelagem.

4.4 ENGENHARIA DE FEATURES

A Engenharia de Features é o processo de utilizar o conhecimento de domínio para criar variáveis que auxiliam os algoritmos de *machine learning* a preverem melhor. O *dataset* consolidado, embora limpo, continha colunas que não representavam diretamente o "valor real"de um imóvel em um determinado momento no tempo. Nesta etapa, portanto, os dados foram transformados para criar *features* mais robustas e informativas.

4.4.1 CRIAÇÃO DA VARIÁVEL-ALVO E FEATURES DERIVADAS

O processo, focou na criação da variável que o modelo deveria prever (a variável-alvo) e de outras *features* relevantes:

1. **Recálculo da Base de Cálculo:** Primeiramente, foi implementada a regra de negócio que define a `base_calculo` do imposto como o valor máximo entre o valor declarado da transação e o valor venal de referência proporcional do imóvel. Este passo garante que a base para o valor do imóvel seja a mais conservadora e próxima da realidade fiscal possível.

2. **Criação da Variável-Alvo:** O passo mais crítico foi o ajuste da `base_calculo` pela inflação. Utilizando o Índice Nacional de Preços ao Consumidor Amplo (IPCA) como deflator, todos os valores de transação, desde 2006, foram trazidos para uma base monetária equivalente ao ano de 2024. O resultado foi a criação da variável-alvo do projeto.

3. **Tratamento de Outliers:** Para aumentar a estabilidade e a performance dos modelos, foram removidos os valores extremos da variável-alvo. Com base na análise de quantis, os 5% de imóveis mais baratos e os 5% mais caros do *dataset* foram excluídos, focando a modelagem na faixa de valores mais representativa do mercado.

4.4.2 CLUSTERIZAÇÃO GEOESPACIAL COM K-MEANS

Para além das coordenadas exatas, hipotetizou-se que a criação de uma *feature* categórica representando macrorregiões (ou "zonas") poderia agregar valor preditivo ao modelo. Para isso, foi utilizada a técnica de clusterização não supervisionada **K-Means**.

O processo foi executado da seguinte forma:

1. **Preparação dos Dados Geoespaciais:** As coordenadas de `Latitude` e `Longitude` de todos os imóveis foram extraídas e padronizadas utilizando um `StandardScaler`. Este passo garante que ambas as dimensões tenham o mesmo peso no cálculo de distância, evitando que uma influencie o algoritmo mais do que a outra.

2. **Aplicação do K-Means:** O algoritmo K-Means foi então treinado para agrupar os imóveis em 8 *clusters* distintos (k=8). O número de *clusters* foi escolhido empiricamente para buscar um bom equilíbrio entre granularidade e representatividade das zonas.

3. **Criação da Feature `localizacao_cluster`:** Ao final do processo, cada imóvel no *dataset* recebeu um rótulo de 0 a 7, correspondente ao *cluster* geográfico ao qual foi atribuído. Esta nova coluna, `localizacao_cluster`, passou a representar a macrorregião do imóvel. A Figura 1 ilustra a distribuição espacial dos *clusters* formados na cidade de São Paulo.

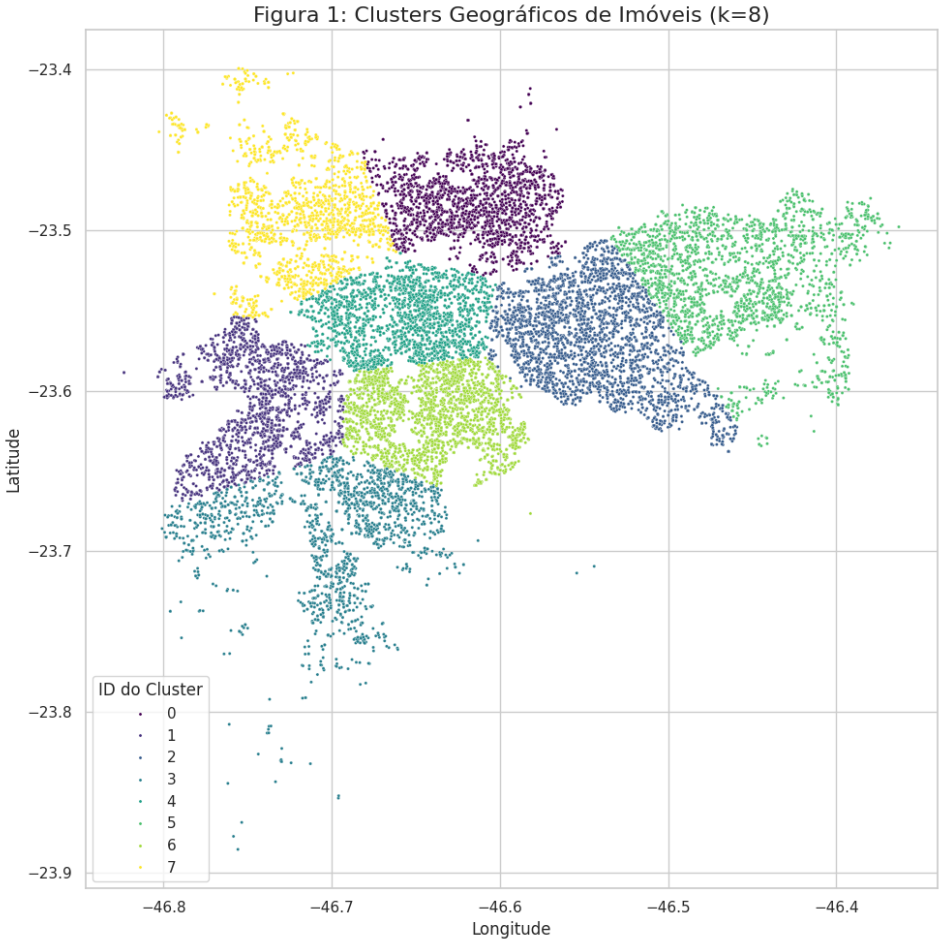

Figura 1 – Mapa de dispersão dos imóveis em São Paulo, coloridos de acordo com os 8 clusters geográficos atribuídos pelo algoritmo K-Means.

Fonte: Autor (2025)

4.5 ANÁLISE EXPLORATÓRIA DE DADOS (AED)

Antes da construção de qualquer modelo preditivo, uma etapa fundamental e indispensável é a Análise Exploratória de Dados (AED). O principal objetivo desta fase foi investigar o conjunto de dados a fim de resumir suas principais características, descobrir padrões e, sobretudo, testar hipóteses iniciais que pudessem orientar as decisões de modelagem.

