UNIVERSIDADE FEDERAL DO TOCANTINS
CAMPUS UNIVERSITÁRIO DE PALMAS
PROGRAMA DE PÓS-GRADUAÇÃO EM DESENVOLVIMENTO REGIONAL

RAFAEL ALVES AMORIM

APRENDIZAGEM DE MÁQUINA E MODELAGEM ESTATÍSTICA NA AVALIAÇÃO DE IMÓVEIS URBANOS: A INFLUÊNCIA DE VARIÁVEIS AMBIENTAIS EM PALMAS-TO

Palmas, TO

2025

Rafael Alves Amorim

Aprendizagem de Máquina e Modelagem Estatística na Avaliação de Imóveis Urbanos: A Influência de Variáveis Ambientais em Palmas-TO

Tese apresentada ao Programa de Pós-Graduação em Desenvolvimento Regional, da Universidade Federal do Tocantins, como requisito para a obtenção do título de Doutor em Desenvolvimento Regional.

Orientador (a): Prof. Dr. Adriano Nascimento da Paixão

Palmas, TO

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

A524a Amorim, Rafael Alves.

Aprendizagem de Máquina e Modelagem Estatística na Avaliação de Imóveis Urbanos: A Influência de Variáveis Ambientais em Palmas-TO. / Rafael Alves Amorim. – Palmas, TO, 2025.

110 f.

Tese (Doutorado) - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Pós-Graduação (Doutorado) em Desenvolvimento Regional, 2025.

Orientador: Adriano Nascimento da Paixão

1. avaliação de imóveis. 2. variáveis ambientais. 3. planejamento urbano. 4. aprendizagem de Máquina. I. Título

CDD 338.9

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Rafael Alves Amorim

Aprendizagem de Máquina e Modelagem Estatística na Avaliação de Imóveis Urbanos: A Influência De Variáveis Ambientais em Palmas-TO

Tese apresentada ao Programa de Pós-Graduação em Desenvolvimento Regional, da Universidade Federal do Tocantins, como requisito para a obtenção do título de Doutor em Desenvolvimento Regional.

Tese defendida e aprovada em 04 de setembro 2025.

Banca Examinadora:

gov.br Documento assinado digitalmente
ADRIANO NASCIMENTO DA PAIXAO
Data: 17/09/2025 09:19:07-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Adriano Nascimento da Paixão (orientador), PPGDR - UFT

gov.br Documento assinado digitalmente
ADRIANO FIRMINO VALDEVINO DE ARAUJO
Data: 17/09/2025 17:13:03-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Adriano Firmino Valdevino de Araújo, UFPB

gov.br Documento assinado digitalmente
FLAVIO ROLDAO DE CARVALHO LELIS
Data: 17/09/2025 09:48:06-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Flávio Roldão de Carvalho Lelis, IFG

gov.br Documento assinado digitalmente
NILTON MARQUES DE OLIVEIRA
Data: 18/09/2025 16:03:54-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Nilton Marques de Oliveira, PPGDR - UFT

gov.br Documento assinado digitalmente
WALDECY RODRIGUES
Data: 22/09/2025 11:47:39-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Waldecy Rodrigues, PPGDR - UFT

AGRADECIMENTOS

Agradeço primeiramente a Jesus Cristo, por ter me sustentado em todos os momentos desta caminhada. Foi Ele quem me deu forças nos dias difíceis e serenidade nas horas de incerteza.

À minha esposa, Marianna Dornelles, minha companheira de vida, pelo amor, paciência e apoio incondicional em cada etapa deste percurso. Aos meus filhos, Bento e Dante, que me enchem de alegria e são a razão maior para que eu siga sempre em frente, buscando ser alguém melhor a cada dia.

Aos meus pais, Maristélia Alves e Domingos Araújo, deixo minha gratidão eterna pelos valores, ensinamentos e pela dedicação incansável que moldaram o que sou. Aos meus irmãos, Gustavo e Gabriel, agradeço pela amizade, pelo incentivo e por estarem sempre presentes.

Ao meu orientador, Professor Dr. Adriano Nascimento da Paixão, manifesto meu profundo reconhecimento pela confiança, pela paciência e pela firmeza de suas orientações, que foram essenciais para a construção e amadurecimento desta tese.

Também registro meu agradecimento aos amigos da Superintendência do Patrimônio da União no Tocantins, pela amizade, pelo incentivo e pela parceria no dia a dia, que fizeram toda diferença nessa trajetória.

Ao Programa de Pós-graduação em Desenvolvimento Regional da Universidade Federal do Tocantins, aos servidores, professores e colegas de turma, agradeço pelo apoio constante e pela convivência enriquecedora ao longo desses anos de estudo e pesquisa.

A cada pessoa que, de alguma forma, contribuiu com este trabalho, deixo aqui o meu sincero muito obrigado.

RESUMO

Esta pesquisa analisou a influência de variáveis ambientais nos preços de imóveis urbanos na cidade de Palmas-TO, com ênfase na proximidade ao Parque Cesamar e ao lago da Usina Hidrelétrica de Lajeado, por meio da aplicação de modelos estatísticos tradicionais e técnicas de Aprendizagem de Máquina (AM). A motivação central está na busca por modelos capazes de captar com maior precisão o efeito das amenidades ambientais sobre o mercado imobiliário, contribuindo com a incorporação do valor ambiental na análise econômica e, eventualmente, em políticas públicas, como a cobrança do IPTU. A investigação também se insere no campo do planejamento urbano, ao reconhecer que a valorização imobiliária está diretamente vinculada à localização, acessibilidade e disponibilidade de serviços e infraestruturas urbanas, além da presença de ativos ambientais. Assim, compreender como esses elementos influenciam os preços dos imóveis permite qualificar as estratégias de ordenamento territorial, subsidiando políticas públicas voltadas à equidade socioespacial e à sustentabilidade da expansão urbana de Palmas. Foram estimados Modelos Lineares (ML) e Lineares Generalizados (MLG), além de quatorze algoritmos de AM, com o objetivo de comparar o desempenho preditivo de cada abordagem. As variáveis explicativas incluíram atributos físicos, de localização, socioeconômicos e ambientais, sendo estas últimas representadas por distâncias contínuas e *dummies* de proximidade a ativos ambientais. Os resultados indicam que os modelos de AM superaram os modelos tradicionais quanto à capacidade preditiva e à sensibilidade na identificação dos efeitos das variáveis ambientais sobre os preços dos imóveis. Esses modelos demonstraram maior robustez e flexibilidade para capturar relações complexas e não lineares, ampliando as possibilidades analíticas no campo da avaliação imobiliária. Dentre os algoritmos testados, destacaram-se *o Bagged Trees, o Random Forest e o Boosted Trees*, que se revelaram particularmente eficazes na modelagem de relações não lineares e na identificação de interações complexas entre os atributos. Os resultados reforçam que a aplicação de técnicas de Aprendizagem de Máquina configura uma abordagem metodológica consistente e inovadora para a avaliação de imóveis urbanos, especialmente quando se busca incorporar variáveis ambientais à dinâmica de formação de preços no mercado imobiliário, superando as limitações dos modelos tradicionais.

Palavras-chave: avaliação de imóveis. variáveis ambientais. planejamento urbano. aprendizagem de Máquina.

ABSTRACT

This research analyzed the influence of environmental variables on urban property prices in the city of Palmas-TO, with emphasis on the proximity to Cesamar Park and the reservoir of the Lajeado Hydroelectric Plant, through the application of traditional statistical models and Machine Learning (ML) techniques. The central motivation lies in the search for models capable of more accurately capturing the effect of environmental amenities on the real estate market, contributing to the incorporation of environmental value into economic analysis and, eventually, into public policies, such as the collection of property tax (IPTU). The investigation also falls within the field of urban planning, by recognizing that real estate appreciation is directly linked to location, accessibility, and the availability of urban services and infrastructure, in addition to the presence of environmental assets. Thus, understanding how these elements influence property prices allows for better-informed territorial planning strategies, supporting public policies aimed at socio-spatial equity and the sustainability of Palmas' urban expansion. Linear Models (LM) and Generalized Linear Models (GLM), as well as fourteen ML algorithms, were estimated with the purpose of comparing the predictive performance of each approach. The explanatory variables included physical, locational, socioeconomic, and environmental attributes, the latter represented by continuous distances and proximity dummies to environmental assets. The results indicate that ML models outperformed traditional models in terms of predictive capacity and sensitivity in identifying the effects of environmental variables on property prices. These models demonstrated greater robustness and flexibility in capturing complex and nonlinear relationships, expanding analytical possibilities in the field of real estate valuation. Among the tested algorithms, Bagged Trees, Random Forest, and Boosted Trees stood out as particularly effective in modeling nonlinear relationships and identifying complex interactions among attributes. The findings reinforce that the application of Machine Learning techniques constitutes a consistent and innovative methodological approach to urban property valuation, especially when seeking to incorporate environmental variables into the price formation dynamics of the real estate market, overcoming the limitations of traditional models.

Keywords: real estate valuation. environmental variables. urban planning. machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
EC	Excedente de Consumidor
IPTU	Imposto Predial e Territorial Urbano
ITBI	Imposto sobre a Transmissão de Bens Imóveis
MLG	Modelo Linear Generalizado
MPH	Método dos preços Hedônicos
NBR	Norma Brasileira
PVG	Planta de valores genéricos
UHE	Usina Hidrelétrica de Lajeado
VE	Valor de existência
VERA	Valor econômico do Recurso Ambiental
VNU	Valor de não-uso
VO	Valor de opção
VU	Valor de uso
VUD	Valor de uso direto
VUI	Valor de uso indireto

SUMÁRIO

1 INTRODUÇÃO

O acelerado processo de ocupação dos centros urbanos está viabilizando uma deterioração progressiva da qualidade do meio ambiente. Tal indicação, sustentada pelas poluições urbanas, induzem a efeitos deletérios, conferindo assim, elevados custos à sociedade. Por outro lado, nos últimos anos, observou-se um aumento na conscientização sobre a importância do espaço verde urbano e da qualidade ambiental.

Isso ocorre, em grande parte, devido à rápida urbanização, que tem ocorrido em muitas cidades em todo o mundo. Esse processo leva à redução da área verde e aumento da poluição do ar e da água e, por isso, vem despertando preocupações relacionadas às questões ambientais. Os ativos ambientais próximos ou dentro de áreas residenciais em cidades em desenvolvimento são particularmente preocupantes devido à sua vulnerabilidade a danos e perda, e estão intimamente ligados ao bem-estar e saúde dos residentes.

Diamantoudi e Sartzetakis (2001) argumentam que o estabelecimento de convenções entre países, fundamentado nas pesquisas desenvolvidas recentemente, apontam uma ampla inquietação com a conservação dos recursos naturais simultaneamente à expansão da produção para atender ao crescimento célere da demanda por bens e serviços.

O esclarecimento da sociedade mundial sobre a temática do meio ambiente, impulsionado pelo nível e quantidade de informações à disposição, tem conduzido ao entendimento de que os recursos derivados do meio ambiente são limitados e, logo, devem ser empregados com prudência. Por exemplo, a teoria do desenvolvimento sustentável surge alicerçada nestas preocupações (COMISSÃO MUNDIAL SOBRE MEIO AMBIENTE E DESENVOLVIMENTO, 1988, p. 49).

Motta (1997) reforça que o uso inapropriado e/ou a degradação do recurso natural advém especialmente pela ausência de um mercado associado a esse recurso, o que acarreta a indefinição de seu preço. Esta escassez de preço resulta na não inclusão desses valores na tomada de decisão, seja como benefício social, quando se escolhe por conservá-lo, seja como custo social quando se decide por degradá-lo.

Para Marques e Comune (1997) *apud* Uberti (2000), a necessidade de conceituar o valor econômico do meio ambiente, bem como de desenvolver técnicas para estimar este valor, surge basicamente, do fato incontestável de que a maioria dos bens e serviços ambientais e das funções providas ao homem pelo ambiente não é transacionada pelo mercado. Pode-se inclusive ponderar que a necessidade de estimar valores para os ativos ambientais atende às necessidades

da adoção de medidas que visem à utilização sustentável do recurso. A crescente complexidade do mercado imobiliário urbano e a necessidade de captar relações não lineares entre variáveis locacionais, físicas e ambientais têm impulsionado a adoção de métodos mais flexíveis na modelagem dos valores dos imóveis.

Nesse sentido, conforme citado por Paixão (2025) e Hansen (2021) entende a aprendizagem de máquina como um conjunto de procedimentos voltados principalmente para estimações pontuais, com a capacidade de lidar com grandes volumes de dados, elevado número de variáveis e estruturas funcionais desconhecidas. Essa característica se mostra especialmente relevante no âmbito da avaliação de imóveis urbanos, onde modelos como Random Forest, XGBoost, Support Vector Machines e Redes Neurais vêm sendo utilizados para superar as limitações dos modelos paramétricos tradicionais. A aplicação desses algoritmos permite maior acurácia na previsão dos preços e oferece, adicionalmente, subsídios para a análise da importância relativa dos atributos considerados, contribuindo para uma abordagem mais robusta e aderente à complexidade do espaço urbano avaliado.

Nessa seara, e especificamente para o caso do ambiente natural urbano, estudos têm sido realizados buscando relacionar o valor de mercado dos imóveis à existência de atributos ambientais próximos a estes (BATALHONE *et al*., 2002; DANTAS, 2005). O intento destes estudos foi a obtenção do valor que os indivíduos estariam dispostos a pagar pela presença das amenidades ambientais, por meio das diferenças dos preços dos imóveis. A diferença paga é uma estimativa, acredita-se, da influência do ambiente natural no bem-estar dos indivíduos e, portanto, do valor monetário associado ao recurso ambiental.

1.1 Contextualização e Justificativa

Quando se trata de desenvolvimento econômico e social não se pode deixar de fora desse diálogo o meio ambiente. Esse entrelaçamento deve ser tratado mediante a mudança do conteúdo, das modalidades e das utilizações do crescimento. A preocupação com os problemas ambientais aparece como um elemento importante a respeito do crescimento material e econômico e da qualidade de vida.

Bresser-Pereira (2014) define desenvolvimento econômico como a melhoria dos padrões de vida motivada pela acumulação de capital com a incorporação do progresso técnico. Já o desenvolvimento humano, segundo o autor, é definido como o avanço das sociedades modernas em direção a seus cinco objetivos políticos autodefinidos. O autor busca ainda, distinguir as formas de desenvolvimento relativas a cada um desses objetivos: desenvolvimento

da segurança (maior paz entre as nações e menos crimes), desenvolvimento econômico (maior bem-estar), desenvolvimento político (maior igualdade política e maior participação no governo), desenvolvimento social (maior igualdade econômica) e desenvolvimento ambiental (maior proteção do ambiente). Diante do exposto, percebe-se que o desenvolvimento ambiental figura entre um dos pilares que alicerçam tanto o desenvolvimento econômico, quanto o desenvolvimento humano.

Segundo Uberti (2000), as metodologias utilizadas atualmente para precificar os bens e serviços sem valor no mercado são ainda muito controversas e, muitas vezes, pouco representativas para países em desenvolvimento, uma vez que foram elaboradas para atender às necessidades dos países desenvolvidos. Por isso, é preciso que haja estudos mais detalhados das mesmas, com aplicações prévias de testes para a identificação das possibilidades de adaptação às realidades sociais e econômicas do Brasil.

Este estudo é fundamentado na latente necessidade de conhecer os valores associados ao meio ambiente que influenciam de forma direta e indireta na valoração dos imóveis urbanos. Associado a isso, pretende-se contribuir com os estudos de valoração econômica do meio ambiente, desenvolvendo um modelo estatístico que inclua variáveis relativas ao meio ambiente urbano a partir de conceitos da engenharia de avaliações e modelagem estatística aliada à pesquisa na variação dos valores imobiliários.

2 OBJETIVOS

2.1 Objetivo Geral

Avaliar a influência de variáveis ambientais nos preços de imóveis urbanos na cidade de Palmas-TO, considerando a proximidade ao Parque Cesamar e ao lago da Usina Hidrelétrica de Lajeado (Praia da Graciosa), por meio da aplicação do Modelo Linear Generalizado (MLG), Modelo Linear (ML) e de Modelos de Aprendizagem de Máquina (AM).

2.2 Objetivos Específicos

1. Identificar a influência dos atributos ambientais - Parque Cesamar e lago da Usina Hidrelétrica de Lajeado (Praia da Graciosa) na precificação dos imóveis transacionados e localizados no Plano Diretor de Palmas-TO, utilizando a base de

dados do Imposto sobre Transmissão de Bens Imóveis (ITBI), referente aos anos de 2019, 2020 e 2021;

2. Estimar um modelo de valoração ambiental a partir da aplicação do Modelo Linear Generalizado (MLG) e do Modelo Linear (ML), mensurando o impacto da proximidade aos ativos ambientais sobre os preços dos imóveis;
3. Estimar um modelo de valoração ambiental utilizando técnicas de Aprendizagem de Máquina, visando à identificação da influência dos ativos ambientais nos valores imobiliários urbanos;
4. Avaliar e comparar a relevância das variáveis ambientais, fiscais e de renda na determinação dos preços dos imóveis urbanos, a partir das abordagens modeladas;
5. Comparar o desempenho preditivo entre o Modelo Linear Generalizado, Modelo Linear e os Modelos de Aprendizagem de Máquina.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Valor Econômico dos Recursos Ambientais (VERA)

O valor econômico dos recursos ambientais muitas vezes não é facilmente mensurável através dos preços de mercado, os quais não refletem adequadamente o custo de oportunidade desses recursos. Esse valor é baseado em todos os seus atributos, que podem ou não estar associados a um uso específico. Em outras palavras, o consumo de um recurso ambiental pode ocorrer tanto por meio de seu uso quanto de sua não utilização (MOTTA, 1997).

Um bem é homogêneo quando os seus atributos ou características que geram satisfação de consumo não se alteram. Outros bens são, na verdade, parte de classes de bens ou serviços compostos. Nestes casos, cada membro da classe apresenta atributos diferenciados, como, por exemplo, automóveis, casas, viagens de lazer e também recursos ambientais. (MOTTA, 1997).

Ainda segundo Motta (1997), os atributos de um recurso ambiental são definidos pelos fluxos de bens e serviços ambientais que são derivados do seu consumo. No entanto, há também atributos de consumo associados à própria existência do recurso ambiental, independentemente do fluxo atual e futuro de bens e serviços resultantes do seu uso.

Portanto, é comum na literatura decompor o valor econômico do recurso ambiental (VERA) em valor de uso (VU) e valor de não uso (VNU). O valor de uso pode ser dividido em valor de uso direto (VUD), quando um indivíduo usa atualmente um recurso, como, por exemplo, na forma de extração, visitação ou outra atividade de produção ou consumo direto e valor de uso indireto (VUI), quando se refere ao benefício atual de um recurso que é derivado das funções ecossistêmicas, tais como a preservação das florestas, que proporciona proteção do solo e estabilidade climática.

Tem-se ainda o valor de opção (VO) quando o indivíduo atribui valor em usos direto e indireto que poderão ser optados em futuro próximo e cuja preservação pode ser ameaçada. Por exemplo, o benefício advindo de fármacos desenvolvidos com base em propriedades medicinais ainda não descobertas de plantas em florestas tropicais.

O valor de não-uso, também conhecido como valor passivo, está associado ao valor de existência (VE) que não depende do uso do recurso, mas sim de uma posição moral, cultural, ética ou altruística em relação aos direitos de existência de espécies não-humanas ou à preservação de outras riquezas naturais. Mesmo que esses recursos não representem o uso atual ou futuro para o indivíduo, ainda assim, o valor de não-uso representa um consumo ambiental.

3.2 Método dos Preços Hedônicos para Valoração de Ativos Ambientais

Para Marques e Comune (1997) *apud* Uberti (2000), a necessidade de conceituar o valor econômico do meio ambiente, bem como de desenvolver técnicas para estimar este valor, surge basicamente, do fato incontestável de que a maioria dos bens e serviços ambientais e das funções providas ao homem pelo ambiente não é transacionada pelo mercado. Pode-se inclusive ponderar que a necessidade de estimar valores para os ativos ambientais atende às necessidades da adoção de medidas que visem à utilização sustentável do recurso.

Por meio do método dos valores hedônicos, Borba (1992) definiu um modelo de avaliação da propriedade imobiliária referenciado à qualidade ambiental, como instrumento para estudos de impacto ambiental. O autor empregou o modelo no problema do odor exalado pela operação de compostagem de lixo, impacto causado no meio ambiente pela Usina de Compostagem da Vila Leopoldina em São Paulo.

Análise hedônica consiste primordialmente na determinação do impacto de uma característica na formação do valor de um imóvel, mantendo os outros atributos constantes. A identificação correta das variáveis independentes que serão incluídas no estudo passa a ser importante, uma vez que deve ser verificada a possibilidade de existência de multicolinearidade entre as características identificadas nos edifícios.

Uberti (2000) afirma que a distinção entre os valores que o ambiente detém por si próprio pode ser dividida em dois grandes grupos, os chamados valores de uso e valores intrínsecos. Os valores de uso referem-se ao uso efetivo ou potencial que o recurso pode prover, enquanto que os valores intrínsecos não estão associados nem com uso efetivo presente do recurso e nem com as possibilidades de uso futuro. O valor intrínseco reflete o valor que reside nos recursos ambientais, independente de uma relação com os seres humanos, sendo captado pelas pessoas através de suas preferências na forma de não-uso do recurso.

Merico (1996) relata que evidenciar os valores monetários do ambiente natural pode parecer, sob alguns aspectos, imoral, mas se justifica pelo fato de que estes valores monetários podem ser utilizados como padrão de medida, indicando ganhos e perdas em utilidade ou bem-estar. São bastante úteis no âmbito da discussão dos rumos e alternativas de desenvolvimento, demonstrando caminhos que levem à feita de sustentabilidade ambiental e produzindo uma análise que consegue capturar elementos não incorporados pela análise econômica tradicional.

O Método dos Preços Hedônicos (MPH) é referenciado como método do preço implícito (MOTTA, 1997) ou técnica do preço da propriedade (BELLIA, 1996; MARGULIS, 1996). A base deste método é a identificação de atributos ou características de um bem ambiental (bonita

paisagem, alto risco, etc.) que tenham relacionamento com o preço da terra ou do trabalho. Duas técnicas podem ser utilizadas neste método: diferenças nos preços dos imóveis e diferenças nos níveis salariais.

Segundo Hochheim (2001) e Uberti (2000), o valor de um imóvel possui afinidade com diversas variáveis como área, padrão de acabamento, número de quartos, localização, entre outros, e pela qualidade ambiental do seu entorno, tais como vista panorâmica, poluição do ar ou sonora. As diferenças nos preços dos imóveis derivadas destas variáveis ambientais podem ser utilizadas para mensurá-las.

Este método quando aplicado às diferenças nos preços dos imóveis, parte do pressuposto que a qualidade ambiental afeta os preços de venda da terra e dos imóveis. O valor de um imóvel está relacionado às vantagens que dele provêm em relação a outros imóveis. Um apartamento que oferece, por exemplo, uma linda vista, com certeza terá valor maior do que aquele que não tem este atributo.

Paixão (2015) ressalta que o modelo de preços hedônicos tem sido utilizado como metodologia para se gerar índices de preços para bens imóveis. Por esse modelo, o preço do bem é uma função dos atributos que o compõe. Cada atributo, portanto, teria um preço embutido no preço final do bem ("preço sombra") e o preço do imóvel seria o somatório do preço sombra de cada um de seus atributos.

Desse modo, a utilização do modelo de preços hedônicos é uma forma de se obter índices de preço para imóveis controlado pelas diferenças nos atributos de cada observação. Em 2006, a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) e o Fundo Monetário Internacional (FMI) realizaram o *Workshop on Real Estate Price Indexes*, onde foram apresentadas as experiências de vários países na construção de índice de preços para imóveis. Em um balanço das apresentações, Diewert (2009) constatou que o modelo de preços hedônicos é o mais utilizado e, segundo esse autor, é o mais recomendando para a construção dos índices.

No Brasil, o modelo de preços hedônicos também tem sido aplicado para mensurar diversos aspectos do mercado imobiliário. Alguns estudos focam em um conjunto de variáveis ambientais, como Hermann e Haddad (2005) para a cidade de São Paulo, Albuquerque *et al*., (2007), para Recife, e Batalhone *et al*., (2002), para Brasília. Outros estudos focam em aspectos mais específicos. Teixeira e Serra (2006) avaliaram o impacto da criminalidade urbana para a cidade de Curitiba; Amrein (2010) teve como foco a qualidade das escolas no município de São Paulo; Aguirre e Faria (1997) e Faria (2008) utilizaram o modelo para avaliar políticas públicas.

Em seu estudo, Paixão (2015), relata que as características de vizinhança são importantes na formação do valor dos imóveis urbanos. Brueckner *et al*.(1999) consideram a renda das famílias a melhor proxy para a qualidade dos serviços localizados em um bairro. Nesse sentido, incluindo na análise um indicador de hierarquia socioespacial, pode-se esperar que quanto maior for o índice para o bairro, maior será o preço do imóvel.

A grande maioria das cidades brasileiras é marcada por acentuada desigualdade no que se refere às condições de moradia da população e à disponibilidade de infraestrutura urbana básica. Nesse cenário, o índice de inadequação de moradias é utilizado como uma variável proxy para representar a presença de vilas e favelas no entorno imediato do imóvel avaliado. Conforme apontado por Uberti (2000) e Hochheim (2001), a precariedade dos assentamentos informais caracterizada pela ausência de infraestrutura, baixa qualidade construtiva e elevado grau de vulnerabilidade social, tende a gerar externalidades negativas que impactam diretamente a valoração dos imóveis localizados nas proximidades. Por esse motivo, espera-se que tal indicador apresente coeficiente negativo nos modelos de regressão, sinalizando a depreciação dos preços de imóveis adjacentes a grandes assentamentos informais. Essa abordagem é compatível com a modelagem estatística discutida por Morettin e Singer (2020), que reforçam a importância de se incorporar variáveis de caráter socioambiental nos modelos de precificação, a fim de captar os efeitos estruturais e espaciais que influenciam a formação dos preços no mercado imobiliário urbano.

A ausência de infraestrutura urbana adequada também é fator que desvaloriza o imóvel. Dentre a infraestrutura urbana incluem-se a presença de iluminação pública, saneamento básico e condições adequadas de coleta de lixo, fatores que influenciam diretamente o conforto, a sensação de segurança e a saúde da população local. Por isso, espera-se que o indicador de inadequação de infraestrutura urbana tenha um sinal negativo.

De acordo com Jim e Chen (2006), o Método dos Preços Hedônicos é considerado como a abordagem mais convincente para estimar o valor de um ativo ambiental, pois sua técnica se baseia em comportamentos transacionais reais do mercado. O MPH utiliza dados de mercado para estimar o impacto dos atributos ambientais sobre o preço de um bem ou serviço, permitindo uma avaliação precisa do valor dos ativos ambientais.

Esse método é baseado na premissa de que os compradores e vendedores de um bem ou serviço são capazes de incorporar as informações ambientais disponíveis na formação dos preços dos bens e serviços, refletindo assim o valor que os consumidores atribuem aos ativos ambientais. Em suma, o MPH é uma técnica que se baseia em dados reais do mercado e permite

a estimativa precisa do valor dos ativos ambientais, tornando-o uma abordagem convincente para avaliar o impacto ambiental de projetos e políticas públicas adequadas.

Existem duas abordagens principais para estimar o valor monetário dos serviços ambientais: preferência declarada e preferência revelada (ADAMOWICZ *et al*., 1994; PAREDES, 2005). Na abordagem de preferência declarada, as técnicas de pesquisa são usadas para obter as preferências individuais e os valores pelos serviços ambientais (BATEMAN *et al*., 2002). Os entrevistados são questionados sobre quanto estão dispostos a pagar para preservar as boas características para recreação e comodidade (TYRVÄINEN e VÄÄNÄNEN, 1998, KWAK *et al*., 2003; JIM e CHEN, 2006).

O valor total da disposição a pagar em conjunto com a população em questão é usado para estimar o valor dos serviços. A segunda abordagem (preferência revelada) é conhecida como método de precificação hedônico. O valor dos serviços ambientais é inferido pela estimativa do preço de venda ou valor do imóvel em função dos atributos ambientais (como a proximidade de parques urbanos e vista para um jardim), em associação com outras características do imóvel e do bairro. Por meio de cálculos estatísticos, é possível atribuir partes do valor da propriedade aos espaços verdes e outras dotações paisagísticas.

3.3 Modelos Estatísticos Aplicados

3.3.1 Modelo Linear Generalizado (MLG)

O MLG é composto por três componentes: uma função de ligação, uma distribuição de probabilidade da família exponencial e um modelo de variância. A função de ligação é usada para relacionar a média da variável de resposta com uma combinação linear das variáveis explicativas.

A distribuição de probabilidade da família exponencial é escolhida para permitir que a variável resposta tenha uma ampla gama de distribuições, incluindo a normal, a binomial e a *Poisson*, entre outras. O modelo de variância é usado para levar em conta a heterogeneidade da variância da variável resposta (MCCULLAGH e NELDER, 1999).

Nelder e Wedderburn (1972) propuseram os Modelos Lineares Generalizados (MLGs) como uma extensão dos modelos clássicos. A ideia principal é aumentar a variedade de distribuições da variável resposta, permitindo que pertença à família exponencial de

distribuições e tornando a relação funcional entre a média da variável resposta (μ) e o preditor linear η mais flexível.

A ligação entre a média e o preditor linear não precisa ser a identidade e pode ter qualquer forma monótona não-linear (NELDER e WEDDERBURN, 1972). Além disso, os autores introduziram um processo iterativo para a estimação dos parâmetros e o conceito de desvio, que é amplamente utilizado na avaliação da qualidade do ajuste dos MLGs, bem como no desenvolvimento de resíduos e medidas de diagnóstico.

McCullagh e Nelder (1999) reforçam que os modelos lineares generalizados são uma generalização dos modelos lineares clássicos e incorporam uma função de ligação que relaciona o valor esperado da variável resposta com a combinação linear das variáveis explicativas. Além disso, o MLG permite o ajuste de modelos de regressão para dados com distribuições não normais, sendo amplamente utilizado em diversas áreas, como biologia, medicina, finanças e engenharia.

O MLG é definido por uma distribuição de probabilidade, membro da família exponencial de distribuições, para a variável resposta, um conjunto de variáveis independentes descrevendo a estrutura linear do modelo e uma função de ligação entre a média da variável resposta e a estrutura linear.

De forma geral, a estrutura de um MLG é formada por três partes:

- Componente aleatória: composta de uma variável aleatória *Y* com *n* observações independentes, um vetor de médias μ e uma distribuição pertencente à família exponencial;
- Componente sistemática: composta por variáveis explicativas X_1, ... ,*Xp* tais que, produzem um preditor linear η;
- Uma função monotônica diferenciável, conhecida como função de ligação, que relaciona estas duas componentes por meio de η e a média da variável resposta.

$$\mu_y = \mu$$

Pode-se dizer que o termo “generalizado” no MLG significa uma distribuição mais ampla do que a normal para a variável e uma função não-linear relacionando à média dessa variável e a parte determinística do modelo (MCCULLAGH e NELDER, 1999).

O modelo linear generalizado é amplamente utilizado em análises de dados, especialmente quando os dados têm uma distribuição não normal. De acordo com Dobson (2002), os modelos lineares generalizados constituem o principal instrumento para ajustar modelos para dados em que a distribuição da resposta é qualquer uma das distribuições exponencialmente familiares.

De acordo com Cordeiro e Demétrio (2008), a seleção do modelo é uma etapa crucial em qualquer pesquisa, que envolve a busca por um modelo simples e razoável que possa descrever adequadamente os dados observados. Na maioria dos casos, a variável resposta é composta por duas partes distintas:

- Um componente sistemático, que é estabelecido durante o planejamento do experimento (fundamental para obter conclusões confiáveis), resultando em modelos de regressão (linear simples, múltipla, não linear, entre outros), modelos de análise de variância (como delineamentos inteiramente casualizados, blocos casualizados, quadrados latinos com estrutura de tratamentos fatoriais, parcelas subdivididas, entre outros) e modelos de análise de covariância;
- Um componente aleatório, que é determinado quando são definidas as medidas a serem realizadas, que podem ser contínuas ou discretas, e exigem o ajuste de diferentes distribuições. Um mesmo experimento pode envolver medidas de diferentes tipos, como altura, número de lesões e proporção de plantas doentes. No modelo linear clássico, é considerado:

$$Y = \mu + \epsilon \tag{3.1}$$

Sendo, Y o vetor, de dimensões $n \times 1$, da variável resposta, $\mu = E(Y) = X\beta$ o componente sistemático, X a matriz, de dimensões $n \times p$, do modelo, $\beta = (\beta_1, \dots, \beta_p)^T$ o vetor dos parâmetros, $\epsilon = (\epsilon_1, \dots, \epsilon_p)^T$, o componente aleatório com $\epsilon \sim N(0, \sigma^2)$, $i = 1,\dots,n$.

Em diversos cenários, contudo, a relação aditiva entre o componente sistemático e o componente aleatório não é atendida. Além disso, não há razão para se restringir à estrutura simples dada por μ para o componente sistemático e nem para se restringir à $\mu = E(Y) = X\beta$ distribuição normal para o componente aleatório e à suposição de homogeneidade de variâncias (DEMÉTRIO, 2002).

3.3.1.1 Definição

Sejam $Y_1 \ldots, Y_n$ variáveis aleatórias independentes, cada uma com função densidade ou função de probabilidades na forma dada abaixo:

$$f(Y_i;\ \theta_i; \emptyset) = exp\ [\emptyset\{y_i\theta_i - b(\theta_i) + c(y_i, \emptyset)]$$

(3.2)

Em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas, pode-se mostrar sob as condições usuais de regularidade.

$$E\left\{\frac{\partial log f(Y_i; \theta_i;\ \emptyset}{\partial \theta_i}\right\} = 0$$

(3.3)

$$E\left\{\frac{\partial^2 log f(Y_i; \theta_i; \emptyset)}{\partial \theta^2 i}\right\} = -E\left\{\left[\frac{\partial\ log f(Y_i; \theta_i; \emptyset)}{\partial \theta_i}\right]^2\right\}$$

(3.4)

$\forall i$, que $E(Y_i) = \mu_i = b'(\theta i)$ e $Var(Y_i) = \phi^{-1} V\ \mu_i$, em que $V_i = V\ (\mu_i) = d\mu_i/d\theta_i$ é a função de variância de $\phi^{-1} > 0$ é o parâmetro de dispersão. A função de variância desempenha um papel fundamental na família exponencial, uma vez que a mesma caracteriza a distribuição.

Assim, dado uma função de variância, é possível encontrar uma classe correspondente de distribuições e vice-versa. Essa característica possibilita a comparação de distribuições por meio de testes simples da função de variância. Por exemplo, considerando a função de variância definida por $V\ \mu_i = \mu\ (1 - \mu)$, $0 < \mu < 1$, caracteriza-se a classe de distribuições binomiais com probabilidades de sucesso (μ ou $1 - \mu$) (PAULA, 2004).

Uma propriedade interessante envolvendo a distribuição de Y e a função de variância é a seguinte: $\sqrt{\emptyset}(Y - \mu) \xrightarrow{d} N\big(0, V(\mu)\big)$ quando $\phi \dashrightarrow \infty$ ou seja, para ϕ grande Y segue distribuição aproximadamente normal de média μ e variância ϕ -1V (μi). Os modelos lineares generalizados são definidos pela equação (3.2) e pela parte sistemática:

$$g(\mu_i) = \eta_i$$

(3.5)

Sendo que $\eta_i = X_i^T\beta$ é o preditor linear, $\beta = (\beta_1 \dots , \beta_p)^T$ e $p < n$, é um vetor de parâmetros desconhecidos a serem estimados, $X_i = (X_{i1}, \dots , X_{ip})T$ representa os valores de variáveis explicativas e $g(\cdot)$ é uma função monótona e diferenciável, denominada função de ligação (PAULA, 2004). Apresenta-se a seguir o caso particular da Distribuição Gama. Seja Y uma variável aleatória com distribuição gama de média μ e coeficiente de variação $\phi^{-1/2}$, denota-se $Y \sim G(\mu, \phi)$. A função densidade de Y é dada por:

$$\frac{1}{\Gamma_{(\phi)}}\left(\frac{\phi y}{\mu}\right)^{\phi} exp\left(-\frac{\phi y}{\mu}\right) d(logy) = exp[\phi\{(-y/\mu) - log\mu\} - log\Gamma(\phi) + \phi log(\phi y) - logy] \quad (3.6)$$

Em que $y > 0$, $\phi > 0$, $\mu > 0$ e $\Gamma(\phi) = \int_0^{\infty} t^{\phi-1}dt$ é a função gama assim, fazendo $\theta = -1/\mu$, $b(\theta) = -log(-\theta)$ e $c(y, \phi) = (\phi - 1)logy + \phi\ log\ \phi - log\Gamma(\phi)$, assim satisfazendo a equação (3.6) (PAULA, 2004).

Para $0 < \phi < 1$ a densidade da gama tem uma *pole* na origem e decresce monotonicamente quando $\gamma \dashrightarrow \infty$. A exponencial é um caso especial quando $\phi = 1$. Para $\phi > 1$ a função densidade assume zero na origem, tem um máximo em $y = \varphi\mu\text{-}\varphi/$ despois decresce para $\gamma \dashrightarrow \infty$. A χ^2 é um outro caso especial quando $\phi = k/2$ e $\mu = k$. A distribuição normal é obtida fazendo $\phi \dashrightarrow \infty$. Isto é, quando ϕ é grande $Y \sim N(\mu, \phi\ \text{-}1V\ (\mu\mu_i))$.

Tem-se que $\phi = E2(Y)/V\ ar(Y)$ é o inverso do coeficiente de variação da Y ao quadrado, ou seja, $\phi = 1/(CV)^2$, em que CV: $\frac{\sqrt{Var\ (Y)}}{E(Y)}$. A função de variância da gama é dada por $V\ (\mu_i) = \mu^2$ (PAULA, 2004).