A análise focou em investigar a distribuição da variável-alvo, as correlações entre os atributos e a evolução dos preços ao longo do tempo. O *insight* mais relevante obtido nesta etapa foi a identificação de comportamentos de preço marcadamente distintos entre os imóveis classificados como "Residencial Vertical"(apartamentos) e "Residencial Horizontal"(casas).

Como ilustrado na Figura 2, a evolução da média da base de cálculo corrigida ao longo dos anos mostra tendências e níveis de valorização diferentes para cada tipo de imóvel. Essa observação levantou a hipótese de que um modelo único teria dificuldade em capturar as particularidades de cada segmento, potencialmente resultando em uma performance preditiva inferior.

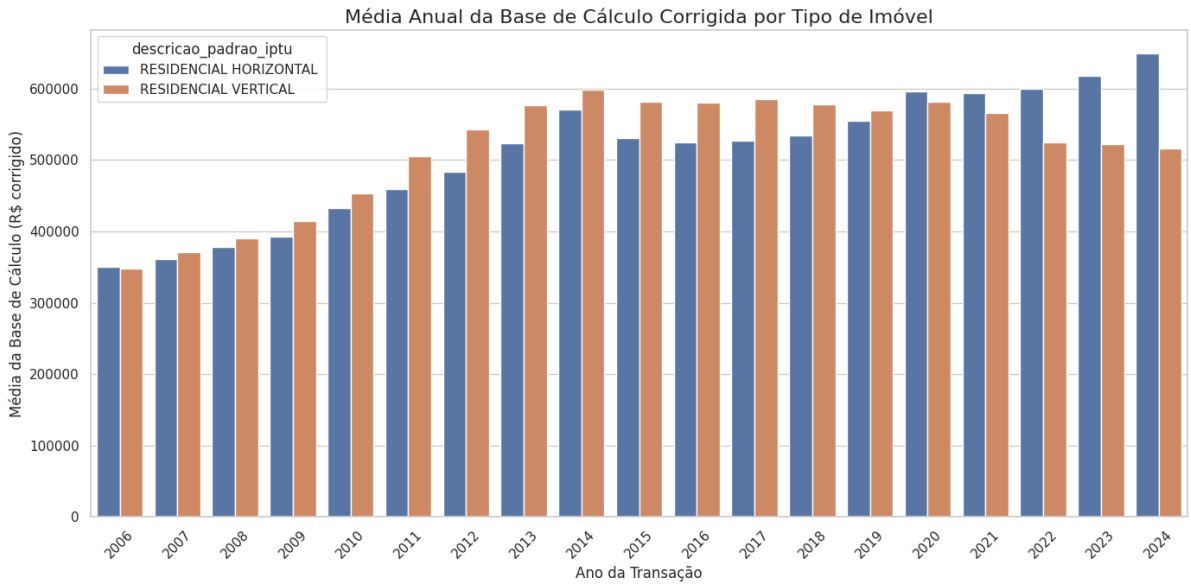

Figura 2 – Evolução da média anual da base de cálculo corrigida, segmentada por tipo de imóvel (Residencial Vertical vs. Horizontal).

Fonte: Autor (2025)

Diante dessa evidência, a principal decisão metodológica derivada da AED foi a de não utilizar um modelo único. Em vez disso, optou-se por segmentar o problema e desenvolver modelos especialistas e independentes: um treinado exclusivamente com dados de imóveis verticais e outro com dados de imóveis horizontais. Esta abordagem, detalhada na seção seguinte, permite que cada modelo se ajuste às dinâmicas de precificação específicas do seu respectivo segmento de mercado.

4.6 PREPARAÇÃO DOS DATASETS FINAIS PARA MODELAGEM

Com base no principal *insight* gerado na Análise Exploratória de Dados, a necessidade de tratar imóveis verticais e horizontais separadamente, a etapa final de preparação consistiu em segmentar e otimizar os *datasets* que seriam efetivamente utilizados para treinar e avaliar os modelos de *machine learning*.

4.6.1 SEGMENTAÇÃO DO DATASET

O primeiro passo foi dividir o *dataset* principal em dois subconjuntos distintos, utilizando a coluna `descricao_padrao_iptu` como critério de separação. Foram criados dois DataFrames independentes:

- Um contendo apenas os registros de imóveis classificados como **Residencial Vertical**.
- Outro contendo apenas os registros de imóveis classificados como **Residencial Horizontal**.

4.6.2 OTIMIZAÇÃO DE FEATURES POR SEGMENTO

Após a segmentação, foram aplicadas otimizações específicas para cada *dataset*, removendo-se colunas que eram irrelevantes ou redundantes para um determinado segmento, a fim de simplificar os modelos e reduzir o ruído. As principais remoções foram:

- **Para o dataset Vertical (apartamentos):** Foram removidas as colunas `area_terreno` e `testada`, uma vez que a área do terreno e a frente do lote são características inerentes ao condomínio como um todo e não possuem valor preditivo direto para uma unidade de apartamento individual.
- **Para o dataset Horizontal (casas):** A coluna `fracao_ideal` foi removida. Para imóveis horizontais, a `area_terreno` é uma *feature* muito mais informativa e direta, tornando a fração ideal redundante.

Adicionalmente, um conjunto de colunas de baixo valor preditivo ou com informações cadastrais (como `n_cadastro`, `matricula_imovel`, `natureza_transacao`, etc.) foi removido de ambos os *datasets* para enxugar a base de dados final.

4.6.3 GERAÇÃO DAS VARIAÇÕES DE FEATURES PARA EXPERIMENTAÇÃO

Para testar as hipóteses deste trabalho e avaliar o impacto das diferentes abordagens de representação geoespacial, foram preparadas, a partir dos *datasets* segmentados, diferentes variações de conjuntos de *features*. Embora os resultados detalhados sejam apresentados no Capítulo 5, a preparação dessas variações é uma etapa metodológica. As duas principais configurações de *features* geoespaciais criadas para os experimentos foram:

1. **Baseline Geoespacial (Apenas Lat/Lon):** Nesta configuração, apenas as colunas numéricas `Latitude` e `Longitude` foram mantidas como representantes da localização do imóvel. A *feature* de *cluster* foi removida.
2. **Híbrida Geoespacial (Lat/Lon + Cluster):** Esta abordagem combinou o dois atributos geoespaciais, mantendo tanto as coordenadas exatas (`Latitude` e `Longitude`) quanto a *feature* categórica de macrorregião (`localizacao_cluster`).

Esses conjuntos de dados segmentados e otimizados foram então salvos em formato Parquet, prontos para a etapa de desenho experimental e modelagem.

4.7 DESENHO EXPERIMENTAL E MODELAGEM

Com os *datasets* devidamente preparados e segmentados, esta seção descreve o desenho experimental adotado para treinar, validar e comparar os modelos de *machine learning*. A metodologia foi estruturada para garantir a robustez dos resultados e uma comparação justa entre as diferentes abordagens.

4.7.1 DIVISÃO TEMPORAL DOS DADOS

Em problemas que envolvem dados com uma dimensão temporal, como transações imobiliárias ao longo dos anos, uma divisão aleatória dos dados para treino e teste pode levar a um problema conhecido como *data leakage* (vazamento de dados). O modelo poderia, por exemplo, ser treinado com dados de 2022 e testado com dados de 2018, aprendendo padrões futuros para prever o passado, o que não reflete um cenário de previsão real.

Para evitar esse viés e garantir que os modelos fossem avaliados em sua capacidade de prever valores em períodos futuros, foi adotada uma estratégia de divisão temporal estrita:

- **Conjunto de Treinamento:** Composto por todas as transações ocorridas entre **2006 e 2015**. Este conjunto foi utilizado para o ajuste dos parâmetros dos modelos.
- **Conjunto de Validação:** Composto pelas transações entre **2016 e 2020**. Foi utilizado durante o desenvolvimento para avaliar e ajustar os modelos sem tocar no conjunto de teste final.
- **Conjunto de Teste:** Composto pelas transações mais recentes, de **2021 a 2024**. Este conjunto foi mantido isolado durante todo o processo e utilizado apenas uma vez para a avaliação final da performance dos modelos campeões.

Para o experimento com Validação Cruzada (CV) usando H2O AutoML , os conjuntos de treino e validação (2006 a 2020) foram combinados em um único *dataset* de treinamento, sobre o qual a CV foi aplicada, mantendo o mesmo conjunto de teste (2021 a 2024) para a avaliação final.

4.7.2 MODELOS PARA ANÁLISE NO TCC

Para a análise central deste trabalho, foram selecionados três algoritmos de regressão supervisionada, representando diferentes níveis de complexidade e interpretabilidade:

- **Regressão Linear e Ridge:** Como modelos de base, foram implementadas a `LinearRegression` e a `Ridge Regression`. Por serem modelos lineares sensíveis à

escala das *features*, uma etapa de pré-processamento com `StandardScaler` foi incluída na *pipeline* para padronizar as variáveis numéricas.

- **Árvore de Decisão:** O `DecisionTreeRegressor` foi escolhido como um modelo não-linear simples e altamente interpretável. Sua estrutura baseada em regras permite entender como as decisões de precificação são tomadas, e sua capacidade de capturar relações complexas serve como um contraponto aos modelos lineares. O hiperparâmetro `max_depth` foi limitado a 10 para evitar sobreajuste.

4.7.3 MODELO AVANÇADO COM AUTOML

Para explorar o potencial máximo dos dados e estabelecer um teto de performance, foi utilizada a plataforma de *Machine Learning* Automatizado **H2O AutoML**. Esta ferramenta automatiza o processo de seleção e otimização de modelos, executando os seguintes passos:

- **Treinamento de Múltiplos Algoritmos:** O AutoML treinou uma variedade de modelos, incluindo Gradient Boosting Machines (GBM), Random Forests (DRF), Redes Neurais e Modelos Lineares Generalizados (GLM).
- **Validação Cruzada (5-fold):** Para garantir uma estimativa de erro robusta, o treinamento foi configurado para usar uma estratégia de Validação Cruzada de 5 *folds* (`nfolds=5`).
- **Otimização e Seleção:** Com um tempo de execução máximo de 600 segundos por experimento, o H2O avaliou os modelos e os classificou em um *leaderboard* com base na métrica de otimização escolhida, o RMSE. O modelo com o melhor desempenho médio na validação cruzada foi então selecionado para a avaliação final.

4.7.4 MÉTRICAS DE AVALIAÇÃO

Para quantificar e comparar a performance dos diferentes modelos, foram utilizadas duas métricas padrão para problemas de regressão:

- **RMSE (Root Mean Squared Error – Raiz do Erro Quadrático Médio):** Mede a magnitude média do erro de previsão, expressa na mesma unidade da variável-alvo (Reais). O RMSE penaliza erros maiores com mais intensidade e foi a principal métrica para otimização e comparação.
- **R^2 (Coeficiente de Determinação):** Mede a proporção da variância no preço do imóvel que é explicada pelo modelo. Varia de $-\infty$ a 1, onde valores mais próximos de 1 indicam um melhor ajuste do modelo aos dados.

5 RESULTADOS E DISCUSSÕES

5.1 INTRODUÇÃO AO CAPÍTULO

Este capítulo apresenta e analisa os resultados obtidos a partir da aplicação da metodologia detalhada no Capítulo 4. Vamos começar com os *insights* da Análise Exploratória de Dados (AED), que foram fundamentais para as decisões estratégicas de modelagem.

Em seguida, será apresentado o desempenho comparativo dos modelos de *machine learning* implementados. A análise iniciará com os modelos de base, mais simples e interpretáveis, e progredirá para os resultados obtidos com a abordagem avançada de *AutoML*, que buscou otimizar a performance preditiva. O objetivo final é consolidar os achados para identificar o modelo e a estratégia de *features* mais eficazes para a precificação de imóveis no complexo mercado de São Paulo.

5.2 ANÁLISE EXPLORATÓRIA DE DADOS E INSIGHTS PARA MODELAGEM

Após a consolidação e o enriquecimento dos dados, foi conduzida uma Análise Exploratória de Dados (AED) para extrair *insights* e orientar a estratégia de modelagem. Esta seção apresenta os resultados mais relevantes dessa análise, que foram fundamentais para as decisões metodológicas subsequentes.

5.2.1 CARACTERIZAÇÃO DA VARIÁVEL-ALVO

O primeiro passo da análise foi investigar a distribuição da variável-alvo, `base_calculo_corrigida_2024`, que representa o valor dos imóveis ajustado pela inflação. A Figura 3 apresenta o histograma e o boxplot desta variável após a aplicação do tratamento de *outliers* (conforme descrito na Seção 4.4.1).