3.3.1.2 Ligações Canônicas

Supondo ϕ conhecido, o logaritmo da função de verossimilhança de um modelo linear generalizado com respostas independentes pode ser expresso na forma:

$$L(\beta; y) = \sum_{i=1}^{n} \phi\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi) \quad (3.7)$$

Um caso particular importante ocorre quando o parâmetro canônico (θ) coincide com o preditor linear, isto é, quando $\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$. Nesse caso, $L(\beta)$ é dado por:

$$L(\beta) = \sum_{i=1}^{n} \quad \phi \left\{ y_i \sum_{j=1}^{p} \quad xij_{\beta j} - b\left(\sum_{j=1}^{p} \quad x_{ij}\beta_j \right) \right\} + \sum_{i=1}^{n} \quad \left(yi_{\phi}\right)$$

(3.8)

Definindo a estatística $S_j = \phi \sum_{i=1}^{n} Y_i x_{ij}$, $L(\beta)$ fica então reescrito na forma:

$$L(\beta) = \sum_{j=1}^{p} \quad S_j\beta_j - \phi \sum_{i=1}^{n} \quad b\left(\sum_{j=1}^{p} \quad x_{ij}\beta_j \right) + \sum_{i=1}^{n} \quad c(y_i, \phi)$$

(3.9)

Assim, pelo teorema da fatoração a estatística $S = (S1,...,Sp)^T$ é suficiente minimal para o vetor $\beta = (\beta_1,...,\beta_p)^T$. As conexões que se referem a essas estatísticas são denominadas conexões ou ligações canônicas e têm um papel crucial na teoria dos modelos lineares generalizados. (PAULA, 2004).

3.3.1.3 Função de Desvio

Assumindo a mesma generalidade, considere agora que o logaritmo da função de verossimilhança seja definido por:

$$L(\mu; y) = \sum_{i=1}^{n} L(\mu_i; y_i)$$

(3.10)

Em que $\mu i = g\text{-}1(\eta i)$ e $\eta i = xT\ i\ \beta$. Para o modelo saturado ($p = n$) a função $L(\mu; y)$ é estimada por:

$$L(y; y) = \sum_{i=1}^{n} \quad L(y_i; y_i)$$

(3.11)

Ou seja, a estimativa de máxima verossimilhança de μ_i fica nesse caso dada por $\tilde{\mu}i = yi$. Quando $p < n$, denota-se a estimativa de $L(\mu; y)$ por $L(\hat{\mu}; y)$. Aqui, a estimativa de máxima verossimilhança de μ_i será dada por $\hat{\mu} = g\text{-}1(\hat{\eta}_i)$ em que $\hat{\eta}_i = xT_{i\,\beta}$ (PAULA, 2004).

A qualidade do ajuste de um modelo linear generalizado é avaliada por meio da função desvio:

$$D^* = (y; \hat{\mu}) = \phi D(y; \hat{\mu}) = 2\{L(y; y) - L(\hat{\mu}; y)\}$$

(3.12)

Onde:

$$D(y; \hat{\mu}) = 2\sum_{i=1}^{n} \left[yi\left(\tilde{\theta} - \hat{\theta}\right) + \left(b\left(\hat{\theta}\right) - b\left(\hat{\theta}_i\right)\right)\right]$$

(3.13)

Denotando por $\widehat{\theta\imath} = \theta i(\hat{\mu}i)$ e $\tilde{\theta}i = \theta i(\tilde{\mu}i)$ respectivamente, as estimativas de máxima verossimilhança de θi para os p modelos com p parâmetros ($p < n$) e saturado ($p = n$).

Apresenta-se a seguir a função desvio da Gama:

$$D(y; \hat{\mu}) = 2\sum_{i=1}^{n} \left[-log\left(\frac{yi}{\hat{\mu}_i}\right) + \frac{(yi - \hat{\mu}_i)}{\hat{\mu}_i}\right]$$

(3.14)

Um valor reduzido para a função desvio indica que, mesmo com menos parâmetros, é possível obter um ajuste tão bom quanto o modelo saturado. Embora seja usual comparar os valores observados da função desvio com os percentis da distribuição *qui- quadrado* com *n - p* graus de liberdade, em geral, $D(y; \hat{\mu})$ não segue assintoticamente uma distribuição χ^2 *n-p*.

No caso do modelo gama, o desvio estará bem aproximado por uma *qui-quadrado* com *n - p* graus de liberdade à medida que o coeficiente de variação ficar próximo de zero (PAULA, 2004).

3.3.1.4 Análise do Desvio

Suponha para o vetor de parâmetros β a partição $\beta = (\beta_1 , \beta_2)^T$, em que β_1 é um vetor q-dimensional enquanto β_2 tem dimensão $p - q$ e ϕ é conhecido(ou fixo). Logo pode-se estar interessados em testar as hipóteses H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$. As funções desvio correspondentes aos modelos sob H_0 e H_1 serão denotadas por $D(y; \mu_0)$ e $D(y;\hat{\mu})$, respectivamente, em que μ_0 é a estimativa de máxima verossimilhança sob H_0. A análise de desvio (ANODEV) é uma generalização da análise de variância para os MLGs. Pode-se definir a seguinte estatística.

$$F = \frac{\{D(y;\hat{\mu}^0) - D(y;\hat{\mu})/q\}}{\big(D(y;\hat{\mu})\big)/(q-p)} \tag{3.15}$$

Cuja distribuição nula assintótica é $Fq,(n\text{-}p)$. Não depende de ϕ e é invariante sob reparametrização, pode ser obtida diretamente de funções desvio, é muito conveniente para uso prático (PAULA, 2004).

3.3.1.5 Função Escore e Informação de Fisher

Segundo Paula (2004) o logaritmo da função de verossimilhança de um modelo linear generalizado definido por equação (3.2) e equação (3.5) pode ser expresso na forma:

$$L(\beta; y) = \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i) + c(y_i) + a(y_i;\phi)] \tag{3.16}$$

A função escore total e a matriz de informação total de Fisher para o parâmetro β são dadas por:

$$U(\beta) = \frac{\partial L(\beta; y)}{\partial \beta} = \phi X^T W^{1/2} V^{-1/2} (y - \mu) \tag{3.17}$$

$$K(\beta) = E\left\{\frac{\partial^2 L(\beta; y)}{\partial\beta\partial\beta^2}\right\} = \phi X^T W X$$

(3.18)

Em que *X* é a matriz modelo $n \times p$ de posto completo cujas linhas são representadas por $X_i^T, i = 1, \dots, n, W = diag\{W_i, \dots, W_n\}$ com $wi = \left(\frac{d\mu_i}{dn_i}\right)^2$ é a matriz de pesos $V = diag\{V_1, \dots V_n\}, y = (y_i, \dots y_n)^T$ com $Vi = \left(\frac{d\mu_i}{d\theta_i}\right)^2$ e $\mu = (\mu_i, \dots, \mu_i)^T$.

3.3.1.6 Estimação de β

De acordo com Paula (2004) para a obtenção da estimativa de máxima verossimilhança de β utiliza-se o processo iterativo de Newton-Raphson, que pode ser reescrito como um processo iterativo de mínimos quadrados reponderados, dado por: *m* = 0, 1,...,onde = η + *W* - 1 2 *V* - 1 2 (*y* - μ).

$$\beta^{m+1} = (X^T W^m X)^{-1} X^T W^m z^m$$

(3.19)

Observe que a quantidade *z* faz o papel de uma variável dependente modificada, enquanto que *W* é uma matriz de pesos que muda a cada passo do procedimento iterativo. A convergência de equação (3.19) ocorre em geral em um número finito de passos, independente dos valores iniciais utilizados. É usual iniciar equação (3.19) com $\eta i(0) = g(yi)$ para *i* = 1,...,*n.* Nesse caso, $\hat{\beta}$ assume a forma fechada $\hat{\beta}$ = (X^T *X*)-1$X^T y$ (PAULA, 2004).

3.3.1.7 Estimação de ϕ

Segundo Paula (2004) igualar a função escore Uϕ a zero chega no seguinte resultado:

$$\sum_{i=1}^{n} \acute{c}\left(y_{i,\hat{\phi}}\right) = \frac{1}{2} D(y, \hat{y}) - \sum_{i=1}^{n} \left\{y_i \tilde{\theta}_i - b\tilde{\theta}_i\right\}$$

(3.20)

Onde $D(y;\hat{\mu})$) representa o desvio do modelo observado. Analisa-se que a estimativa de máxima verossimilhança para φ nos casos normal e normal inversa, igualando Uϕ a zero, e é dada por:

$$\hat{\phi} = \frac{n}{D(y;\hat{\mu})} \tag{3.21}$$

No caso da gama, a estimativa de máxima verossimilhança de ϕ é obtida pela equação:

$$2n\{log(\hat{\phi}) - \psi(\hat{\phi}) = D(y;\hat{\mu})\} \tag{3.22}$$

3.3.1.8 Estimação de Modelos

De acordo com Florencio (2010), a especificação de modelos para a estimação empírica da equação de preços hedônicos requer compreensão, intuição e habilidade, e não pode ser realizada mecanicamente. Embora o senso comum, a lógica e a experiência de outros pesquisadores possam fornecer orientações para selecionar o melhor método para explicar a formação dos preços, tais teorias devem ser testadas com base em dados de mercado para verificar sua adequação.

Florencio (2010) apresenta de uma forma bastante didática a interpretação geométrica nas equações de preços hedônicos voltadas para o mercado imobiliário. Elas têm sido, em sua maioria, formuladas com base no modelo normal de regressão linear clássico e adotam uma forma linear, log-linear.

3.3.1.9 Teste de Hipóteses

Paula (2004) apresenta de forma simples as generalizações para os MLGs. Supõe-se, inicialmente, a seguinte situação de hipóteses simples: $H_0 : \beta = \beta_0$: contra $H_1 : \beta\ 6= \beta_0$, em que β_0 é um vetor *p*-dimensional conhecido e ϕ é também assumido conhecido. A estatística da razão de verossimilhança para o teste de H_0 pode ser escrita da seguinte forma:

$$\xi_{RV} = \phi\{D((y;\hat{\mu}^0)) - D(y;\hat{\mu})\} \tag{3.23}$$

Em que $\hat{\mu}^0 = g^{-1}(\hat{n}^0), \hat{n}^0 = X\beta^0$. Assintoticamente e sob *H₀*, temos que X_q^2.

3.3.1.10 Critérios de Seleção Dos Modelos

Na maior parte das situações pode-se pensar em critérios para penalizar a inclusão de novas variáveis. Assim cada variável adicionada ao modelo, apresenta 1 grau de liberdade a menos. Dentre estes critérios temos:

1 - Critério de informação de Akaike (AIC) (1987), o qual usa a seguinte expressão:

$$AIC = log\left(\frac{SQRes}{n}\right) + \frac{2p}{n} \tag{3.24}$$

Onde *p*, é o número de regressores do modelo, *n*, é o total de dados, e *SQRes*, é a soma de quadrados do resíduo.

2 - Schwarz (1978) propôs o critério de informação Bayesiana, definido como:

$$BIC = log\left(\frac{SQRes}{n}\right) + \frac{2p}{n} log(n) \tag{3.25}$$

O BIC penaliza mais fortemente que o AIC, quando n > 8.

3.3.2 Modelo Linear (ML)

O Modelo Linear configura-se como uma das metodologias mais consolidadas na literatura de avaliação imobiliária, notadamente no âmbito do Método dos Preços Hedônicos (MPH), pois, permite a decomposição do valor de mercado dos imóveis em função de seus atributos físicos, locacionais, ambientais e socioeconômicos.

A aplicação de modelos lineares, sobretudo mediante a regressão linear múltipla, possibilita estimar o impacto marginal de cada característica sobre o preço dos imóveis, sendo amplamente utilizada tanto em pesquisas acadêmicas quanto na prática pericial.

Ainda que se reconheça a eficácia do modelo linear na identificação das principais variáveis que influenciam o valor dos imóveis, conforme demonstrado por Brueckner *et al*. (1999). Ressalta-se que sua aplicação exige a observância de pressupostos clássicos, como linearidade, homocedasticidade e normalidade dos resíduos, condições que nem sempre são plenamente verificadas em mercados caracterizados por elevada heterogeneidade e complexidade.

Nesse sentido, Jim e Chen (2006) enfatizam que, em mercados urbanos heterogêneos, a capacidade explicativa dos modelos hedônicos lineares tende a variar entre 20% e 50%, em termos do coeficiente de determinação (R^2), evidenciando que embora o modelo linear seja útil para análises exploratórias e comparativas, ele possui limitações no tratamento de relações não lineares ou efeitos interativos entre variáveis.

Apesar dessas restrições, Dantas (2005) reforça que a regressão linear permanece como um referencial metodológico fundamental, servindo como modelo base para comparação com técnicas mais sofisticadas, como os modelos generalizados ou os algoritmos de aprendizagem de máquina, que buscam superar as limitações impostas pelos pressupostos restritivos da regressão linear clássica.

Dessa forma, a inclusão do Modelo Linear nesta pesquisa justifica-se não apenas por sua tradição e simplicidade, mas também por sua utilidade como parâmetro de comparação, fornecendo uma base sólida para a análise e interpretação econômica dos fatores que determinam o preço dos imóveis urbanos.

3.4 Aprendizagem de Máquina (AM)

A avaliação de imóveis urbanos é uma atividade que demanda precisão e interpretação de múltiplas variáveis heterogêneas. Diante da crescente disponibilidade de dados e da complexidade do comportamento do mercado imobiliário, os modelos de aprendizagem de máquina (AM) têm se consolidado como ferramentas robustas para a modelagem e previsão de preços de imóveis, superando muitas das limitações dos métodos estatísticos tradicionais, como os modelos hedônicos clássicos.

Baur *et al*., (2022) analisaram a aplicação de diferentes modelos de aprendizagem de máquina, incluindo *Random Forest* e *Support Vector Machine*, para avaliar imóveis com base nas descrições textuais das propriedades. Os autores observaram ganhos expressivos de acurácia com os métodos não lineares e baseados em árvores de decisão, especialmente em bases de dados com variáveis textuais ou não estruturadas.

Choy e Ho (2023), ao avaliarem o desempenho de modelos como *Extra Trees*, k-*Nearest Neighbors* e *Random Forest,* constataram resultados superiores aos modelos hedônicos tradicionais. A precisão dos modelos foi significativamente elevada, especialmente em ambientes urbanos complexos e com alta variação de atributos.

Deng (2025) propôs um modelo de avaliação automatizada de imóveis urbanos que incorpora imagens de fachadas, mapas de ruas e imagens aéreas, utilizando redes neurais profundas para capturar padrões espaciais e visuais relevantes. Os resultados mostraram que a inclusão de dados visuais melhora substancialmente o desempenho dos modelos.

Rodriguez-Serrano (2024) desenvolveu uma abordagem baseada em protótipos e aprendizagem de máquina explicável (XAI), permitindo não apenas previsões acuradas, mas também transparência na explicação dos valores preditos. A técnica mostrou desempenho semelhante ao de *Random Forest* e *XGBoost*, mas com maior interpretabilidade.

Root, Strader e Huang (2023), realizaram uma revisão sistemática das técnicas de aprendizagem de máquina aplicadas ao mercado imobiliário, destacando *Random Forest*, *Gradient Boosting* e Redes Neurais como os modelos mais robustos em diferentes contextos urbanos. O estudo também abordou desafios como sobreajuste, interpretabilidade e viés algorítmico.

Mouna *et al*. (2023) realizaram uma comparação entre Regressão Linear, *Random Forest* e *Gradient Boosting* em um estudo empírico na cidade de Melbourne, Austrália. O modelo *Gradient Boosting* apresentou o melhor desempenho preditivo (R^2 e MAE), sendo recomendado para avaliações em mercados com alta variabilidade.

Em síntese, os modelos de aprendizagem de máquina constituem uma importante evolução metodológica para a avaliação de imóveis urbanos. A literatura nacional e internacional apresenta evidências consistentes de sua superioridade em termos de desempenho preditivo, especialmente em cenários caracterizados por grande volume de dados, alta complexidade estrutural e múltiplos atributos correlacionados.

O presente estudo insere-se nesse debate ao propor uma abordagem de valoração que incorpora atributos ambientais, utilizando algoritmos de AM para captar, de forma mais eficiente, as externalidades positivas decorrentes da proximidade ao Parque Cesamar e à Praia da Graciosa, em Palmas-TO.

3.7 Variáveis ambientais

Segundo Gazola (2002) a variável meio ambiente representa o nível de poluição, obtido por meio de notas dadas por engenheiros de diversas áreas e calculadas utilizando a estatística descritiva. Para seu estudo de caso, a variável meio ambiente torna-se relevante, pois a cidade estudada está localizada em área crítica de poluição, devido à exploração de carvão, curtume, cerâmica e metalurgia dentro do perímetro urbano. Em outro estudo de caso, o município encontrava-se na praia, e a variável distância ao mar era indispensável na lista de variáveis explicativas relevantes.

Moro (2008) emprega o termo "amenidades ambientais" para destacar a importância desse fator no valor de um imóvel. Já Uberti (2000) e Hochheim (2001) enfatizam que a poluição, seja sonora, do ar, das praias ou dos rios, são atributos que podem depreciar o valor do imóvel e não podem ser ignorados. Além disso, o tráfego, os engarrafamentos e o lixo produzido por hospitais também podem afetar negativamente o valor do imóvel.

Favero (2008) ainda acrescenta que a proximidade de pontos de inundação e aterros sanitários ou lixões também são fatores desfavoráveis que devem ser considerados. A presença de áreas verdes nas cidades é essencial devido aos diversos benefícios que proporcionam em relação à qualidade ambiental. A arborização urbana é capaz de controlar a radiação solar, temperatura e umidade do ar, bem como a ação dos ventos e chuvas, além de reduzir a poluição do ar e sonora, oferecer sombra, valorizar economicamente as propriedades e oferecer espaços de lazer para a população.

O conceito de áreas verdes urbanas pode variar entre especialistas, porém, a definição de Lorusso (1992) é a mais específica e engloba três setores distintos: áreas verdes públicas (1), que incluem espaços públicos destinados ao lazer ou contato direto com a natureza; áreas verdes privadas (2), que compreendem remanescentes vegetais significativos incorporados à malha urbana; e a arborização de ruas e vias públicas (3). Este conceito permite uma compreensão clara das características funcionais das áreas verdes urbanas.

Como exemplo de variáveis ambientais pode-se citar as emissões atmosféricas, poluição do ar, poluição do som, poluição visual, lançamentos em corpos d´água, alterações do solo, segurança, entre outros. Variáveis ambientais relevantes podem estar presentes no avaliando, devendo ser consideradas. Porém, estas possuem características próprias e subjetivas e as pessoas possuem percepções diferentes quanto a elas, sendo de difícil quantificação, devido a esse subjetivismo expresso, pelo senso comum através de termos linguísticos como “péssimo”, “ruim”, bom”, “longe”, para citar alguns.

Arraes e Souza Filho (2008) afirmam que os consumidores criam sua própria lista de desejos ao adquirir um imóvel, buscando a melhor combinação entre as diversas características

disponíveis em cada opção. Essas características são relacionadas a três fatores fundamentalmente importantes: os atributos físicos do imóvel, a localização ou vizinhança ideal e as características ambientais intrínsecas à propriedade, com o objetivo de obter uma escolha ótima.

3.8 Avaliação de Imóveis Voltada a Engenharia de Avaliações

O método comparativo de dados de mercado é aquele que define o valor através da comparação com dados de mercado assemelhados quanto às características intrínsecas e extrínsecas dos imóveis. Uma condição fundamental para a aplicação deste método é a existência de um conjunto de dados que possa ser tomado, estatisticamente, como amostra do mercado imobiliário.

A aplicação do método de regressão linear permite estimar o valor de um imóvel a partir de uma amostra de dados de mercado, representativa da população, através de modelos do tipo (HOFFMAN e VIEIRA, 1977; WONNACOT e WONNACOT-TH, 1978).