Figura 2: Distribuição da Variável-Alvo

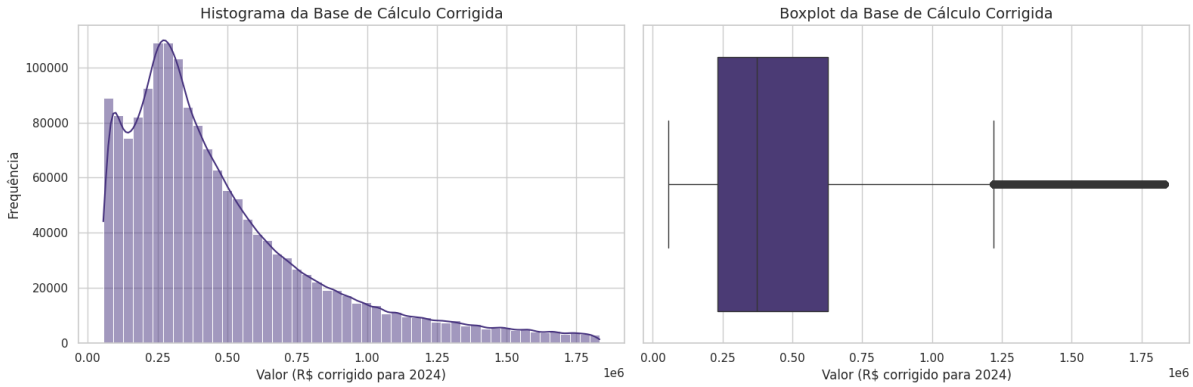

Figura 3 – Histograma e Boxplot da variável-alvo (`base_calculo_corrigida_2024`) após o tratamento de outliers.

Fonte: Autor (2025)

A análise visual revela uma forte assimetria positiva, com uma longa cauda à direita, o que é característico de variáveis de preço. Mesmo após a remoção dos valores extremos, a concentração de imóveis ocorre na faixa de valores mais baixos. Essa distribuição reforça a complexidade do problema e sugere que modelos não-lineares podem ser mais eficazes para capturar a dinâmica de preços.

5.2.2 JUSTIFICATIVA PARA A SEGMENTAÇÃO DOS MODELOS

O *insight* mais impactante da Análise Exploratória surgiu ao se comparar a evolução dos preços entre os dois tipos de imóveis residenciais: "Vertical"(apartamentos) e "Horizontal"(casas). A Figura 4 exibe a evolução da média da base de cálculo corrigida para cada segmento ao longo dos anos.

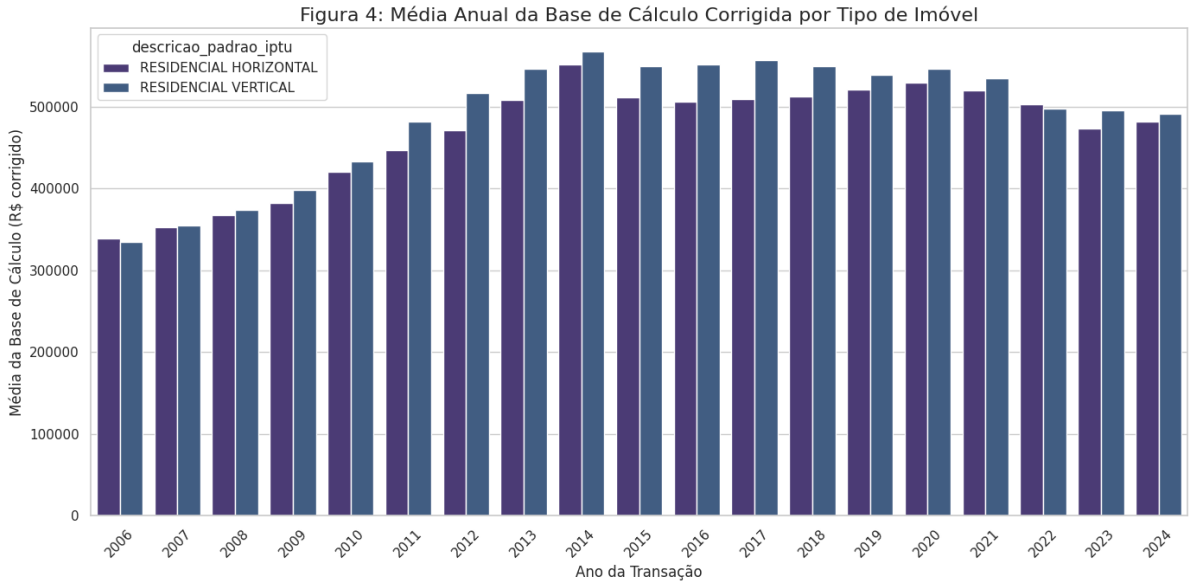

Figura 4 – Evolução da média anual da base de cálculo corrigida por tipo de imóvel.

Fonte: Autor (2025)

Observa-se que, embora ambos os segmentos apresentem uma tendência de valorização, os níveis de preço e as dinâmicas anuais são marcadamente distintos. Para normalizar o efeito do tamanho dos imóveis, que pode variar significativamente entre os dois tipos, a mesma análise foi refeita utilizando o preço por metro quadrado, como mostra a Figura 5.

A análise do preço por metro quadrado não apenas confirma, mas intensifica a observação de que os dois mercados se comportam de maneira diferente. Essa evidência visual fundamentou a principal decisão estratégica da metodologia deste trabalho, que foi abandonar a abordagem de um modelo único e, em vez disso, **segmentar o problema**. Foram desenvolvidos modelos preditivos especialistas, um treinado exclusivamente com dados de imóveis verticais e outro com dados de imóveis horizontais, para que cada um pudesse se ajustar às particularidades de seu respectivo mercado.

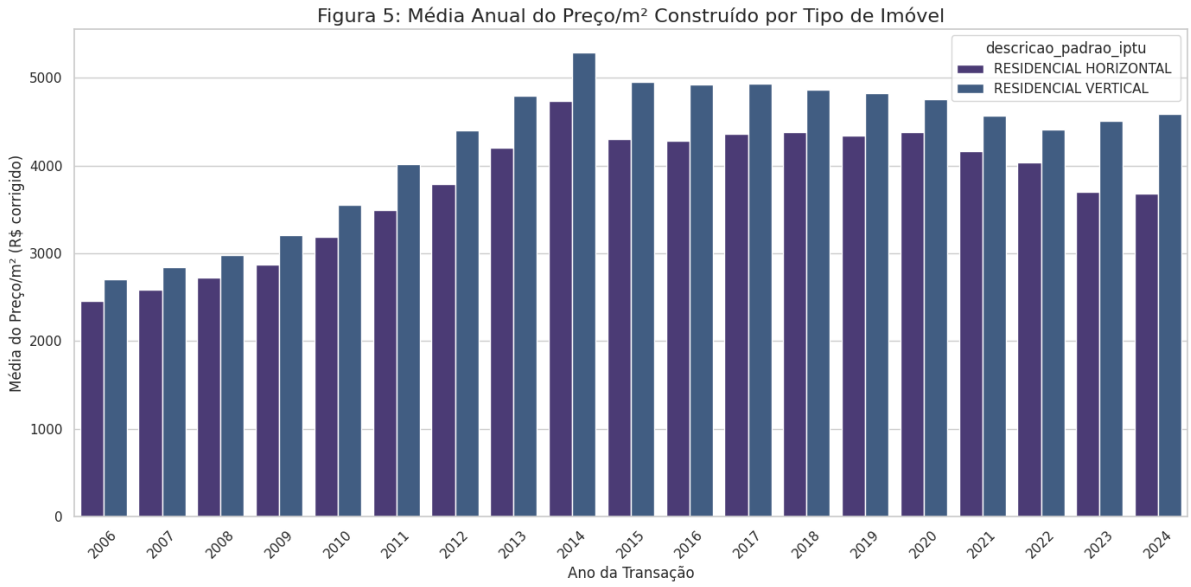

Figura 5 – Evolução da média anual do preço por metro quadrado construído por tipo de imóvel.

Fonte: Autor (2025)

5.3 DESEMPENHO DOS MODELOS INTERPRETÁVEIS (BASELINE)

Após a análise exploratória, o passo seguinte foi estabelecer uma linha de base (*baseline*) de performance, utilizando os modelos mais simples e interpretáveis descritos na metodologia. Estes experimentos, foram configurados utilizando apenas as coordenadas de latitude e longitude como *features* preditoras, a fim de isolar e avaliar a capacidade de cada algoritmo em modelar a relação entre a localização e o preço do imóvel.

Os resultados obtidos no conjunto de teste para cada modelo e segmento estão consolidados na Tabela 2.

Tabela 2 – Resultados de performance dos modelos simples no conjunto de teste.

Tipo de Imóvel	Modelo	Estratégia Geo.	R^2 (Teste)	RMSE (Teste)
Vertical	Regressão Linear	Apenas Lat/Lon	-3.3887	R$ 775.984,49
Vertical	Regressão Ridge	Apenas Lat/Lon	-3.3887	R$ 775.983,58
Vertical	Árvore de Decisão	Apenas Lat/Lon	0.6346	R$ 223.912,43
Horizontal	Regressão Linear	Apenas Lat/Lon	-1.6360	R$ 542.616,05
Horizontal	Regressão Ridge	Apenas Lat/Lon	-1.6360	R$ 542.614,65
Horizontal	Árvore de Decisão	Apenas Lat/Lon	-0.0329	R$ 339.665,36

Fonte: Autor (2025)

5.3.1 DISCUSSÃO DOS RESULTADOS DE BASELINE

A Tabela 2 revela dois pontos fundamentais para a compreensão do problema.

Primeiramente, observa-se o completo insucesso dos modelos lineares (`LinearRegression` e `Ridge`). Para ambos os segmentos, eles produziram um coeficiente de determinação (R^2) negativo, o que indica que seu desempenho foi pior do que um modelo ingênuo que simplesmente previsse o preço médio para todos os imóveis. Este resultado evidencia a principal limitação desses algoritmos: sua incapacidade de capturar as complexas relações não-lineares que existem entre a localização geográfica e o valor dos imóveis, conforme discutido na Seção 4.7.2. A performance da Regressão Ridge, virtualmente idêntica à da Linear, sugere que a multicolinearidade não era o problema dominante, mas sim a natureza não-linear dos dados.

Em contrapartida, o modelo de **Árvore de Decisão** (`DecisionTreeRegressor`) apresentou um resultado significativamente superior, especialmente para o segmento de imóveis verticais, onde alcançou um R^2 de 0.6346 no conjunto de teste. Este desempenho demonstra sua capacidade de modelar padrões não-lineares ao particionar o espaço das coordenadas geográficas em diferentes regiões de preço, superando a principal limitação dos modelos lineares. Para o segmento horizontal, embora o R^2 ainda seja negativo (-0.0329), a drástica redução no RMSE em comparação com os modelos lineares (de R$ 542 mil para R$ 339 mil) indica que o modelo conseguiu aprender padrões relevantes, ainda que de forma insuficiente.

Esta análise valida a hipótese de que modelos não-lineares são mais adequados para este problema e estabelece um *baseline* de performance claro: um modelo minimamente eficaz deve superar o desempenho da Árvore de Decisão.

5.4 DESEMPENHO DO MODELO AVANÇADO COM H2O AUTOML

Para explorar o potencial máximo dos dados e estabelecer um teto de performance, foi empregada uma abordagem de *Machine Learning* Automatizado (AutoML) com a plataforma H2O. Conforme detalhado na metodologia (Seção 4.7.3), esta etapa utilizou uma robusta estratégia de Validação Cruzada de 5 *folds* para treinar e avaliar uma gama de algoritmos nos *datasets* segmentados. Os resultados são apresentados a seguir.

5.4.1 SELEÇÃO DO MELHOR MODELO PARA IMÓVEIS VERTICAIS

Para o segmento de imóveis verticais (apartamentos), o H2O AutoML treinou diversos modelos, incluindo Gradient Boosting Machines (GBM), Random Forests (DRF) e Ensembles. A Tabela 3 apresenta o *leaderboard* final, classificando os modelos com base em sua performance média de RMSE na validação cruzada.