Gomide (2008) ressalta que toda avaliação imobiliária é norteada em quatro pilares fundamentais: o objetivo da avaliação, os informes sobre o imóvel avaliando, os informes do mercado e o tratamento científico aplicado a esses informes.

A avaliação de imóveis é utilizada na grande maioria dos negócios, discussões e pendências interpessoais e sociais em nossas comunidades, tais como na compra ou na venda de casas, lojas comerciais, instalações industriais, aluguéis, na reavaliação de ativos de empresas, em atendimento à legislação vigente, na partilha oriunda de heranças, meações ou divórcios, no lançamento de impostos, nas hipotecas imobiliárias, nas divergências que originam ações demarcatórias, possessórias, nas indenizações, nas desapropriações e servidões, enfim, em um número expressivo de ações oriundas de problemas inerentes aos relacionamentos humanos, onde o valor de um bem assume importância fundamental (NADAL *et al*., 2003).

Dantas (2005) destaca que o principal objetivo da Engenharia de Avaliações é a determinação técnica do valor de um bem, dos seus custos, frutos ou direitos sobre ele. O valor que se pretende determinar numa avaliação é o Valor de Mercado, o qual é definido pela Norma Brasileira para Avaliação de Bens NBR 14653-1 (2019): Quantia mais provável pela qual se negociaria voluntariamente e conscientemente um bem, numa data de referência, dentro das condições do mercado vigente. Valor de mercado é o preço justo pago por um imóvel por um comprador desejoso de comprar para um vendedor desejoso de vender, ambos com pleno

conhecimento do seu aproveitamento eficiente (THOFEHRN, 2010 *apud* TRIVELLONI e HOCHHEIM, 1998, p. 26).

A NBR 14653-2 (2011) em seu item 3.1 define aproveitamento eficiente como sendo aquele recomendável e tecnicamente possível para o local, numa data de referência, observado a atual e efetiva tendência mercadológica nas circunvizinhanças, entre os diversos usos permitidos pela legislação pertinente.

Dantas (2005) reforça que o mercado é formado por três componentes: os bens levados a mercado, as partes desejosas em vendê-los e as partes interessadas em adquiri-los. Quando se tratam de bens imóveis, esses três componentes forAM o mercado imobiliário.

Na seara das metodologias normalizadas, as mesmas procuram atender às necessidades do mercado imobiliário tradicional, priorizando os imóveis de utilidade particular, desejabilidade econômica do lucro e temporaneidade conhecida (GOMIDE, 2008).

A Associação Brasileira de Normas Técnicas (ABNT), por meio da NBR 14653-1 (2019), relaciona e define os métodos para identificar o valor de um bem, de seus frutos e direitos:

- Método comparativo direto de dados de mercado: Identifica o valor de mercado do bem por meio de tratamento técnico dos atributos dos elementos comparáveis, constituintes da amostra.
- Método involutivo: Identifica o valor de mercado do bem, alicerçado no seu aproveitamento eficiente, baseado em modelo de estudo de viabilidade técnico-econômica, mediante hipotético empreendimento compatível com as características do bem e com as condições do mercado no qual está inserido, considerando-se cenários viáveis para execução e comercialização do produto.
- Método evolutivo: Identifica o valor do bem pelo somatório dos valores de seus componentes. Caso a finalidade seja a identificação do valor de mercado, deve ser considerado o fator de comercialização. Método da capitalização da renda: Identifica o valor do bem, com base na capitalização presente da sua renda líquida prevista, considerando-se cenários viáveis.

A NBR 14653-1 (2019) também aborda em seu texto os métodos para identificar o custo de um bem. Eles se dividem em Método comparativo direto e Método da quantificação de custo. O primeiro visa identificar o custo do bem por meio de tratamento técnico dos atributos dos elementos comparáveis, constituintes da amostra, já o segundo busca identificar o custo do bem

ou de suas partes por meio de orçamentos sintéticos ou analíticos a partir das quantidades de serviços e respectivos custos diretos e indiretos.

Outro conceito importante trazido pela NBR 14653-1 (2019) versa sobre os métodos para identificar indicadores de viabilidade da utilização econômica de um empreendimento. Neste quesito, os procedimentos avaliatórios usuais com a finalidade de determinar indicadores de viabilidade da utilização econômica de um empreendimento são baseados no seu fluxo de caixa projetado, a partir do qual são determinados indicadores de decisão baseados no valor presente líquido, taxas internas de retorno, tempos de retorno, dentre outros.

Sobre qual metodologia deve ser escolhida para valorar um determinado imóvel, a NBR 14653-1 (2019) ressalta que a mesma deve ser compatível com a natureza do bem avaliando, a finalidade da avaliação e os dados de mercado disponíveis. Para a identificação do valor de mercado, sempre que possível preferir o método comparativo direto de dados de mercado.

De acordo com Dantas (2005), as variáveis potencialmente influentes são estabelecidas, a *priori*, com base em teorias existentes, conhecimentos adquiridos em trabalhos anteriores, entre outros. Contudo, no decorrer dos trabalhos, outras variáveis podem se revelar como importantes.

No meio técnico é consenso que os imóveis possuem um comportamento distinto economicamente de outros bens, devido aos efeitos de seus atributos especiais, especialmente o custo elevado, a heterogeneidade, a imobilidade e a durabilidade.

Em relação à inferência estatística, tem-se que seu principal objetivo é estimar as características da população (parâmetros) a partir do conhecimento das características de uma amostra dela extraída (estatísticas) (DANTAS, 2005).

Para Radegaz (2011), a inferência estatística envolve a formulação de certos julgamentos (ou conclusões) sobre um todo, após examinar apenas uma parte ou amostra dele. Para que a inferência estatística seja válida, a amostra deve ser representativa da população, e a probabilidade do erro, ser especificada.

Dantas (2005) ressalta que inferir significa concluir. Assim, inferir estatisticamente significa tirar conclusões com base em medidas estatísticas. Em Engenharia de Avaliações o que se pretende é explicar o comportamento do mercado que se analisa com base em alguns dados levantados no mesmo. Neste caso a inferência estatística é fundamental para solucionar a questão, pois conhecendo-se apenas uma parte do mercado pode-se concluir sobre o seu comportamento, com determinado grau de confiança.

Radegaz (2011) reforça que o objetivo da inferência por meio da análise de regressão é encontrar uma função linear que permita compreender a relação entre os elementos, além de estimar uma variável em função de uma ou mais variáveis.

Quando o valor da variável desconhecida é dependente dos valores de mais de uma variável conhecida, chama-se de regressão múltipla. Gonzáles (2003), em sua explanação sobre os coeficientes da equação de regressão, ressalta que os mesmos, geralmente são estimados por meio do Método dos Mínimos Quadrados, que minimiza a soma dos quadrados dos resíduos.

A inferência estatística exige muita experiência em avaliação de imóveis porque a ausência de variáveis importantes ou a inclusão de variáveis inadequadas pode conduzir a erros gravíssimos (THOFEHRN, 2010 *apud* FIKER, 2008).

Em muitas situações, os modelos de avaliação de imóveis têm dificuldades na determinação das variáveis que influenciam no seu valor, sendo que para obter precisão na avaliação muitos fatores devem ser considerados, mas nem sempre é possível chegar a um modelo único que represente a realidade do mercado. As variáveis que influenciam o valor de uma amostra podem não ser as mesmas para outra, inclusive localizada na mesma região.

Em muitos casos, é necessário excluir elementos da amostra, por serem muito diferentes dos demais e por influenciarem fortemente nos valores gerais da equação de regressão (ROCHA, 2005 *apud* TRIVELLONI e HOCHHEIM, 1998).

Na prática trabalha-se com modelos lineares ou linearizáveis, por facilidades no cálculo das estimativas das médias e facilidades de interpretação. Os modelos linearizáveis são aqueles que podem ser transformados em lineares pela simples transformação nas escalas das variáveis envolvidas (DANTAS, 2005).

Na maioria das situações, o Engenheiro de Avaliações vai observar que são diversas as variáveis que influenciam na formação do valor de mercado de um imóvel. Assim, o profissional deve procurar identificar estas variáveis e encontrar o modelo explicativo do valor por meio das regressões múltiplas (MENDONÇA, 1998).

Dantas (2005) afirma que o modelo de regressão linear múltipla deve ser adotado quando mais de uma variável independente é necessária para explicar a variabilidade dos preços praticados no mercado.

Uma linha de regressão, também chamada de linha de melhor ajuste, é a linha para qual a soma dos quadrados dos resíduos é um mínimo e sua equação pode ser usada para prever os valores de y para um dado valor de x (LARSON, 2010). Na modelagem, devem ser expostas as hipóteses relativas aos comportamentos das variáveis dependentes e independentes, com base

no conhecimento que o engenheiro de avaliações tem a respeito do mercado, quando serão formuladas as hipóteses nulas e alternativas para cada parâmetro (NBR 14653-2, 2011).

Gonzáles (2003) destaca que o engenheiro de avaliações deve estipular modelos com as hipóteses de relacionamento entre as variáveis, que devem ser testadas pelos critérios estatísticos, verificando-se a validade destas hipóteses, ou seja, se os modelos são capazes de representar o segmento de mercado em questão.

Para tanto, devem ser coletados dados de transações (evidências do mercado), analisando-se o ajuste dos modelos considerados a estes dados, dentro de um determinado grau de precisão. Os testes estatísticos permitem avaliar o próprio modelo e a importância individual das variáveis incluídas, indicando a qualidade geral do modelo formulado.

Após a sumarização dos dados, o avaliador deverá partir em busca de modelos explicativos do mercado, utilizando técnicas da inferência estatística. Na realidade, estes modelos são uma representação simplificada do mercado, uma vez que não levam em conta todas as suas informações (população), mas é construído considerando-se apenas uma parte do mesmo (amostra), por isso precisa de cuidados científicos na sua elaboração, para fornecer respostas confiáveis (DANTAS, 2005).

Sobre a transformação logarítmica, Dantas (2005) enfatiza que a mesma é a escolhida quando se procura ajustar modelos a dados de valores imobiliários. É bastante coerente a sua utilização uma vez que as variáveis explicadas possuindo valores no campo dos reais positivos garante que o campo de variação dos valores ajustados correspondentes também será reais positivos.

Outro aspecto importante é que a transformação logarítmica na variável explicada torna o modelo multiplicativo, característica esta sugerida pelas próprias normas brasileiras que versam sobre avaliações, bem como pode estabilizar a variância do modelo. Conforme o mesmo autor, durante a modelagem muitas etapas de análise são necessárias, tais como a verificação de dados atípicos, variáveis pouco significativas, tendências, pressupostos sobre os erros aleatórios etc.

O processo de análise de regressão exige o respeito aos chamados “pressupostos básicos", e outras condições relacionadas, que precisam ser atendidos para que a análise seja válida, e possam ser realizadas inferências (previsões) com a equação determinada (GONZÁLEZ, 2003).

Para que os modelos sejam considerados aptos, deve-se garantir que: exista homocedasticidade dos resíduos (a variância é constante), haja independência serial dos resíduos (não há autocorrelação), os resíduos obedeçam a distribuição normal, que a relação

entre as variáveis independentes e a variável dependente seja linear, que não ocorra colinearidade perfeita entre quaisquer variáveis independentes.

Além destes, o modelo deve ainda atender a outros requisitos em parte decorrentes dos próprios pressupostos básicos, tais como: as variáveis importantes devem ser incluídas (modelo especificado é similar ao real), que não existam observações espúrias (elementos claramente não adaptados ao modelo, chamados de *outliers*), que as variáveis independentes não sejam aleatórias (somente a variável dependente pode ser estocástica), os resíduos terem média nula e número de observações (tamanho da amostra) maior que o de coeficientes a ser estimado (GONZÁLEZ, 2003).

De acordo com Dantas (2005), escolhido o modelo, parte o avaliador para a interpretação dos parâmetros quanto aos aspectos de sensibilidade e elasticidade, bem como do comportamento do mercado em relação a cada variável, qualitativa e quantitativamente. Segundo o mesmo autor, a interpretação do modelo deve ocorrer quando a variável resposta encontra-se devidamente explicada na escala original.

Radegaz (2011) define o seguinte roteiro para a análise de regressão: análise do coeficiente de determinação, análise da significância dos regressores, análise dos valores do "t de *Student*", análise da coerência da equação, análise dos resíduos e gráficos, análise da autocorrelação (série temporal), Durbins-Watson, análise (verificação) da homocedasticidade, análise (verificação) da multicolinearidade, análise (verificação) da normalidade dos resíduos e análise do intervalo de confiança.

3.9 Planejamento Urbano e Avaliação de Imóveis Urbanos

O crescimento populacional, concentrado sobretudo em cidades médias e grandes, trouxe à tona questões ligadas à infraestrutura, habitação e uso do solo (MARICATO, 2013; BONDUKI, 2010). Nesse contexto, formou-se uma tradição de estudos urbanos marcada pela preocupação com a desigualdade na produção do espaço e com o papel das políticas públicas na orientação do desenvolvimento urbano (ROLNIK, 2015; MARICATO, 2013).

Villaça (1998), por sua vez, destacou que a segregação socioespacial é um elemento constitutivo da estrutura das cidades, lembrando que a valorização imobiliária depende, em grande medida, da localização e da acessibilidade. Maricato (2013) chamou atenção para a dualidade entre a cidade formal, regulada pelo Estado, e a cidade informal, marcada pela

autoconstrução e pela ausência de infraestrutura adequada. Já Rolnik (2015) trouxe o debate sobre a financeirização da cidade, mostrando como a lógica do crédito e do mercado imobiliário passou a moldar preços e oportunidades urbanas. Bonduki (2010), por fim, resgatou a trajetória das políticas habitacionais e urbanísticas, ressaltando a interação entre Estado, mercado e sociedade civil no desenho da cidade contemporânea.

Essas contribuições ajudam a compreender que o valor de um imóvel urbano não é determinado apenas por suas características físicas ou ambientais. Ele também resulta de escolhas políticas e de instrumentos de planejamento que organizam o espaço urbano. Planos diretores, zonas especiais de interesse social (ZEIS) e operações urbanas consorciadas, previstos no Estatuto da Cidade (Lei n.º 10.257/2001), são exemplos de mecanismos que interferem diretamente na valorização ou desvalorização de áreas. Em muitos casos, a simples proximidade a equipamentos públicos, áreas verdes ou centralidades comerciais pode elevar o valor imobiliário, enquanto a ausência desses elementos gera o efeito oposto.

A literatura de engenharia de avaliações corrobora essa perspectiva. Dantas (2005) lembra que a definição do valor de mercado, de acordo com a NBR 14.653, deve considerar não apenas os atributos intrínsecos do imóvel, como área ou padrão construtivo, mas também fatores ligados ao entorno urbano. Estudos como os de Batalhone *et al*. (2002) e Motta (1997) reforçam que amenidades ambientais, a exemplo de parques ou corpos hídricos, são incorporadas pelo mercado e se refletem no preço de transação.

Dessa forma, o diálogo entre planejamento urbano e avaliação imobiliária torna-se fundamental. A análise do valor do solo urbano não pode ser dissociada das políticas de planejamento, que moldam a expansão e a organização da cidade, nem tampouco das condições ambientais que influenciam a qualidade de vida. Enquanto o planejamento urbano busca assegurar a função social da propriedade e o direito à cidade, a avaliação de imóveis fornece métricas objetivas que podem orientar desde a tributação imobiliária (IPTU, ITBI) até políticas de habitação e regularização fundiária.

Por fim, observa-se que a incorporação de métodos quantitativos mais sofisticados, como os modelos lineares generalizados e os algoritmos de aprendizagem de máquina, cria uma ponte entre o urbanismo crítico brasileiro e a engenharia de avaliações. Essa integração abre espaço para uma agenda interdisciplinar, capaz de articular dimensões sociais, políticas e econômicas da cidade com instrumentos técnicos de avaliação mais transparentes e fundamentados.

4 METODOLOGIA

Com o intuito de aprimorar a capacidade preditiva e ampliar a compreensão dos fatores que influenciam os preços dos imóveis urbanos, esta pesquisa adotou uma abordagem comparativa entre modelos estatísticos tradicionais e técnicas de aprendizagem de máquina (AM). A utilização de diferentes métodos permitiu avaliar a robustez dos resultados, a sensibilidade das variáveis ambientais e a adequação dos modelos às especificidades do mercado imobiliário local. A hipótese central sustenta a existência de uma relação positiva entre atributos ambientais e os valores de mercado dos imóveis na cidade de Palmas-TO, especialmente no entorno de ativos naturais relevantes. Para isso, foram empregados o Modelo Linear (ML) e o Modelo Linear Generalizado (MLG), no contexto do Método dos Preços Hedônicos (MPH), além de quatorze algoritmos de AM, voltados à modelagem não linear e à identificação de padrões complexos nos dados.

A estimação da função hedônica pode ser realizada por meio de diversas técnicas estatísticas, e, uma delas é o Modelo Linear Generalizado (MLG), que foi um dos métodos utilizados para alcançar esse objetivo. O método dos preços hedônicos é uma técnica de estimação de preços que se concentra nas características individuais de um item para determinar seu valor. Em vez de simplesmente considerar o valor de mercado, o preço hedônico analisa como as características do item (tamanho, localização, qualidade, entre outros) afetam o preço dos imóveis.

Outro método empregado consistiu na utilização de modelos de Aprendizagem de Máquina. De acordo com Lantz (2015), o campo do Aprendizagem de Máquina é responsável pelo desenvolvimento de algoritmos capazes de transformar dados em conhecimento acionável, por meio de processos de generalização e abstração que permitem a computadores aprenderem com experiências passadas e realizarem previsões com base em padrões identificados.

De acordo com Jiang (2021), a Aprendizagem de Máquina corresponde à capacidade de um sistema computacional de imitar comportamentos inteligentes tipicamente humanos. Trata-se, em essência, de uma abordagem fundamentada na análise de dados, caracterizada pela identificação e reconhecimento de padrões. Diferentemente dos métodos tradicionais de previsão, amplamente utilizados na economia e em diversas áreas das ciências sociais aplicadas, os quais partem do pressuposto de relações pré-especificadas entre variáveis dependentes e independentes. Os modelos de Aprendizagem de Máquina flexibilizam essas relações,

permitindo a modelagem de padrões mais complexos sem a necessidade de hipóteses estatísticas rígidas.

4.1 Base de Dados

A base de dados utilizada nesta pesquisa foi fornecida pela Prefeitura Municipal de Palmas-TO, contendo informações referentes às transações imobiliárias registradas no âmbito do Imposto sobre Transmissão de Bens Imóveis (ITBI). As variáveis disponíveis para cada observação incluíram: a área do terreno, a área da benfeitoria e o valor da transação. Tais dados constituem uma importante fonte para a aplicação do Método dos Preços Hedônicos, bem como para os algoritmos de Aprendizagem de Máquina empregados nesta pesquisa, por refletirem valores efetivamente praticados no mercado imobiliário local.

Inicialmente, a base continha 12.518 registros. Contudo, considerando a natureza administrativa da fonte, foi necessário um rigoroso processo de filtragem e depuração dos dados, de modo a assegurar a consistência estatística e a representatividade da amostra. Diversas simulações foram conduzidas no ambiente R, com o objetivo de identificar e remover inconsistências, duplicidades e valores atípicos (*outliers*) que poderiam comprometer a validade dos modelos.

O tratamento dos *outliers* envolveu a utilização de técnicas gráficas, como boxplots aplicados ao valor unitário do metro quadrado e gráficos de dispersão do logaritmo do valor por metro quadrado em relação às principais variáveis explicativas, além de análises estatísticas mais robustas, como a padronização dos resíduos studentizados, a aplicação do critério interquartil (IQR) e a avaliação da distância de Cook. Esses procedimentos permitiram identificar observações extremas que, em grande medida, resultavam de erros de registro ou de transações atípicas em relação ao padrão do mercado.

Apesar do processo de refinamento, a base de dados apresenta algumas limitações. A principal delas refere-se à ausência de características construtivas dos imóveis, como padrão construtivo, estado de conservação e número de quartos, variáveis relevantes na explicação da formação dos preços. Outra limitação decorre da inclusão conjunta de diferentes tipologias de imóveis (lotes comerciais e residenciais, bem como edificações comerciais e residenciais) o que introduz heterogeneidade adicional à amostra e dificulta a modelagem de forma plenamente ajustada às especificidades de cada segmento. Mesmo diante dessas restrições, a base consolidada representa um esforço metodológico consistente para a obtenção de informações

confiáveis, permitindo a aplicação de modelos estatísticos e de Aprendizagem de Máquina em condições adequadas de análise.

Ao final desse processo de limpeza, obteve-se uma base refinada composta por 6.692 observações válidas, consideradas apropriadas para a estimação dos modelos. A base final foi então utilizada em duas frentes metodológicas: (i) para a estimação da função hedônica via Modelo Linear Generalizado (MLG) e Modelo Linear (ML), considerando a especificidade das distribuições não normais e a presença de heterocedasticidade; e (ii) para o treinamento e validação de diferentes algoritmos de Aprendizagem de Máquina (AM), visando ampliar a capacidade preditiva e explorar a complexidade estrutural das variáveis envolvidas.

Esse tratamento criterioso da base de dados assegurou maior robustez às análises realizadas e contribuiu significativamente para a consistência dos resultados apresentados, permitindo uma comparação válida entre os modelos estatísticos tradicionais e os métodos computacionais mais avançados.

4.2 Construção e Especificação dos Modelos

4.2.1 Modelo Linear Generalizado (MLG)

O trabalho de Rosen (1974) sugere que a estimação dos preços implícitos dos atributos deve ser realizada em duas etapas distintas. Na primeira, é importante selecionar a forma funcional que melhor se adequa aos dados, utilizando técnicas como a análise de regressão, por exemplo. Na segunda etapa, as derivadas parciais da função selecionada são calculadas para avaliar o impacto de um aumento adicional de uma determinada característica sobre o preço do produto.

O resultado deste cálculo é o preço implícito do atributo. Este método é amplamente utilizado na área de economia para estimar o valor dos atributos de bens e serviços, como, por exemplo, o espaço adicional em um automóvel, o tempo de viagem ao trabalho, entre outros. A estimativa dos preços implícitos dos atributos permite aos economistas avaliar de forma mais precisa o valor que os consumidores atribuem aos diferentes aspectos de um produto ou serviço.

O método de precificação hedônica é usado para estimar o valor de um bem, como um imóvel, com base em seus atributos, a estimação da função hedônica pode ser realizada por meio de diversas técnicas estatísticas, e, uma delas é o Modelo Linear Generalizado (MLG). Esses atributos podem incluir a localização, tamanho, idade, condição, características

específicas do imóvel e outras variáveis relevantes que afetam seu preço. No entanto, escolher a forma funcional adequada para o modelo de precificação hedônica pode ser desafiador, pois a relação entre o preço e os atributos pode ser complexa e não linear.

Existem várias formas funcionais que podem ser usadas na modelagem da precificação hedônica, incluindo linear (paramétrica, semiparamétrica e não paramétrica), semilogarítmica, logarítmica dupla e Box-Cox. A escolha da forma funcional dependerá do objetivo da pesquisa, da quantidade e qualidade dos dados disponíveis e da seleção de variáveis dependentes e independentes.

A transformação Box-Cox é uma das formas funcionais mais poderosas e flexíveis, pois pode acomodar uma ampla variedade de relações não lineares entre o preço e os atributos. No entanto, a transformação Box-Cox é de difícil aplicação e pode introduzir mais erros aleatórios no modelo. Isso pode afetar a precisão das estimativas e, portanto, pode ser necessário usar outras formas funcionais mais simples e diretas. Por exemplo, estudos mostraram que outras formas funcionais, como a semilogarítmica ou a logarítmica dupla, podem produzir resultados mais precisos em certas situações, especialmente quando há falta de variáveis explicativas relevantes.

Essas formas funcionais são relativamente menos complicadas de aplicar e podem ser mais eficazes na estimativa do preço do imóvel. Em suma, a escolha da forma funcional apropriada na precificação hedônica dependerá de uma variedade de fatores, incluindo o objetivo da pesquisa, a qualidade e quantidade dos dados disponíveis e a seleção de variáveis dependentes e independentes. A seleção deve ser baseada nessas considerações e pode ser necessário testar várias formas funcionais para determinar a que melhor se adequa ao modelo (JIM e CHEN, 2006).

A aplicação de modelos lineares generalizados (MLG) como alternativa ao modelo de regressão linear clássico tem sido explorada na literatura com o objetivo de superar limitações relacionadas à normalidade e homocedasticidade dos resíduos. Jim e Chen (2006), por exemplo, utilizaram o MLG para estimar uma função de preços hedônicos com o intuito de avaliar a influência de variáveis ambientais urbanas sobre os preços dos imóveis na cidade de Guangzhou, na China. Os autores aplicaram o método de máxima verossimilhança para a estimação dos parâmetros e constataram que a presença de determinados elementos ambientais, como parques e áreas verdes, exerceu impacto positivo e significativo sobre os preços dos imóveis, ao passo que atributos como a proximidade a vias de tráfego intenso apresentaram efeitos negativos. Os resultados obtidos mostraram-se consistentes com aqueles observados em

aplicações do modelo hedônico tradicional, evidenciando o potencial do MLG como ferramenta complementar na análise da valoração ambiental.

De acordo com Paixão (2015), a aplicação do modelo de preços hedônicos no mercado imobiliário é diversificada na literatura, incluindo a construção de índices de preços como um de seus subitens. Diversos temas foram incorporados à literatura de preços hedônica, como exemplificada por Freeman (1979), que usou o modelo para medir o impacto da qualidade do ar no preço dos imóveis; Clark (2006), que testou o efeito da poluição sonora no mercado imobiliário; Tita *et al*., (2006), que estimaram a desvalorização dos imóveis devido à criminalidade urbana; Gibbons e Machin (2003), que se concentraram no impacto da qualidade das escolas no valor dos imóveis e Sheppard (2010), que testou o efeito das amenidades culturais no valor das propriedades urbanas.

Uma das formas funcionais de equação estimada do método de preços hedônicos sugerida por Jim e Chen (2006) para avaliar a relação entre elementos ambientais urbanos e os preços dos imóveis na cidade de Guangzhou, na China, quando se assume uma relação linear, é dada pela equação (4.1):

$$P = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_n A_n + \varepsilon \tag{4.1}$$

Em que:

- P: preço do imóvel;
- A_1, A_2, A_n: atributos ambientais urbanos considerados (por exemplo, proximidade a parques, qualidade do ar, ruído, trânsito, etc.);
- β_0: intercepto da equação;
- β_1, β_2, β_n: coeficientes que medem a contribuição de cada atributo ambiental para o preço do imóvel;
- ε: erro.

Essa forma funcional do método de preços hedônicos é comumente utilizada na literatura, pois permite avaliar a contribuição relativa de cada atributo ambiental urbano para o preço dos imóveis, e pode ser facilmente estimada por meio de técnicas estatísticas padrão, como a regressão linear.

Jim e Chen (2006) ainda propuseram uma segunda forma funcional que assumia uma relação não-linear entre os atributos ambientais urbanos e os preços dos imóveis, expressada pela equação semi-logarítmica (4.2):

$$\log(P) = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_n A_n \tag{4.2}$$

4.2.2 Modelo Linear (ML)

A função do Modelo Linear (ML), conforme estabelecido na literatura econométrica, é fundamentada na premissa de que a variável dependente pode ser expressa como uma combinação linear de um conjunto de variáveis independentes, acrescida de um termo de erro aleatório. Assim, o modelo assume a seguinte forma geral:

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j \ X_J + \varepsilon \tag{4.3}$$

Onde:

- Y: representa a variável dependente, no caso, o valor o imóvel;
- β_0: intercepto;
- β_j: coeficientes de regressão associados às variáveis independentes Xj;
- ε: termo de erro aleatório, assumindo com média zero e variância constante.

A estrutura funcional do modelo linear parte do pressuposto de linearidade nos parâmetros, sendo adequado para captar relações proporcionais entre as variáveis explicativas e o fenômeno observado. A estimação dos parâmetros β_j é realizada por meio do método dos mínimos quadrados ordinários (MQO), que busca minimizar a soma dos quadrados dos resíduos, proporcionando estimadores não viesados, consistentes e eficientes sob os pressupostos clássicos.

No contexto desta pesquisa, o Modelo Linear é aplicado para estimar os efeitos de variáveis físicas, ambientais, locacionais e socioeconômicas sobre os preços dos imóveis urbanos, servindo como base comparativa para a aplicação de abordagens mais flexíveis, como o Modelo Linear Generalizado e os Algoritmos de Aprendizagem de Máquina.

4.2.3 Algoritmos de Aprendizagem de Máquina (AM)

Os algoritmos de Aprendizagem de Máquina (AM) vêm se consolidando como ferramentas robustas na modelagem estatística aplicada à análise de fenômenos complexos e multidimensionais, como é o caso da avaliação de imóveis urbanos. Esses algoritmos são capazes de explorar padrões ocultos nos dados, dispensando muitas das suposições restritivas impostas pelos métodos econométricos tradicionais.

Desse modo, no presente estudo, os modelos de AM foram empregados como estratégia complementar ao Modelo Linear Generalizado (MLG) com o intuito de aprimorar a capacidade preditiva dos valores dos imóveis em função de variáveis ambientais, locacionais, física e fiscal.

A AM, conforme descrito por Hansen (2021), compreende um conjunto de procedimentos utilizados, sobretudo, para estimações pontuais em contextos com grande volume de dados, número elevado de variáveis e estruturas relacionais não necessariamente conhecidas *a priori*. Tal abordagem se mostra compatível com a realidade empírica do mercado imobiliário, onde as relações entre atributos dos imóveis e seus respectivos preços nem sempre seguem padrões lineares ou homogêneos.

A estruturação do processo de aprendizagem de máquina pode ser compreendida a partir das etapas fundamentais descritas por Lantz (2015), que envolvem a coleta e preparação dos dados, a exploração inicial, a modelagem por meio do treinamento de algoritmos e, por fim, a avaliação do desempenho preditivo. Essas etapas são essenciais para garantir que o modelo seja capaz de aprender padrões relevantes a partir dos dados. Conforme destacado por James *et al*. (2023), a finalidade central da aprendizagem supervisionada é aplicar métodos estatísticos sobre uma base de dados rotulada para estimar uma função desconhecida *f*, de modo que uma função estimada $\hat{f}$ aproxime a variável resposta Y para qualquer observação (X,Y), isto é, $Y \approx \hat{f}(X)$. Essa abordagem permite prever resultados futuros com base em novos dados, sendo amplamente empregada em contextos de regressão e classificação.

Conforme Lantz (2015), os principais algoritmos aplicados em problemas de aprendizagem supervisionado incluem *Naive Bayes, K-Nearest Neighbor, Decision Trees, Linear Regression, Regression Trees, Model Trees, Neural Networks e Support Vector Machines*. O autor também destaca os algoritmos de meta-aprendizagem, como *Bagging, Boosting e Random Forests*, que têm como objetivo aprimorar a capacidade preditiva por meio da combinação de múltiplos modelos base.

A seguir, apresentam-se os principais algoritmos empregados nesta tese.

4.2.3.1 Modelos de Árvores de Decisão

As árvores de decisão são modelos que procuram reproduzir, de maneira estruturada, o processo de tomada de decisão adotado pelo ser humano, considerando passo a passo as possíveis alternativas e seus desdobramentos, ainda que sob uma lógica não linear (NIELSEN, 2021). Tais modelos apresentam ampla aplicabilidade, podendo ser empregados tanto em problemas de regressão quanto de classificação. Sua organização segue uma estrutura hierárquica, na qual cada nó interno expressa uma condição associada a uma variável, os ramos representam os resultados dessa condição, e as folhas correspondem à decisão ou ao valor predito, conforme o tipo de tarefa.

Como destaca Lantz (2013), trata-se de classificadores com elevada capacidade de interpretação e desempenho, capazes de representar as relações entre atributos e desfechos por meio de regras sucessivas e facilmente compreensíveis.

4.2.3.2 Classification Trees

Nos modelos conhecidos como *Classification Trees*, o principal objetivo é classificar cada observação com base na classe mais frequente entre os dados de treinamento localizados na mesma região da árvore decisória. Ao interpretar os resultados desse tipo de modelo, não se busca apenas identificar a classe atribuída em cada nó terminal, mas também compreender a distribuição das classes entre as observações de treinamento que compõem aquela região específica, permitindo avaliar a homogeneidade e a confiança da predição realizada (JAMES *et al.*, 2013).

Conforme apontam Hastie, Tibshirani e Friedman (2009), em problemas de classificação com múltiplas categorias (1, 2, ..., *K*), as adaptações necessárias no algoritmo das árvores de decisão concentram-se essencialmente nos critérios adotados para a divisão dos nós e para o processo de poda. A tarefa de fazer crescer uma árvore de classificação utiliza a divisão binária recursiva para desenvolver uma árvore de classificação, o critério para realizar a divisão binária é a taxa de erro de classificação.

Como a finalidade do modelo é atribuir a cada observação a classe mais frequente entre as amostras de treinamento presentes em uma determinada região da árvore, a taxa de erro de classificação é simplesmente a fração das observações de treinamento nessa região que não pertencem à classe mais comum:

$$E = 1 - \hat{p}_{mk} \tag{4.4}$$

Em que $\hat{p}_{mk}$ é a proporção de observações de treinamento na m-ésima região que são da *k*-ésima classe. No entanto, o erro de classificação não é suficientemente sensível para o cultivo de árvores e, na prática, duas outras medidas são preferíveis, o índice de *Gini* é um deles, definido por:

$$G = \sum_{k=1}^{k} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.5}$$

O índice de *Gini* é conhecido como uma medida de pureza do nó. Assim, um valor pequeno indica que um nó contém predominantemente observações de uma única classe.

4.2.3.3 Bagging

Segundo James *et al*. (2023), o bagging (abreviação de bootstrap aggregating) é uma técnica de aprendizagem de máquina desenvolvida para reduzir a variância de modelos preditivos, sendo especialmente eficaz no contexto de árvores de decisão, cuja natureza instável tende a amplificar pequenas variações nos dados de entrada. O procedimento consiste em gerar múltiplas amostras de treinamento a partir de reamostragem com reposição (bootstrap) e, em seguida, ajustar um modelo a cada subconjunto. A previsão final é obtida pela média (em problemas de regressão) ou pelo voto majoritário (em problemas de classificação) das previsões individuais. Esse mecanismo explora a propriedade estatística de que, ao se calcular a média de *n* observações independentes $Z_1, Z_2, \ldots, Z_n$, cada uma com variância σ^2, a variância da média $\underline{Z}$ é reduzida para σ^2/n, o que contribui diretamente para a melhoria da estabilidade e da precisão do modelo agregado.

Sabe-se que, dado um conjunto de *n* observações independentes $Z_1, Z_2, \ldots, Z_n$. cada um com variância σ^2, a variância da média das $\underline{Z}$ das observações é dada por σ^2/n.

De acordo com James *et al*. (2013), uma estratégia eficiente para diminuir a variância de um modelo de aprendizagem de máquina e, ao mesmo tempo, elevar sua acurácia preditiva, é construir múltiplos modelos a partir de diferentes subconjuntos da base de dados de treinamento.

Ao combinar as previsões desses modelos, obtém-se uma estimativa mais estável e robusta. Desta maneira, é possível construir um modelo de predição separado, usando cada

conjunto de treinamento e calcular a média das previsões resultantes, que é calculada por: $\hat{f}^1$(x), $\hat{f}^2$(x), ... , $\hat{f}^n$(x), utilizando *B* conjuntos de treinamento separados para calcular a média deles a fim de obter um único modelo de aprendizagem de máquina de baixa variância, dado por:

$$\hat{f}_{médio} \quad (x) = \frac{1}{B} \sum_{b=1}^{B} \quad \hat{f}^b \ (x) \tag{4.6}$$

Contudo, como observam James *et al*., (2013), na prática nem sempre se dispõe de múltiplos conjuntos distintos de treinamento. Diante disso, uma alternativa é gerar amostras repetidas a partir da própria base original de dados, por meio de técnicas de reamostragem como o *bootstrap*. Nessa abordagem, é gerado *B* conjuntos de dados de treinamento, por meio de *bootstrapp* diferentes. Em seguida, treina-se o método no *b*-ésimos conjuntos de treinamentos para obter $\hat{f}^b$(x), e assim, obter as médias de todas as previsões. O resultado final é chamado de *bagging*:

$$\hat{f}_{bag} \quad (x) = \frac{1}{B} \sum_{b=1}^{B} \quad \hat{f}^b(x) \tag{4.7}$$

O método de *bagging*, de modo geral, tende a melhorar a precisão preditiva em relação ao uso isolado de uma única árvore de decisão. No entanto, essa melhoria ocorre à custa da interpretabilidade do modelo. Enquanto as árvores individuais possuem como uma de suas principais vantagens a facilidade de interpretação, o uso de múltiplas árvores dificulta a representação do processo de decisão em um formato único e compreensível. Além disso, tornam-se menos evidente quais variáveis exercem maior influência no modelo final. Assim, conforme apontam James *et al*. (2013), o *bagging* oferece ganhos em desempenho preditivo, mas apresenta uma maior dificuldade na interpretação.

4.2.3.4 Random Forests

O *Random Forest* é um algoritmo estruturado a partir de um conjunto de árvores de decisão. Desenvolvido por Leo Breiman e Adele Cutler em 2001, esse método combina os fundamentos do *bagging* com a seleção aleatória de variáveis em cada divisão da árvore,

promovendo maior diversidade entre os modelos gerados. Após a construção do conjunto de árvores - a chamada "floresta", as previsões individuais são agregadas por meio de um sistema de votação, resultando na predição final. Essa abordagem, conforme explica Lantz (2013), contribui para reduzir a variância e melhorar o desempenho preditivo do modelo.

O método *Random Forest* aprimora o desempenho das árvores geradas via *bagging* ao incorporar uma etapa adicional de aleatoriedade no processo de construção dos modelos. Assim como no *bagging*, diversas árvores de decisão são geradas a partir de amostras *bootstrap* da base de treinamento. No entanto, ao construir essas árvores de decisão, cada vez que uma divisão em uma árvore é considerada uma amostra aleatória de *m* preditores é escolhida como candidata à divisão do conjunto completo de *p* preditores. A divisão pode usar apenas um desses *m* preditores. Assim, uma nova amostra de *m* preditores é obtida em cada divisão e, normalmente, se escolhe $m \approx \sqrt{p}$.