O *leaderboard* demonstra a superioridade dos modelos de *ensemble*, que combinam as previsões de múltiplos algoritmos. O modelo campeão, `StackedEnsemble_BestOfFamily_1`, alcançou o menor erro médio na validação cruzada. Para obter a estimativa final de sua capacidade de generalização, este modelo foi então avaliado no conjunto de teste (dados de 2021 a 2024),

Tabela 3 – Leaderboard do H2O AutoML para o segmento de Imóveis Verticais, classificado por RMSE na validação cruzada.

Model ID	RMSE (CV)
StackedEnsemble_BestOfFamily_1	R$ 184.424,81
StackedEnsemble_AllModels_1	R$ 185.854,28
StackedEnsemble_BestOfFamily_2	R$ 185.854,28
GBM_1	R$ 193.846,63
XGBoost_1	R$ 196.322,29
DRF_1	R$ 205.772,17

Fonte: Autor (2025)

que não foi utilizado durante o treinamento ou a validação. A performance final foi:

- **RMSE no Teste:** R$ 184.424,82
- **R^2 no Teste:** 0.7521

Este resultado representa o melhor desempenho preditivo alcançado para o segmento de imóveis verticais.

5.4.2 SELEÇÃO DO MELHOR MODELO PARA IMÓVEIS HORIZONTAIS

O mesmo processo foi repetido para o segmento de imóveis horizontais (casas). A Tabela 4 exibe o *leaderboard* resultante.

Tabela 4 – Leaderboard do H2O AutoML para o segmento de Imóveis Horizontais, classificado por RMSE na validação cruzada.

Model ID	RMSE (CV)
GBM_4	R$ 302.382,49
GBM_2	R$ 303.057,54
GBM_1	R$ 305.347,50
GBM_3	R$ 305.351,60
XGBoost_2	R$ 306.232,66
StackedEnsemble_BestOfFamily_1	R$ 312.228,47

Fonte: Autor (2025)

Neste caso, os modelos baseados em Gradient Boosting (GBM) dominaram as primeiras posições, superando os *ensembles*. O melhor modelo, `GBM_4`, foi selecionado e avaliado no conjunto de teste, obtendo a seguinte performance final:

- **RMSE no Teste:** R$ 302.382,49

Este valor representa o menor erro médio de previsão para o segmento de casas, demonstrando uma melhora substancial em relação aos modelos de *baseline* e estabelecendo o melhor resultado do projeto para este tipo de imóvel.

5.5 ANÁLISE COMPARATIVA E DISCUSSÃO FINAL DOS RESULTADOS

Após a apresentação individual dos resultados, esta seção realiza uma análise comparativa direta entre a melhor abordagem simples (Árvore de Decisão) e a melhor abordagem avançada (H2O AutoML) para cada segmento de imóvel. O objetivo é sintetizar os achados e extrair as conclusões centrais do projeto.

A Tabela 5 consolida a performance final dos modelos campeões de cada abordagem, permitindo uma comparação direta do erro médio de previsão (RMSE) no conjunto de teste.

Tabela 5 – Resumo comparativo de performance dos melhores modelos no conjunto de teste.

Tipo de Imóvel	Melhor Modelo Simples (RMSE)	Melhor Modelo AutoML (RMSE)
Vertical (Apartamentos)	Árvore de Decisão (R$ 223.912,43)	StackedEnsemble (R$ 184.424,82)
Horizontal (Casas)	Árvore de Decisão (R$ 339.665,36)	GBM (R$ 302.382,49)

Fonte: Autor (2025)

A análise da tabela revela que, para ambos os segmentos, a abordagem avançada com H2O AutoML resultou em uma performance preditiva superior. Para os imóveis verticais, o modelo `StackedEnsemble` conseguiu uma redução de aproximadamente **17,6%** no erro médio em comparação com a Árvore de Decisão. Para os imóveis horizontais, o ganho foi igualmente expressivo, com o modelo `GBM` reduzindo o erro médio em cerca de **11%**. Esses resultados validam a hipótese de que algoritmos mais complexos e a otimização automática de hiperparâmetros são capazes de extrair mais informações dos dados e gerar previsões mais acuradas.

5.5.1 ANÁLISE DA IMPORTÂNCIA DAS FEATURES

Além de avaliar a acurácia preditiva dos modelos, é fundamental analisar quais características (*features*) foram mais influentes para as suas previsões. Uma análise da importância das *features*, extraída do modelo `RandomForestRegressor` (que serve como um bom proxy para os modelos baseados em árvores), revelou uma hierarquia clara de preditores para o segmento de imóveis verticais:

- **Tamanho do Imóvel:** A `area_construida` foi, de forma conclusiva, a variável mais preditiva, corroborando o *insight* da análise de correlação.

- **Idade e Localização Exata:** Em seguida, a idade do imóvel (`acc_iptu`) e as coordenadas geográficas (`Longitude` e `Latitude`) surgiram como os fatores de maior relevância, destacando a tríade "tamanho, idade e localização"como o pilar da precificação.
- **Outras Características:** Variáveis como a `testada` e o padrão do imóvel (ex: `descricao_padrao_iptu_RESIDENCIAL VERTICAL`) também contribuíram para as previsões, indicando que o modelo foi capaz de capturar nuances entre diferentes subcategorias de apartamentos e condomínios.

Uma análise similar para o modelo horizontal (não detalhada aqui) apontou uma alta importância para a *feature* `area_terreno`, validando a decisão metodológica de tratar os segmentos de casas e apartamentos separadamente.

A presença significativa das *features* `Latitude` e `Longitude` nos modelos finais do H2O (que também utilizam a *feature* de *cluster*) sugere que a combinação de informações de localização em diferentes níveis de granularidade, a coordenada exata e a zona geográfica, foram a estratégia mais eficaz para maximizar o poder preditivo.

6 CONSIDERAÇÕES FINAIS

6.1 RECAPITULAÇÃO DO PROBLEMA E DA METODOLOGIA

O presente trabalho de conclusão de curso propôs-se a enfrentar um desafio central do mercado imobiliário de São Paulo: a acentuada assimetria de informações e a subjetividade inerente aos métodos tradicionais de avaliação de imóveis. Conforme discutido, essa lacuna informacional gera desequilíbrios e ineficiências, dificultando a tomada de decisão de todos os agentes envolvidos, desde compradores e vendedores até investidores e o poder público.

Diante deste cenário, o objetivo principal do projeto foi desenvolver e avaliar um modelo de preços hedônicos, baseado em técnicas de *Machine Learning*, capaz de gerar precificações de imóveis mais acuradas e transparentes para o mercado residencial paulistano. A hipótese central era de que a aplicação de algoritmos avançados sobre uma base de dados de transações reais permitiria a criação de uma ferramenta robusta para a redução dessa assimetria informacional.

Para alcançar tal objetivo, foi implementada uma metodologia de ponta a ponta. O processo iniciou-se com a obtenção e validação de uma vasta base de dados do ITBI, contendo registros de 2006 a 2024. Subsequentemente, os dados foram consolidados e enriquecidos com *features* geoespaciais, e uma variável-alvo robusta foi criada através do ajuste dos valores pela inflação. Uma análise exploratória revelou dinâmicas de preço distintas entre imóveis verticais e horizontais, o que fundamentou a decisão estratégica de segmentar a modelagem. Por fim, foram conduzidos experimentos comparativos, avaliando desde modelos interpretáveis, como a Árvore de Decisão, até uma abordagem avançada com *AutoML* e Validação Cruzada, a fim de determinar a solução de melhor performance.

6.2 SÍNTESE E DISCUSSÃO DOS PRINCIPAIS ACHADOS

A execução da metodologia proposta permitiu validar a hipótese central do trabalho e gerou *insights* relevantes sobre a precificação de imóveis no mercado paulistano. Os principais achados são discutidos a seguir.

Primeiramente, os resultados confirmam a hipótese de que a aplicação de técnicas de *Machine Learning* sobre uma base de dados do ITBI permite o desenvolvimento de modelos de preços hedônicos com acurácia preditiva. A capacidade dos modelos de aprender padrões a partir dos dados históricos e generalizar para dados não vistos valida esta abordagem como uma ferramenta eficaz para a análise imobiliária.

Em segundo lugar, a análise comparativa dos modelos de base demonstrou a clara superioridade dos algoritmos não-lineares para este problema. Conforme apresentado na Tabela 2, os modelos de Regressão Linear e Ridge obtiveram um R^2 negativo no conjunto de teste, indicando um desempenho inferior a um modelo que simplesmente previsse o valor médio para todos os imóveis. Em contraste, o modelo de Árvore de Decisão, por sua capacidade de modelar

relações complexas, alcançou um R² de 0.6346 para o segmento vertical, provando ser uma abordagem muito mais adequada para capturar a dinâmica não-linear do mercado imobiliário.

O terceiro achado fundamental foi a eficácia da estratégia de segmentação de mercado. A performance consistentemente diferente entre os modelos para imóveis verticais e horizontais, observada em todos os experimentos, valida a decisão metodológica de criar modelos especialistas. Esta abordagem permitiu que cada modelo se ajustasse às dinâmicas de precificação e à relevância de *features* específicas de seu nicho, como a importância da `area_terreno` para casas, que é irrelevante para apartamentos.

Finalmente, a aplicação de uma abordagem avançada com H2O AutoML e Validação Cruzada estabeleceu o teto de performance do projeto, superando significativamente o *baseline* da Árvore de Decisão. Conforme resume a Tabela 5, o modelo final alcançou uma performance notavelmente superior:

- Para **imóveis verticais**, o modelo `StackedEnsemble` reduziu o erro médio de previsão (RMSE) em aproximadamente **17,6%** em relação à Árvore de Decisão, alcançando um R² final de 0.7521.
- Para **imóveis horizontais**, o modelo `GBM` obteve uma redução de **11%** no RMSE e elevou o R² de um valor negativo para 0.5618, demonstrando uma melhoria substancial na capacidade de explicação e predição para este segmento mais desafiador.

É fundamental, neste ponto, analisar a disparidade de performance entre os dois segmentos. O R² de 0.7521 para imóveis verticais indica um modelo robusto, enquanto o R² de 0.5618 para imóveis horizontais, embora muito superior ao *baseline*, é visivelmente mais baixo. Esta diferença não deve ser interpretada como uma falha do algoritmo, mas sim como um reflexo direto da natureza dos dados. Conforme detalhado na Seção de Limitações (Seção 6.4.1), o segmento "Residencial Horizontal"(casas) é inerentemente mais heterogêneo que o "Vertical"(apartamentos). Mais importante, a base do ITBI carece de *features* estruturais críticas para casas, como número de dormitórios, banheiros, vagas de garagem e padrão de acabamento, que são mais padronizadas em apartamentos (onde a `area_construida` e a localização já explicam grande parte do valor). Portanto, o teto de performance inferior do modelo horizontal é uma consequência direta da ausência dessas variáveis explicativas essenciais.

6.3 CONTRIBUIÇÕES DO TRABALHO

Este trabalho oferece contribuições significativas em três domínios principais: metodológico, prático e acadêmico. Cada uma delas é detalhada a seguir.

6.3.1 CONTRIBUIÇÃO METODOLÓGICA

A principal contribuição metodológica desta pesquisa é a construção de uma *pipeline* de ciência de dados de ponta a ponta, documentada e reprodutível, para o tratamento e a modelagem de uma base de dados pública, complexa e de grande escala. Ao sistematizar o processo desde a validação de mais de 2,3 milhões de registros brutos do ITBI até a implementação de técnicas avançadas de engenharia de *features*, como a clusterização geoespacial com K-Means, o trabalho oferece um roteiro metodológico robusto que pode ser adaptado e replicado para análises similares em outros grandes centros urbanos.

6.3.2 CONTRIBUIÇÃO PRÁTICA

Do ponto de vista prático, a contribuição mais relevante é o desenvolvimento de modelos preditivos que representam uma ferramenta eficaz para mitigar a assimetria de informações no mercado imobiliário paulistano. Conforme exposto no Problema de Pesquisa, a subjetividade e a falta de transparência são desafios recorrentes neste setor. Os modelos finais desenvolvidos, especialmente os gerados pelo H2O AutoML, oferecem uma estimativa de valor baseada em dados históricos de transações reais, constituindo uma referência objetiva que pode beneficiar diretamente:

- **Compradores e Vendedores:** Fornecendo um valor de referência para negociações mais justas e informadas.
- **Agentes do Mercado:** Auxiliando corretores, avaliadores e instituições financeiras em seus processos de precificação e análise de risco.
- **Poder Público:** Oferecendo *insights* para a formulação de políticas urbanas e para a atualização de plantas de valores fiscais.