Na construção de um modelo *Random Forest*, o algoritmo restringe, propositalmente, o número de preditores considerados a cada divisão da árvore, impedindo que a maioria das variáveis disponíveis seja avaliada simultaneamente. Isso é especialmente relevante em situações em que existe um preditor extreAMente dominante, acompanhado de outros com influência moderada. Em modelos baseados em *bagging*, é comum que esse preditor mais forte seja selecionado repetidamente nas divisões iniciais de praticamente todas as árvores, o que resulta em estruturas muito semelhantes entre si.

Como consequência, as previsões geradas por essas árvores tendem a ser altamente correlacionadas. A média de muitas quantidades, altamente correlacionadas, não leva a uma redução tão grande na variância quanto à média de muitas quantidades não correlacionadas. Em particular, isso significa que *bagging* não levará a uma redução substancial na variância em uma única árvore nesse cenário (JAMES *et al*., 2013).

James *et al*. (2013) destacam ainda que, a principal diferença entre *bagging* e o *Random Forest* é a escolha do tamanho do subconjunto do preditor *m*. A título de ilustração, se um *Random Forest* é construído definindo *m* = *p*, isso equivale, simplesmente, ao *bagging*. O *Random Forest*, usando $m = \sqrt{p}$, proporciona uma redução no erro de teste e no erro *outof-bag*, sobre o *bagging*. Usar um pequeno valor de *m* na construção de um *Random Forest* normalmente será útil quando temos um grande número de preditores correlacionados.

4.2.3.5 Algoritmo de Support Vector Machine (SVM)

O SVM, proposto por Cortes e Vapnik (1995), é uma metodologia voltada à tarefa de classificação que ganhou ampla aceitação, especialmente na área da ciência da computação. Segundo James *et al*. (2013), os modelos baseados em SVM têm demonstrado desempenho consistente em diferentes contextos de aplicação e são frequentemente reconhecidos como uma das abordagens mais eficazes para problemas de classificação.

A finalidade do SVM é construir um limite plano, denominado hiperplano, que divida o espaço de forma a gerar partições homogêneas em ambos os lados. O SVM é uma extensão do classificador do vetor suporte. A solução para resolver o problema do classificador envolve apenas os produtos internos das observações (em oposição às próprias observações). O produto interno de dois vetores *a* e *b* é dado por:

$$\langle a, b \rangle = \sum_{i=1}^{r} a_i b_i \tag{4.8}$$

De modo que o produto interno entre duas observações x_i e x_j é:

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_i x_{i'} \tag{4.9}$$

Assim, o classificador de vetor de suporte linear é definido por:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x_i, x'_i \rangle \tag{4.10}$$

Em que temos *n* parâmetros αi, onde $i = 1, \ldots, n$, um por observação.

De forma resumida, ao definir o classificador linear *f(x)* e estimar seus coeficientes, os produtos internos desempenham um papel fundamental. Dessa maneira, sempre que o produto interno da equação (4.9) surgir na equação (4.10) ou em qualquer etapa do cálculo da solução do classificador de vetor de suporte é possível substituí-lo por uma generalização desse produto interno, expressa na forma:

$$k(x_i, x_{i'})$$

(4.11)

Em que o *K* é uma função definida como Kernel. James *et al*. (2013). O Kernel é uma função que quantifica a similaridade entre duas observações. Para exemplificar, tem-se:

$$k(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

(4.12)

A equação (4.12) corresponde a um kernel do tipo linear, uma vez que o classificador de vetor de suporte assume linearidade em relação aos atributos. Esse kernel linear mede a similaridade entre duas observações com base no produto interno, análogo a um coeficiente de correlação. Contudo, é possível adotar outras formas funcionais para substituir a equação (4.12), como, por exemplo:

$$k(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} \quad x_{ij} x_{i'j}\right)^{d}$$

(4.13)

A equação (4.13) é denominada de Kernel Polinomial de grau *d*, em que *d* é um número inteiro positivo. Se utilizado o Kernel com *d* > 1, em vez do Kernel linear padrão, o algoritmo do classificador de vetor de suporte leva a um limite de decisão mais flexível.

4.2.3.6 Algoritmo K-Nearest Neighbors (KNN)

O algoritmo *K-Nearest Neighbors* (KNN), ou K-vizinhos mais próximos, baseia-se na análise dos vizinhos mais próximos de uma determinada observação para realizar a classificação de exemplos ainda não rotulados. O parâmetro *k* representa o número de vizinhos considerados, podendo assumir diferentes valores conforme a especificidade do problema.

Conforme apontam Izbicki e Santos (2019), o KNN é aplicável tanto a tarefas de classificação quanto de regressão. Diferentemente de outros métodos, o KNN não constrói um modelo explícito ou uma estrutura paramétrica, ele simplesmente armazena todas as instâncias do conjunto de treinamento em um espaço n-dimensional.

A predição de novos casos é feita com base em medidas de similaridade entre as observações, e a classificação resulta, geralmente, de uma votação simples entre os *k* vizinhos mais próximos. Uma vez definido o valor de *k,* o algoritmo requer uma base de dados de treinamento com exemplos previamente classificados. Para cada nova entrada no conjunto de teste, o KNN identifica as *k* observações mais próximas e similares no conjunto de treinamento, sobre as quais baseia sua predição.

Para um dado valor de *K* e um ponto de predição x_0, a regressão KNN inicialmente identifica as observações de treinamento *K* que estão mais próximas de x_0, representado por N_0. Em seguida, estima $f(x_0)$ usando a média de todas as respostas de treinamento em N_0. Assim, descrever a estimação de $f_{(x_0)}$:

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in N_0} y_i \tag{4.14}$$

Segundo James *et al*. (2023), a escolha do valor de *K* no algoritmo dos K Vizinhos Mais Próximos (k-NN) está diretamente relacionada ao equilíbrio entre viés e variância. Quando *K* é pequeno, o modelo tende a se ajustar com maior flexibilidade aos dados, resultando em baixo viés, porém elevada variância, já que a previsão em uma dada região do espaço de entrada depende de um número muito reduzido de observações, podendo sofrer grande influência de ruídos ou *outliers*. Por outro lado, valores maiores de *K* produzem predições mais estáveis, com menor variância, pois a resposta em determinada região passa a refletir a média de múltiplos pontos vizinhos. Essa suavização implica em maior viés, mas reduz o risco de sobreajuste, evidenciando a clássica compensação viés-variância presente em métodos de aprendizagem supervisionado. Entretanto, como aponta Lantz (2013), a suavização pode causar algum viés, o que pode ocultar parte da estrutura de $f(x_0)$.

Para o contexto de classificação, a equação (4.14) pode ser adaptada para a seguinte expressão:

$$g(x) = moda_{i \in N_{xyi}}$$

(4.15)

Em que N_x é o conjunto de *k* observações mais próximas de *x*. No qual o objetivo é identificar a classe mais frequentemente observadas entre as observações mais próximas ao vetor de covariáveis *x* de interesse.

4.2.3.7 Redes Neurais Artificiais (RNAs)

As Redes Neurais Artificiais (RNAs) constituem um dos conceitos mais antigos no campo da inteligência artificial, tendo sido inicialmente propostas por McCulloch e Pitts (1943) e posteriormente desenvolvidas por Rosenblatt (1958). De acordo com Géron (2019), as RNAs figuram entre as ferramentas mais relevantes e consolidadas no escopo dos modelos de aprendizagem de máquina, destacando-se por sua versatilidade e elevada capacidade de modelagem em diferentes contextos. Do ponto de vista matemático, no contexto de regressão, trata-se de um estimador não linear de *r(x)*.

Uma Rede Neural Artificial (RNA) pode ser entendida como uma estrutura computacional composta por camadas de unidades interconectadas, organizadas em três blocos fundamentais: a camada de entrada, a(s) camada(s) oculta(s) e a camada de saída. Na camada de entrada, cada unidade representa uma variável explicativa do modelo, que transmite seus valores às unidades da camada oculta por meio de conexões ponderadas. A camada oculta é composta por neurônios artificiais que processam combinações lineares dos sinais recebidos, aplicando funções de ativação não lineares — como a função sigmoide, ReLU ou tangente hiperbólica — que introduzem flexibilidade ao modelo. Por fim, a camada de saída agrega os sinais da camada oculta para produzir a estimativa final da rede, seja para regressão ou classificação.

Segundo James *et al*. (2023), as RNAs com ao menos uma camada oculta e um número suficiente de unidades são capazes de aproximar funções complexas com precisão arbitrária, conforme estabelece o teorema da aproximação universal. A capacidade de capturar relações não lineares e interações de alta ordem entre as variáveis explicativas torna esse modelo particularmente útil em contextos em que os métodos lineares tradicionais apresentam limitações. Como destaca Lantz (2015), o processo de treinamento da rede envolve a propagação direta das entradas (feedforward) seguida pelo ajuste iterativo dos pesos das conexões, com base no erro entre a saída prevista e o valor real, por meio do algoritmo de

retropropagação (backpropagation). Esse procedimento permite que a rede aprenda padrões complexos presentes nos dados, tornando-se uma ferramenta poderosa na modelagem preditiva.

De acordo com Izbicki e Santos (2019), considerando uma Rede Neural Artificial com três variáveis de entrada representadas por $x = (x_1, x_2, x_3)$, o valor de ativação de um dado neurônio *j* na camada oculta é obtido por meio de uma combinação linear dos elementos de entrada ponderados por seus respectivos pesos, somada a um termo de viés. Essa combinação é então transformada por uma função de ativação não linear, conforme a expressão:

$$x_j^1 = f\left(\beta_0^0, j + \sum_{i=1}^{3} \beta_i^0, j, x_i^0\right) \tag{4.16}$$

Em que $x_i^0 = x_i$ para *i* = 1, 2, 3 e a função *f* é uma função definida pelo pesquisador e é chamada de função de ativação.

O Perceptron de Múltiplas Camadas (MLP) configura-se como uma extensão do Perceptron simples e representa uma das arquiteturas mais tradicionais de redes neurais artificiais no campo da Aprendizagem de Máquina. Ao contrário do Perceptron de camada única, o MLP possui uma ou mais camadas ocultas situadas entre as camadas de entrada e saída, compostas por neurônios artificiais responsáveis por transformações não lineares das entradas. Segundo Lantz (2015), essas camadas intermediárias permitem que o modelo aprenda padrões complexos nos dados, sendo especialmente útil para problemas em que as relações entre variáveis não são lineares ou são altamente interdependentes. Cada unidade da rede realiza uma combinação linear dos sinais de entrada, a qual é então processada por uma função de ativação, como a função logística ou ReLU, que introduz não linearidade ao sistema. Conforme exposto por James *et al*. (2023), o treinamento dessas redes é realizado por meio do algoritmo de retropropagação, que ajusta iterativamente os pesos das conexões para minimizar a função de perda. A utilização do MLP é particularmente vantajosa em tarefas de regressão e classificação supervisionadas, sobretudo quando se dispõe de um volume de dados suficiente para capturar a complexidade do fenômeno estudado. Complementarmente, Morettin e Singer (2020) ressaltam que a representação em múltiplas camadas torna esse tipo de rede uma estrutura universal de aproximação de funções, o que explica sua ampla adoção em contextos de previsão e análise preditiva.

4.3 Critérios de Avaliação dos Modelos

A avaliação do desempenho dos modelos estatísticos e de aprendizagem de máquina utilizados nesta pesquisa constitui uma etapa fundamental para verificar a capacidade preditiva das abordagens implementadas.

Tanto para o Modelo Linear (ML), amplamente utilizado na literatura em estudos baseados no Método dos Preços Hedônicos (MPH), quanto para o Modelo Linear Generalizado (GLM), bem como para os Modelos de Aprendizagem de Máquina (AM), a comparação de desempenho baseia-se em métricas estatísticas consolidadas, que expressam a qualidade do ajuste e a precisão das previsões.

Conforme destaca Paixão (2025), a escolha criteriosa dessas métricas permite não apenas identificar o modelo mais eficiente, mas também compreender os limites e vantagens de cada abordagem no contexto da avaliação de imóveis urbanos.

4.3.1 Avaliação do Modelo Linear (ML)

Para o Modelo Linear, as métricas de avaliação adotadas foram: o coeficiente de determinação ajustado (R^2 ajustado), o erro padrão residual (ou Root Mean Square Error - RMSE), a estatística F de significância conjunta e os valores-*p* dos coeficientes estimados.

O coeficiente de determinação ajustado, expressa a proporção da variabilidade dos preços dos imóveis que é explicada pelas variáveis independentes do modelo, ajustada ao número de preditores, evitando assim a superestimação da qualidade do ajuste quando múltiplos preditores são incluídos, conforme equação (4.17):

$$R^2_{ajustado} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right) \tag{4.17}$$

Onde:

- R^2*:* coeficiente de determinação;
- *n*: número de observações;
- *k*: número de preditores (excluindo o intercepto).

Outro critério relevante é o Erro Padrão da Estimativa (EP), ou RMSE, expresso pela equação (4.18):

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - k - 1}}$$

(4.18)

Essa medida reflete a magnitude média dos erros de predição em unidades monetárias, sendo especialmente útil para comparar o desempenho de diferentes modelos sob a mesma base de dados.

Além dessas métricas, a avaliação do Modelo Linear requer a análise da estatística F de significância conjunta, que testa a hipótese nula de que todos os coeficientes das variáveis explicativas (exceto o intercepto) são simultaneamente iguais a zero, ou seja, que o modelo não possui poder explicativo significativo em relação à variável dependente. A equação geral da estatística F é definida por:

$$F = \frac{\left(SQ_{reg}/k\right)}{\left(SQ_{res}/(n - k - 1)\right)}$$

(4.19)

Onde:

- SQ_{reg}: soma de quadrados da regressão;
- SQ_{res}: soma de quadrados dos resíduos.

Um valor de F elevado, com base na distribuição F de Snedecor, indica que ao menos uma das variáveis explicativas está significativamente associada à variável dependente, rejeitando-se, assim, a hipótese nula. O valor-*p* associado à estatística F também é calculado e interpretado conforme o nível de significância adotado (geralmente, 1%, 5% ou 10%).

Complementarmente, a análise dos valores-*p* individuais dos coeficientes estimados fornece evidências sobre a contribuição estatística de cada variável explicativa no modelo. O valor-*p* é derivado do teste *t* para cada coeficiente $\hat{\beta}_j/\beta_j$ sendo calculado com base na estatística:

$$t = \frac{\hat{\beta}_j}{EP\left(\hat{\beta}_j\right)}$$

(4.20)

O valor-*p* resultante indica a probabilidade de se obter um coeficiente tão extremo quanto o observado, caso a hipótese nula ($\beta_j = 0$) seja verdadeira. Valores-*p* inferiores ao nível de significância estabelecido, sugerem que a variável *j* tem efeito estatisticamente significativo sobre a variável dependente e, portanto, deve ser mantida no modelo.

4.3.2 Avaliação do Modelo Linear Generalizado (MLG)

No caso do Modelo Linear Generalizado (GLM), embora a estrutura de avaliação seja análoga ao modelo linear clássico, é importante destacar que o coeficiente de determinação tradicional (R^2) não se aplica diretamente em função da utilização de procedimentos de estimação baseados na máxima verossimilhança. Assim, a qualidade do ajuste foi avaliada por meio de medidas alternativas, com destaque para o *pseudo* R^2, obtido neste estudo com base na correlação quadrada entre os valores ajustados ($\hat{n}$) e a inversa das transações ($\tilde{n}$):

$$Pseudo\ R^2 = \left(cor(\hat{n}, \tilde{n})\right)^2 \tag{4.21}$$

Adicionalmente, outros critérios foram empregados para avaliar o GLM:

- Erro padrão residual (*dispersion parameter*): estima a variância dos resíduos sob a distribuição especificada;
- Valores-*p* dos coeficientes estimados: obtidos com base em testes *z*, derivados das estimativas e seus respectivos erros padrão;
- Diagnóstico de heterocedasticidade: realizado por meio dos testes de Breusch-Pagan e White, essenciais para verificar a adequação das suposições do modelo quanto à homocedasticidade;
- Análise gráfica dos resíduos: incluindo gráficos de resíduos versus ajustados,
- Q-Q Plot, Scale-Location e resíduos versus leverage, fundamentais para avaliar linearidade, normalidade, homocedasticidade e presença de pontos influentes;
- Intervalos de confiança via *bootstrap*: empregados para robustecer a inferência dos coeficientes, especialmente frente à violação de pressupostos clássicos.

Em suma, as métricas e procedimentos adotados permitiram avaliar de forma abrangente o desempenho e a validade dos modelos estatísticos estimados nesta pesquisa, bem como suas limitações frente à complexidade inerente ao mercado imobiliário urbano.

4.3.3 Avaliação dos Modelos de Aprendizagem de Máquina (AM)

Para os modelos de aprendizagem de máquina, a avaliação seguiu critérios consagrados em problemas de regressão supervisionada, permitindo a comparação direta entre abordagens tradicionais e modernas.

As métricas utilizadas incluem o Erro Médio Absoluto (MAE), o Erro Quadrático Médio da Estimativa (RMSE), o Erro Percentual Absoluto Médio (MAPE) e o Coeficiente de Determinação (R^2). O MAE expressa a média das diferenças absolutas entre os valores observados e previstos:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.22}$$

Onde:

- y_i: valor observado da variável dependente;
- $\hat{y}_i$: valor previsto pelo modelo;
- n: número total de observações.

Já o RMSE, conforme descrito anteriormente, penaliza erros maiores, sendo especialmente sensível a *outliers*. O MAPE, por sua vez, indica o erro percentual médio em relação ao valor real, sendo útil para avaliar a precisão relativa:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i} \tag{4.23}$$

Onde:

- y_i: valor observado da variável dependente;
- $\hat{y}_i$: valor previsto pelo modelo;
- n: número total de observações.

O coeficiente de determinação R^2 também é amplamente utilizado na literatura para quantificar o grau de explicação da variabilidade do valor dos imóveis pelos atributos incorporados ao modelo de AM:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4.24}$$

Onde:

- y_i: valor observado da variável dependente;
- $\hat{y}_i$: valor previsto pelo modelo;
- $\overline{y}$: média dos valores observados;
- n: número total de observações.

Conforme discutido por James *et al*. (2023), embora o coeficiente de determinação R^2 forneça uma medida intuitiva da qualidade do ajuste, sua interpretação isolada pode ser enganosa, especialmente em contextos de alta dimensionalidade ou modelos altamente flexíveis. Dessa forma, Lantz (2015) e Morettin e Singer (2020) recomendam a análise conjunta de diferentes métricas de desempenho, como o erro quadrático médio (RMSE), o erro absoluto médio (MAE) e o erro percentual absoluto médio (MAPE), para avaliar a precisão e a generalização do modelo.

Adicionalmente, conforme orientações metodológicas presentes em Lantz (2015), a avaliação dos modelos de AM foi complementada pela técnica de validação cruzada *k*-fold e pela análise do erro *out-of-bag* (OOB) nos métodos de ensamble, como *Random Forest* e *Bagged Trees*. Tais procedimentos conferem maior robustez às estimativas ao reduzir o risco de *overfitting* e assegurar que os resultados obtidos são representativos para novas observações.

4.3.4 A variável dependente: preço da transação

A variável dependente utilizada foi o preço nominal da transação, constante na base de dados do ITBI. No entanto, trabalhar com os dados do ITBI apresenta alguns desafios, como a possibilidade de subdeclaração ou de valores muito diferentes da realidade de mercado.