6.3.3 CONTRIBUIÇÃO ACADÊMICA

No campo acadêmico, este trabalho contribui ao validar a aplicação de técnicas avançadas de *Machine Learning*, como *ensembles* e AutoML com Validação Cruzada, para a precificação hedônica no contexto específico do mercado imobiliário brasileiro. A pesquisa demonstra empiricamente a superioridade dessas abordagens em relação a modelos mais simples, quantificando o ganho de performance e estabelecendo um *baseline* robusto para futuras investigações na área. Além disso, a análise comparativa entre diferentes estratégias de *features* geoespaciais (coordenadas exatas vs. *clusters*) oferece um *insight* valioso sobre a interação entre a complexidade das *features* e a capacidade de cada algoritmo, enriquecendo a literatura sobre modelagem de dados espaciais.

6.4 LIMITAÇÕES DO ESTUDO

Apesar dos resultados robustos e das contribuições relevantes, é fundamental reconhecer as limitações inerentes a este estudo. A identificação transparente dessas limitações não diminui o valor do trabalho, mas, ao contrário, contextualiza os achados e orienta futuras investigações na área.

6.4.1 LIMITAÇÃO DAS FEATURES DISPONÍVEIS

A principal limitação da pesquisa reside na granularidade das *features* disponíveis na base de dados do ITBI. Embora esta fonte de dados seja extremamente valiosa por conter os valores reais de transação, ela carece de atributos estruturais detalhados que são sabidamente importantes na precificação hedônica. Informações como o número de dormitórios, banheiros, suítes, vagas de garagem, e a presença de áreas de lazer no condomínio (como piscina ou academia) não estavam disponíveis. A ausência dessas variáveis preditivas certamente impôs um teto à performance dos modelos, sendo uma provável explicação para a dificuldade em se obter um R^2 mais elevado para o segmento de imóveis horizontais, que é naturalmente mais heterogêneo em suas características.

6.4.2 AUSÊNCIA DE VARIÁVEIS DE VIZINHANÇA

Outra limitação relevante foi a não inclusão de variáveis contextuais ou de vizinhança. Fatores como a qualidade das escolas no entorno, os índices de segurança pública (criminalidade), a proximidade a parques, hospitais e estações de metrô, e o nível socioeconômico médio da região são elementos que influenciam de forma significativa a percepção de valor de um imóvel. A coleta e a integração dessas *features* demandariam um esforço considerável de cruzamento com outras bases de dados (públicas ou privadas), o que fugiu ao escopo deste trabalho. A incorporação futura dessas variáveis representa uma oportunidade clara para o aprimoramento dos modelos.

6.4.3 INTERPRETABILIDADE DOS MODELOS AVANÇADOS

Por fim, é importante notar que os modelos de melhor performance, o `StackedEnsemble` e o `GBM`, são inerentemente complexos e operam como "caixas-pretas"(*black boxes*). Embora a análise de importância de *features* (Seção 5.5.1) ofereça uma visão geral e agregada sobre quais variáveis são mais influentes, ela não permite uma explicação detalhada de como uma previsão individual é feita. A dificuldade em se interpretar as decisões internas desses algoritmos representa um *trade-off* comum em *Machine Learning* entre a acurácia do modelo e a sua transparência.

6.5 SUGESTÕES PARA TRABALHOS FUTUROS

Com os objetivos deste trabalho alcançados, uma direção natural e promissora para a continuidade da pesquisa reside na incorporação de uma análise de séries temporais. Enquanto o presente estudo focou no desenvolvimento de um modelo de precificação hedônico, uma abordagem futura poderia se concentrar em modelar a evolução temporal dos preços dos imóveis.

Isso permitiria não apenas prever valores, mas também investigar tendências, sazonalidades e o impacto de variáveis macroeconômicas, como a taxa de juros e a inflação setorial, sobre o mercado imobiliário. Tal análise representaria uma valiosa expansão do escopo atual, transitando de um modelo de avaliação estático para um modelo dinâmico de previsão de mercado.

REFERÊNCIAS

BORGES, W. V.; SALVIATO, R. B.; GOES, T. H. M. Machine learning e a precificação de imóveis: um estudo comparativo entre modelos de preços hedônicos. In: **XXVII Seminários em Administração (SemeAd)**. [S.l.: s.n.], 2024.

CHAPMAN, P. et al. **CRISP-DM 1.0: Step-by-step data mining guide**. [S.l.], 2000.

CHIN, T.-L.; CHAU, K. W. A critical review of literature on the hedonic price model. **International Journal for Housing Science and Its Applications**, v. 27, n. 2, p. 145–165, 2003.

GAZOLA, S. **Construção de um Modelo de Regressão para Avaliação de Imóveis**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Santa Catarina, 2002.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. 2nd. ed. [S.l.]: O'Reilly Media, 2019.

LANCASTER, K. J. A new approach to consumer theory. **Journal of Political Economy**, v. 74, n. 2, p. 132–157, 1966.

LIMSOMBUNCHAI, V.; GAN, C.; LEE, M. House price prediction: Hedonic price model vs. artificial neural network. **American Journal of Applied Sciences**, v. 1, n. 3, p. 193–201, 2004.

MATTA, T. A. Monografia (Engenharia de Produção), **Avaliação do Valor de Imóveis por Analise de Regressão: Um Estudo de Caso para a Cidade de Juiz de Fora**. 2007.

ROSEN, S. Hedonic prices and implicit markets: Product differentiation in pure competition. **Journal of Political Economy**, v. 82, n. 1, p. 34–55, 1974.

SIEBRA, N. V. A. **Aprimoramento de Modelo de Regressão Linear para Precificação e Investimentos Imobiliários em Fortaleza**. Dissertação (Dissertação de Mestrado) — Universidade Federal do Ceará, 2024.

SILVA, C. H. d. S. Monografia (Bacharelado em Estatística), **Modelo de Regressão Múltipla para Avaliação de Imóveis na Cidade de Aracaju - SE**. 2016.

SILVA, G. H. P. d. Trabalho de Conclusão de Curso (Engenharia Civil), **Modelos de Aprendizagem de Máquina para Precificação de Imóveis na Cidade de Fortaleza**. 2019.

TEODORO, L. d. A.; KAPPEL, M. A. A. Modelo hedônico para estimação do valor de imóveis: Aplicação em nova friburgo-rj. **Vetor, Rio Grande**, v. 30, n. 1, p. 28–37, 2020.

ZULKIFLEY, N. H. et al. House price prediction using a machine learning model: A survey of literature. **International Journal of Modern Education and Computer Science**, v. 12, n. 6, p. 46–54, 2020.