Para evitar problemas de subdeclaração, as prefeituras mantêm um cadastro de preços avaliados para cada imóvel. Se houver divergência entre o valor declarado e o valor avaliado, o ITBI é calculado com base no valor mais alto. Além disso, a informação de preço no ITBI pode sofrer com erros de leitura, gravação, transcrição e cálculo de dados, especificidade do imóvel ou da transação e deficiências na coleta de dados.

Para contornar esses problemas, González (1997a) sugere o uso de técnicas estatísticas simples, como gráficos de dispersão do preço por metro quadrado e/ou do logaritmo do preço. O autor também recomenda a remoção dos dados discrepantes ou a inclusão de *dummies* em casos em que o pesquisador consegue identificar a razão para o preço discrepante.

Para o presente estudo, foram suprimidas da base de dados do ITBI todas as transações em que o preço declarado foi menor que o preço avaliado pela prefeitura. Em seguida, foi realizada a análise gráfica de dispersão dos valores do logaritmo e, quando necessário, a análise de dispersão do preço do metro quadrado, de acordo com a proposta de González (1997a).

Os dados utilizados referem-se às transações imobiliárias registradas nos anos de 2019, 2020 e 2021 no município de Palmas-TO. Considerando que os valores das transações ocorreram em momentos distintos no tempo, foi necessário realizar a atualização monetária de todos os preços para a data-base de junho de 2023, utilizando-se como referência o Índice Nacional de Preços ao Consumidor Amplo (IPCA). Essa correção teve como objetivo eliminar os efeitos da inflação acumulada no período e assegurar a comparabilidade entre os valores analisados, independentemente do ano em que a transação foi efetivada.

4.3.5 Variáveis independentes

As variáveis independentes que foram incluídas nos modelos testados encontram-se no Quadro 1, que apresenta a sistematização das mesmas.

Quadro 1: Conjunto de variáveis independentes

Conjunto	**Variável**	**Descrição**	**Unidade de medida**	**Tipo**	**Sinal**	**Fonte de dados**
Atributos físicos do imóvel	Imóvel	Área do Imóvel	m^2	Contínua	+	ITBI/SEFAZ
	Terreno	Área do Terreno	m^2	Contínua	+	
Ativos Ambientais	Água	Distância à Praia da Graciosa	km	Contínua	-	Sistema de Informações Geográficas de Palmas - GEOPALMAS
	Verde	Distância ao Parque Cesamar	km	Contínua	-	

Fiscal	Zona Fiscal	Índice Fiscal	1 para melhor região e 5 para pior região	Escalar	-	Lei n.º 2.428, de 20 de dezembro de 2018. Dispõe sobre a Planta de Valores Genéricos do município de Palmas-TO.
	PVG	Valor do m^2 por quadra	R$/$m^2$	Contínua	+	
Dummy's – Proximidade ao Parque Cesamar	dummycp	Perto_Parque	1 se a distância ≤ 3,5 km, 0, caso contrário	Escalar		Sistema de Informações Geográficas de Palmas - GEOPALMAS
	dummycp	Médio_Praia	1 se 3,5 km < distância ≤ 7 km, 0 caso contrário.	Escalar		Sistema de Informações Geográficas de Palmas - GEOPALMAS
Econômico	Renda	Renda Familiar Mensal	R$	Contínua	+	IBGE

Fonte: Elaborada pelo autor. * m^2: metro quadrado; km: quilômetros.

Essas variáveis foram organizadas em cinco conjuntos temáticos: atributos físicos do imóvel, ativos ambientais, atributos fiscais, indicadores econômicos e variáveis *dummies* de proximidade. Cada uma delas foi selecionada com base na literatura correlata e na sua relevância para a formação do valor de mercado dos imóveis urbanos no contexto do plano diretor de Palmas-TO.

No que se refere aos atributos físicos do imóvel, foram consideradas duas variáveis contínuas: a área do terreno (m²) e a área construída da benfeitoria (m²). Ambas foram extraídas da base de dados do ITBI/SEFAZ e pressupõem relação positiva com o valor do imóvel, uma vez que se referem às características intrínsecas ao bem avaliado.

Foram eleitas duas variáveis explicativas para representar as amenidades ambientais ao longo do plano diretor da cidade de Palmas-TO: distância dos imóveis ao Parque Cesamar (VERDE) e à Praia da Graciosa (ÁGUA), como podem ser observadas na Figura 1.

Figura 1: Localização dos ativos ambientais no município de Palmas-TO

Fonte: Sistema de Informações Geográficas de Palmas – GEOPALMAS.

Para representar as amenidades ambientais em um modelo de precificação de imóveis, é comum utilizar variáveis que capturem a distância entre os imóveis e esses recursos.

Algumas das variáveis mais utilizadas incluem:

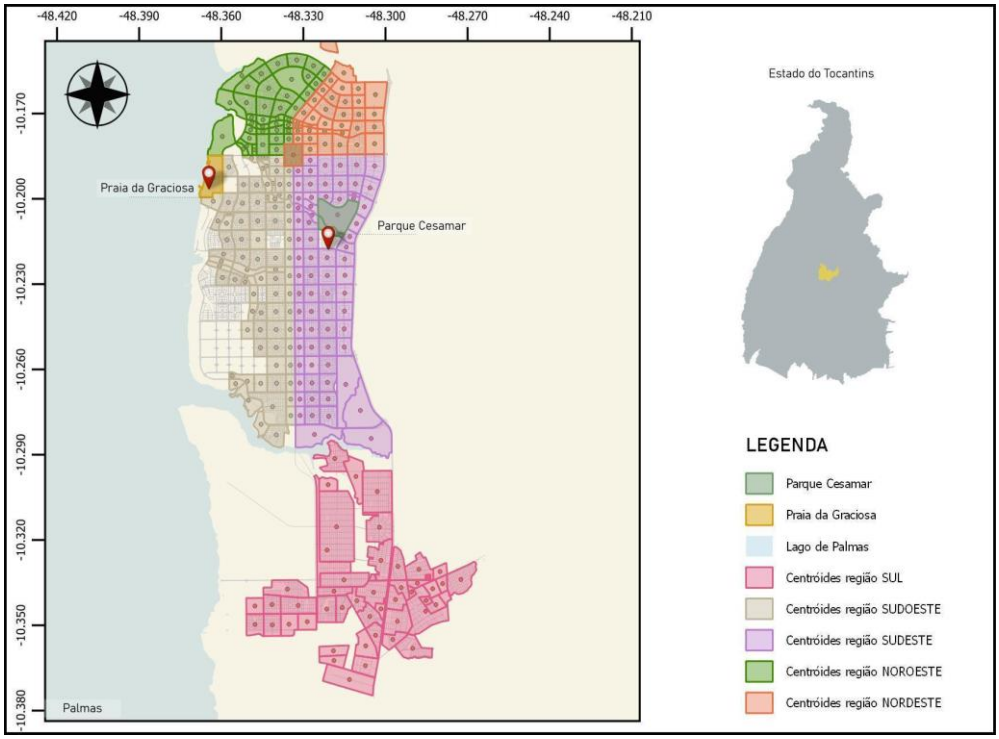
- Distância aos parques: pode ser medida em metros ou quilômetros a partir da localização do imóvel até o parque mais próximo. Essa variável pode capturar o acesso dos moradores a áreas verdes para lazer e recreação;
- Distância às áreas verdes: assim como para parques, pode ser medida em metros ou quilômetros a partir da localização do imóvel até a área verde mais próxima. Essa variável pode capturar a proximidade do imóvel às áreas verdes como bosques, praças e jardins;
- Distância às praias: pode ser medida em metros ou quilômetros a partir da localização do imóvel até a praia mais próxima. Essa variável é especialmente relevante em regiões litorâneas, onde a proximidade do imóvel à praia pode afetar significativamente seu valor;
- Distância a recursos hídricos: pode ser medida em metros ou quilômetros a partir da localização do imóvel até o rio, lago ou oceano mais próximo. Essa variável pode capturar a proximidade do imóvel a recursos hídricos que oferecem oportunidades para atividades de lazer e recreação, além de afetar a qualidade de vida dos moradores.

A proximidade dos imóveis aos ativos ambientais ÁGUA e VERDE foi determinada com base no cálculo da distância euclidiana entre os centróides das quadras que possuem imóveis inseridos nos dados do ITBI e os respectivos pontos de referência ambiental.

Essas distâncias, expressas em quilômetros, foram obtidas por meio do Sistema de Informações Geográficas de Palmas (GEOPALMAS) e incorporadas à modelagem como variáveis contínuas: a distância até a Praia da Graciosa (variável ÁGUA) e a distância até o Parque Cesamar (variável VERDE).

Parte-se do pressuposto, de que existe uma relação inversa entre a distância ao ativo ambiental e o valor de mercado do imóvel, ou seja, quanto mais próximo o imóvel estiver desses elementos naturais, maior tende a ser sua valorização. A localização dos centróides utilizados na construção dessas variáveis pode ser visualizada na Figura 2.

Figura 2: Centróides das quadras



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

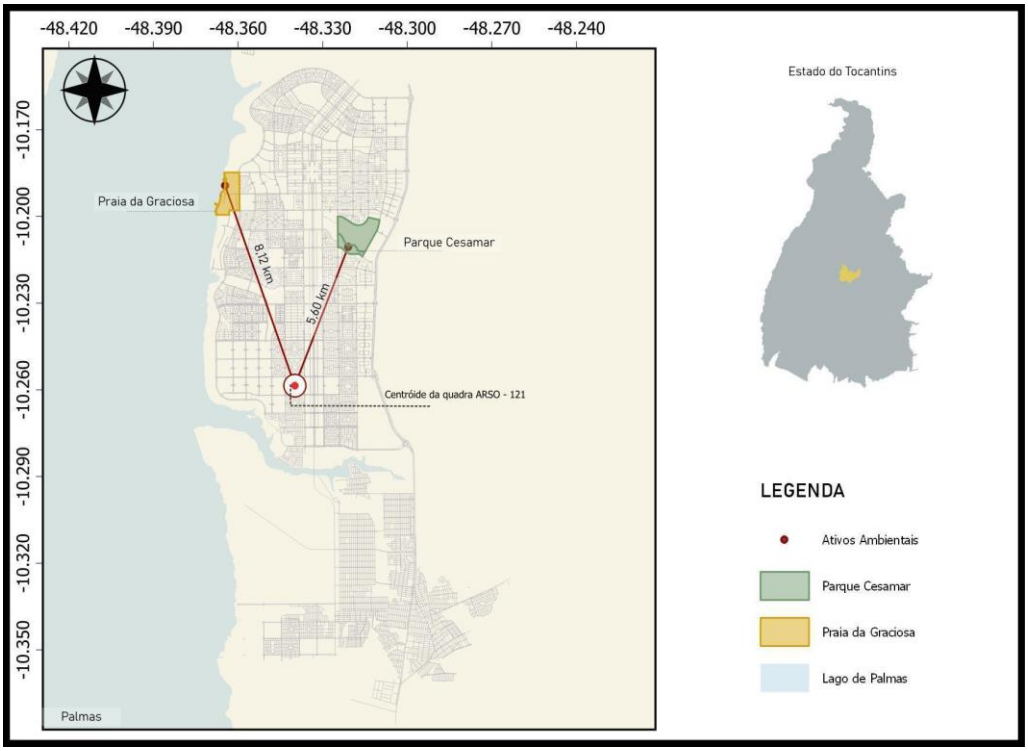
A distância euclidiana é o comprimento da linha reta que une dois pontos em um mapa. Em um plano com o ponto P1 localizado nas coordenadas do mapa (x_1, y_1) e o ponto P2 em (x_2, y_2).

A distância euclidiana entre os dois pontos é calculada por meio da equação (4.25). A Figura 3 mostra a determinação dessa distância.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

(4.25)

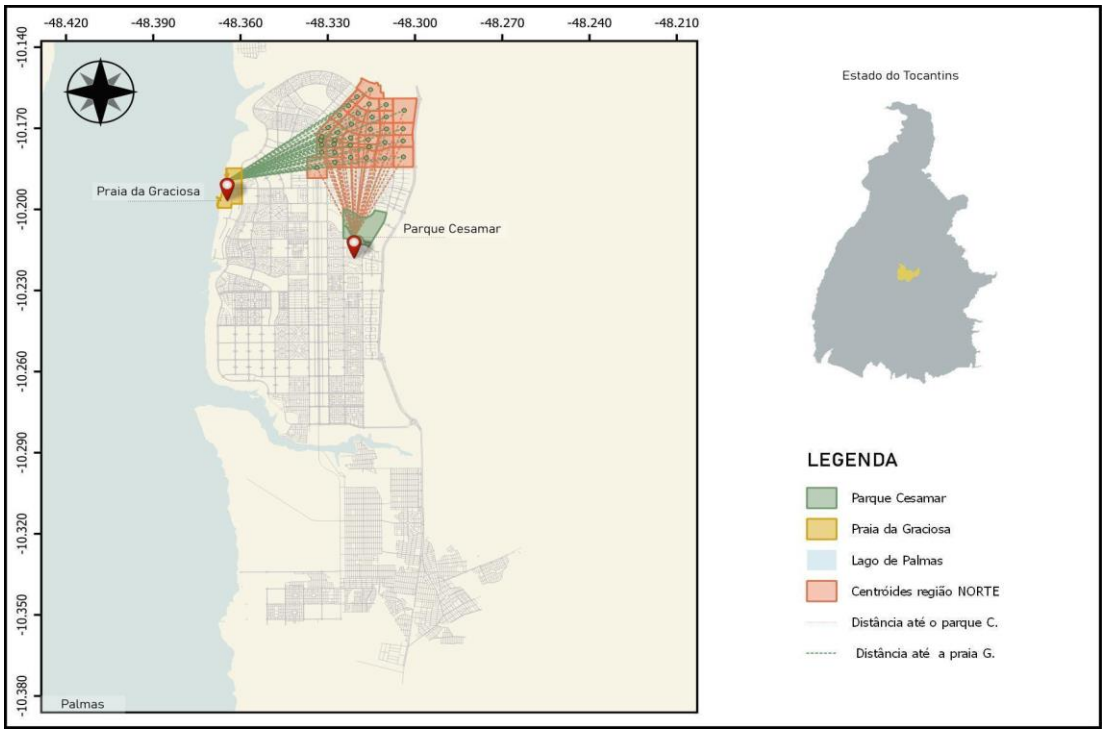
Figura 3: Determinação da Distância Euclidiana



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

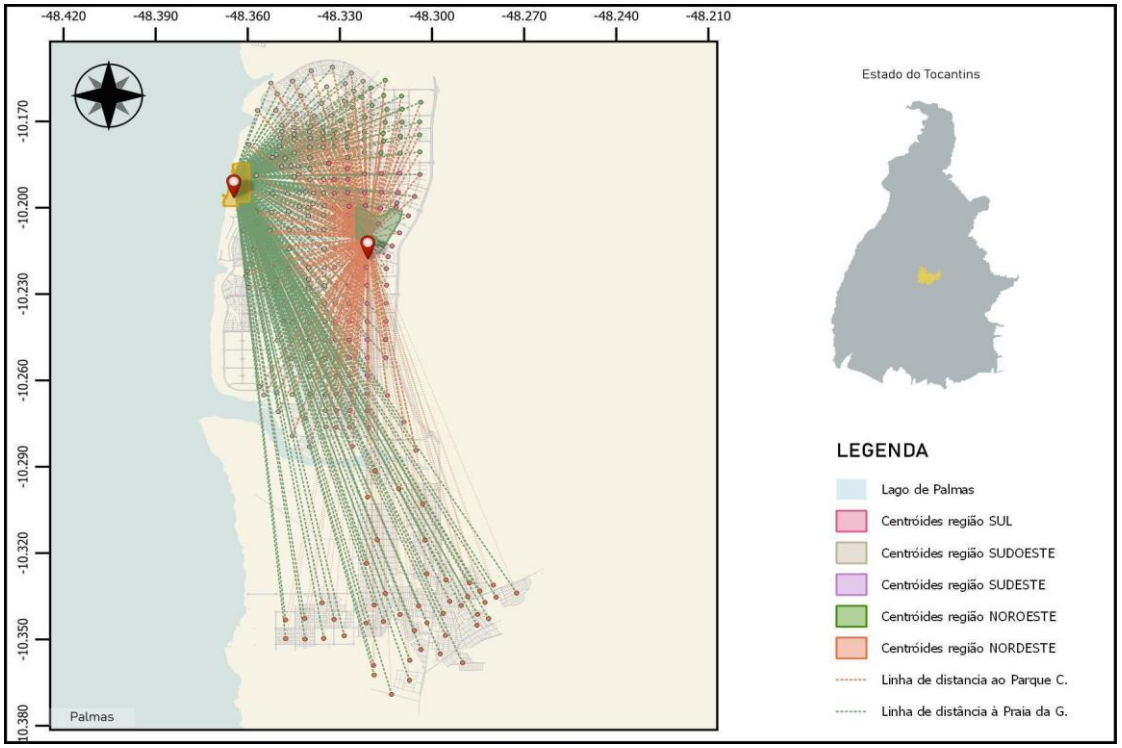
A Figura 4 apresenta a trama de distâncias entre os ativos ambientais e os centróides das quadras da Região Norte, já a Figura 5 indica a trama de distâncias de todas as quadras que compõem o presente estudo.

Figura 4: Trama de Distâncias entre os Centróides das Quadras da Região Norte e os Ativos Ambientais



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Figura 5: Trama de Distâncias entre os Centróides e os Ativos Ambientais



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

O conjunto fiscal contempla duas variáveis comumente utilizadas na literatura de avaliação imobiliária em contexto urbano: o Índice Fiscal da quadra (variável Zona Fiscal), que varia de 1 a 5 conforme a classificação da Planta de Valores Genéricos (PVG), e o próprio valor do metro quadrado constante na PVG (variável PVG), expresso em reais por metro quadrado (R$/m²). A primeira tem sinal esperado negativo, enquanto a segunda, positivo, refletindo, respectivamente, o nível de adensamento e o potencial de valorização da localidade.

No que tange às variáveis *dummies* de proximidade, foram criadas categorias para captar os efeitos de localização relativa dos imóveis em relação aos dois ativos ambientais mencionados. Para o Parque Cesamar, definiram-se as *dummies* "*Perto_Parque"* (distância menor ou igual a 3,5 km) e "*Médio_Parque"* (distância entre 3,5 km e 7 km). Da mesma forma, para a Praia da Graciosa, foram criadas as variáveis "*Perto_Praia"* e "*Médio_Praia"* com a mesma lógica de distanciamento. Essas variáveis, de natureza escalar, foram geradas a partir do GEOPALMAS e visam capturar efeitos não lineares da localização, especialmente relevantes nos modelos de aprendizagem de máquina.

Por fim, incluiu-se uma variável de natureza econômica: a renda familiar mensal dos domicílios, expressa em reais, extraída de bases do IBGE e agregada por setor censitário. A expectativa é que essa variável apresente correlação positiva com os preços dos imóveis, refletindo a capacidade de pagamento das famílias residentes na região.

Esse conjunto de variáveis foi utilizado tanto nos modelos de regressão por MLG e ML quanto nos algoritmos de aprendizagem de máquina, permitindo a comparação entre abordagens tradicionais e preditivas quanto à capacidade de explicar a variação dos preços dos imóveis urbanos em Palmas-TO.

5 RESULTADOS E DISCUSSÃO

A análise dos resultados está organizada em três blocos principais: inicialmente, são apresentados e discutidos dois Modelos Lineares Generalizados (GLM), denominados Modelos A e B, em seguida, dois Modelos Lineares (ML), identificados como Modelos C e D, e, por fim, procede-se à apresentação e discussão dos 14 modelos de Aprendizagem de Máquina (AM) empregados nesta pesquisa, com ênfase na comparação de seus desempenhos e na interpretação dos principais achados.

5.1 Resultados dos Modelos Lineares Generalizados (MLG)

5.1.1 - Modelo A (MLG) - Considerando as Distâncias - Parque e Praia

O Modelo Linear Generalizado (MLG), que incorporou variáveis contínuas de distância, referentes à proximidade dos imóveis em relação à praia e ao parque, apresentou resultados coerentes com aqueles obtidos nos modelos lineares previamente estimados, tanto no que diz respeito à direção dos efeitos quanto à magnitude dos coeficientes. A especificação do modelo considerou como variável dependente o valor da transação imobiliária e, como variáveis explicativas, a área do terreno, a área da benfeitoria, as distâncias em relação aos ativos ambientais (praia e parque), bem como variáveis de natureza fiscal (zona fiscal e PVG) e socioeconômica (renda média per capita). Os resultados completos estão apresentados na Tabela 1.

Tabela 1: Resultados do Modelo A (MLG)

Modelo A (MLG) – Considerando as distâncias Parque-Praia	
	Variável Dependente:
	Modelo A
Área do Terreno	253.260***
	(9.922)
Área da Benfeitoria	207.176
	(9.886)
Distância à Praia	1,017.636
	(764.524)
Distância ao Parque	-2,286.149
	(1,893.833)
Zona Fiscal	-52,992.050***
	(3,221.546)
PVG	179.800***
	(11.860)
Renda Per capita	2.457***
	(0.478)
Constant	158,294.900***
	(13,947.300)
Observações	6,692
Critério de Informação de Akaike	182,555.200
Nota:	*p<0.1; **p<0.05;*** p<0.01

Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Os resultados indicaram que as variáveis físicas (terreno e benfeitoria) mantiveram efeitos positivos sobre o valor de transação, com coeficientes estimados em R$ 253,26 e R$ 207,17 por metro quadrado adicional, respectivamente.

A variável terreno revelou-se estatisticamente significativa ao nível de 1%, indicando forte evidência de associação com os preços dos imóveis analisados. Em contrapartida, a variável benfeitoria, após a aplicação da correção robusta dos erros padrão, deixou de apresentar significância estatística, sugerindo que seu efeito estimado não se mostra robusto diante das possíveis limitações da base de dados utilizada.

No que se refere às variáveis ambientais, observou-se que a distância à praia apresentou coeficiente positivo, indicando um acréscimo estimado de R$ 1.017,64 por quilômetro adicional; entretanto, esse efeito não se revelou estatisticamente significativo. De forma semelhante, a distância ao parque apresentou coeficiente negativo, sugerindo uma redução de R$ 2.286,15 por quilômetro, mas também sem significância estatística. Esses achados

corroboram o que já foi verificado em modelagens anteriores. Embora os sinais estimados estejam em consonância com a literatura, que associa a proximidade de áreas verdes e corpos hídricos à valorização imobiliária, tais relações, no presente estudo, não se mostraram estatisticamente robustas.

As variáveis locacionais e fiscais apresentaram efeitos estatisticamente significativos e substantivos sobre os preços dos imóveis. O índice fiscal evidenciou um coeficiente negativo expressivo, estimando-se uma redução de R$ 52.992,05 no valor dos imóveis localizados em zonas fiscais menos qualificadas. Tal resultado confirma a hipótese de que a inserção territorial em regiões com menor valorização fiscal impacta negativamente o valor de mercado dos bens imóveis.

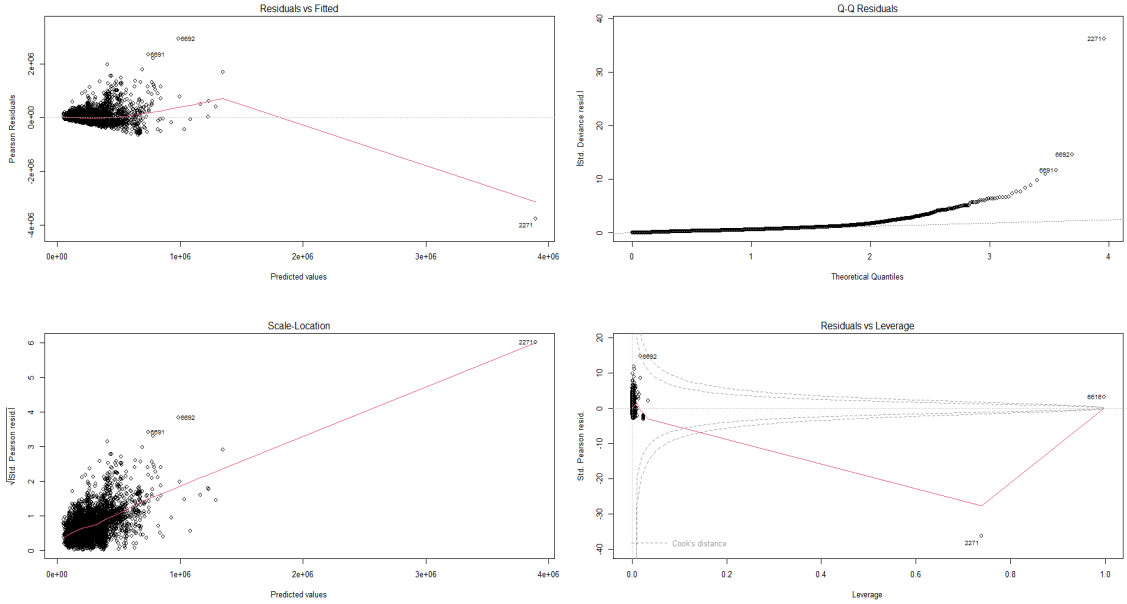
A variável PVG apresentou efeito positivo e estatisticamente significativo, indicando um acréscimo médio de R$ 179,80 no valor do imóvel para cada real adicional no valor de referência por metro quadrado. De igual modo, a variável renda demonstrou-se relevante, com coeficiente positivo de R$ 2,46 por real, evidenciando uma relação direta entre o poder aquisitivo local e a valorização imobiliária. Ambos os resultados confirmam a influência das condições econômicas e dos parâmetros fiscais de referência na formação dos preços dos imóveis urbanos.

A avaliação do modelo evidenciou importantes limitações. O teste de Breusch-Pagan revelou forte evidência de heterocedasticidade ($BP = 3429,8$; $p < 2,2e^{-16}$), resultado corroborado pelo teste de White ($BP = 3349,2$; $p < 2,2e^{-16}$), indicando violação do pressuposto de homocedasticidade. Em função disso, foram aplicadas correções robustas aos erros padrão, que resultaram em alteração da significância de algumas variáveis, especialmente benfeitoria e as variáveis ambientais, que não se mostraram robustamente associadas aos preços dos imóveis.

O diagnóstico gráfico do modelo, apresentado na Figura 6, reforça essas limitações, evidenciando a presença de resíduos com padrão de dispersão heterogênea, conforme observado no gráfico "Residuals vs Fitted", bem como a existência de pontos influentes e potenciais *outliers*, destacados no gráfico "Residuals vs Leverage", com valores elevados de alavancagem.

A distribuição dos resíduos, ilustrada no gráfico Q-Q Plot, afasta-se da normalidade, especialmente nas caudas, indicando assimetria e presença de valores extremos. Ademais, o gráfico "Scale-Location" sugere aumento da variabilidade dos resíduos à medida que os valores ajustados se elevam, corroborando a presença de heterocedasticidade já apontada pelos testes estatísticos.

Figura 6: Gráficos de diagnóstico do Modelo Linear Generalizado



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

O cálculo do *pseudo* R^2 apresentou valor muito baixo (R^2 = 0,0161), reiterando a limitação do modelo linear em capturar a complexidade das relações entre atributos ambientais e preços de imóveis, sobretudo diante da heterogeneidade e não linearidade inerentes ao mercado imobiliário.

Por fim, a análise via *bootstrap* com 5.000 repetições forneceu estimativas mais robustas dos intervalos de confiança. Para as variáveis praia e parque, os intervalos de confiança abrangeram a nulidade (-R$ 4.926,04 até R$ 2.710,37 e -R$ 6.987,09 até R$ 6.629,15, respectivamente), reforçando a ausência de um efeito estatisticamente robusto dessas variáveis sobre o preço de transação dos imóveis. Em contrapartida, as variáveis fiscal, PVG e renda mantiveram intervalos de confiança que não incluem o zero, reiterando sua relevância como determinantes do valor de mercado dos imóveis.

Portanto, os resultados obtidos com o GLM corroboram as limitações dos modelos lineares tradicionais para capturar os efeitos das variáveis ambientais sobre os preços de imóveis, ao mesmo tempo em que destacam a robustez das variáveis físicas, fiscais e socioeconômicas como principais determinantes.

Esses achados reforçam a necessidade de adoção de modelagens mais flexíveis e sofisticadas, como os modelos de aprendizagem de máquina, que podem melhor lidar com não linearidades e interações complexas que caracterizam os mercados imobiliários.

5.1.2 - Modelo B (glm) - Considerando as Dummys - Parque e Praia

O Modelo Linear Generalizado (MLG) com a inclusão de variáveis dummies associadas à localização ambiental dos imóveis (*dummycp*, *dummycm*, *dummypp* e *dummypm*) apresentou resultados coerentes com aqueles verificados nas especificações anteriores, tanto no que se refere à direção dos efeitos quanto à magnitude dos coeficientes estimados. A formulação do modelo adotou como variável dependente o valor da transação imobiliária, sendo consideradas como preditoras a área do terreno, a área da benfeitoria, as variáveis indicadoras de proximidade a ativos ambientais (parque e praia), além das variáveis fiscais (zona fiscal e PVG) e socioeconômicas (renda média per capita). Os resultados completos encontram-se apresentados na Tabela 2.

Tabela 2: Resultados do Modelo B (MLG)

Modelo B (MLG) – Considerando Variáveis *Dummy* para Proximidade

	Variável Dependente:
	Modelo B
Área do Terreno	256.139***
	(9.907)
Área da Benfeitoria	204.725***
	(9.887)
Dummy-Curta Distância à Praia	5,166.690
	(11,505.750))
Dummy-Média Distância à Praia	-14,485.060
	(10,394.500)
Dummy-Curta Distância ao Parque	9,758.836
	(8735.047)
Dummy-Média Distância ao Parque	16,208.090**
	(8,120.013)
Zona Fiscal	-48,470.170***
	(3,465.615)
PVG	180.696***
	(11,844)
Renda per capita	2,520***
	(0.506)
Constant	135,734.800***
	(15,612.140)
Observações	6,692
Critério de Informação de Akaike	(182,546.000)
Nota:	*p<0.1; **p<0.05;*** p<0.01

Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

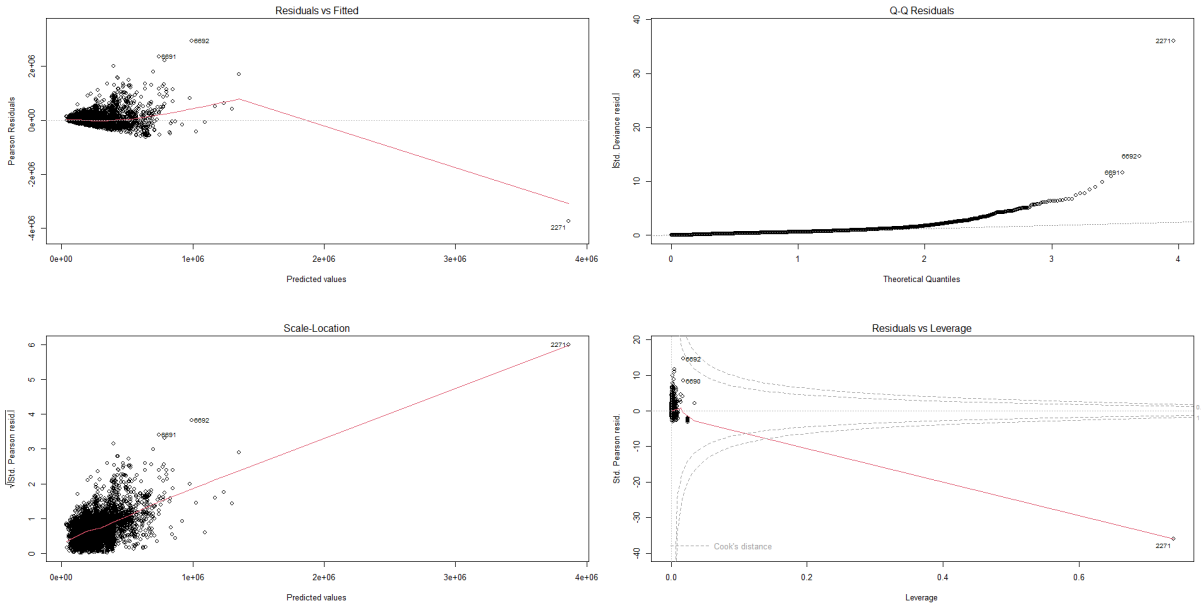
Os resultados evidenciaram que as variáveis físicas terreno e benfeitoria exerceram efeitos positivos sobre o valor de transação dos imóveis, com acréscimos estimados de R$ 256,14 e R$ 204,73 por metro quadrado adicional, respectivamente. A variável terreno apresentou significância estatística ao nível de 1%, indicando forte associação com o valor dos imóveis. Em contrapartida, a variável benfeitoria, após a aplicação da correção robusta dos erros padrão, não se mostrou estatisticamente significativa, comportamento já verificado em modelos anteriores que consideraram variáveis contínuas de distância.

No que se refere às variáveis ambientais modeladas como *dummies*, destaca-se que apenas *dummypm* apresentou significância estatística marginal após a correção robusta (coeficiente de R$ 16.208,09, p = 0,0093), sugerindo um potencial efeito positivo da presença desta característica ambiental específica sobre o valor dos imóveis. As demais variáveis ambientais (*dummycp*, *dummycm* e *dummypp*) não apresentaram significância estatística, e seus intervalos de confiança, obtidos via *bootstrap* com 5.000 repetições, incluíram a nulidade, reforçando a ausência de robustez desses efeitos.

As variáveis fiscais e socioeconômicas mantiveram-se como determinantes relevantes para a explicação do valor de mercado dos imóveis. A variável fiscal apresentou coeficiente negativo expressivo, estimando-se uma redução de R$ 48.470,17 nos imóveis situados em zonas com menor qualificação fiscal, com significância estatística ao nível de 1%. A variável PVG também evidenciou efeito positivo significativo ao nível de 1%, indicando um acréscimo médio de R$ 180,70 para cada real adicional no valor de referência por metro quadrado. Por sua vez, a variável renda média per capita demonstrou significância estatística ao nível de 1%, com coeficiente positivo de R$ 2,52 por real, corroborando a relação direta entre renda local e valorização imobiliária.

A análise diagnóstica do modelo, ilustrada na Figura 7, reforça as limitações já identificadas nos modelos anteriores.

Figura 7: Gráficos de diagnóstico do Modelo Linear Generalizado



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa

O gráfico "*Residuals vs Fitted*" evidencia um padrão de heterocedasticidade, com aumento da dispersão dos resíduos em função dos valores ajustados. O gráfico Q-Q Plot demonstra desvios significativos da normalidade, especialmente nas caudas, indicando a presença de assimetria e valores extremos. O gráfico "*Scale-Location*" confirma a variabilidade não constante dos resíduos, enquanto o gráfico "*Residuals vs Leverage*" aponta a existência de pontos influentes, com destaque para observações com alta alavancagem e potencial impacto na estabilidade do modelo.

O teste de heterocedasticidade de *Breusch-Pagan* indicou forte evidência de violação do pressuposto de homocedasticidade (BP = 3395,4; $p < 2{,}2e^{-16}$), resultado corroborado pelo teste de White (BP = 3325; $p < 2{,}2e^{-16}$), evidenciando a necessidade de correção dos erros padrão, realizada por meio do estimador heterocedasticity-consistent (HC).

O pseudo R^2 calculado para este modelo resultou em um valor muito baixo (R^2 = 0,0164), consistente com os valores obtidos nas modelagens anteriores, e que reforça as limitações do modelo linear tradicional em capturar a complexidade das relações entre as variáveis ambientais e o preço de transação dos imóveis.

Adicionalmente, a análise via *bootstrap* forneceu estimativas mais robustas dos intervalos de confiança. Para as variáveis ambientais, os intervalos incluíram a nulidade, exceto para *dummypm*, cujo intervalo foi (R$ 3.316,03; R$ 28.285,47), sugerindo um possível, embora ainda marginal, efeito positivo. As variáveis fiscal, PVG e renda mantiveram intervalos de

confiança que não incluem o zero, reafirmando sua importância na explicação dos valores transacionados.

Comparativamente aos modelos ajustados com variáveis contínuas de distância, observa-se uma coerência nos principais achados: robustez das variáveis físicas e de localização e ausência de efeito estatisticamente consistente das variáveis ambientais. A única exceção parcial foi a variável *dummypm*, que apresentou um efeito estatisticamente marginal, não identificado nas modelagens com variáveis contínuas.

Portanto, este modelo reitera as conclusões já apontadas: as variáveis físicas e fiscais são os principais determinantes dos valores de transação imobiliária em Palmas-TO, enquanto as variáveis ambientais, quando modeladas linearmente, não apresentam influência estatisticamente robusta. Este cenário reforça a necessidade de adoção de modelos mais sofisticados e flexíveis, como os de aprendizagem de máquina, capazes de lidar com padrões não lineares e interações complexas, conforme já defendido na literatura especializada (Paixão, 2025).

5.2 – Resultados dos Modelos Lineares (ML)

5.2.1 Modelo C (lm) - Considerando as Distâncias - Parque e Praia

O terceiro modelo estimado, baseado em regressão linear múltipla, adotou como variável dependente o valor da transação imobiliária. As variáveis independentes incluíram a área do terreno, a área da benfeitoria, as distâncias euclidianas em relação à praia e ao parque, a zona fiscal, a Planta de Valores Genéricos (PVG) e a renda per capita das respectivas quadras censitárias. A formulação buscou avaliar o efeito conjunto dessas características sobre os preços dos imóveis transacionados.

O modelo apresentou um coeficiente de determinação (R^2) de 0,312 e um R^2 ajustado de 0,311, indicando que aproximadamente 31% da variabilidade do preço de transação dos imóveis é explicada pelas variáveis independentes consideradas. Embora esse valor possa ser classificado como moderado, trata-se de um resultado típico em estudos empíricos na área de avaliação de imóveis urbanos, especialmente quando se trabalha com variáveis físicas e ambientais que, por sua natureza, não capturam toda a complexidade subjetiva envolvida na formação dos preços.

Conforme destacam Jim e Chen (2006), análises hedônicas de imóveis urbanos frequentemente registram valores de R^2 entre 0,20 e 0,50, o que se deve, em grande parte, à

heterogeneidade intrínseca aos mercados imobiliários. O teste F revelou-se altamente significativo (F = 432,5; p-valor < $2e^{-16}$), demonstrando que, de forma global, o modelo possui validade estatística, ou seja, ao menos uma das variáveis independentes contribui efetivamente para explicar o preço de transação dos imóveis analisados.

O erro padrão residual estimado foi de R$ 202.829,70, o que, considerando o valor médio das transações, sugere uma dispersão relevante, porém característica em estudos com dados de mercado imobiliário, cujo comportamento é naturalmente ruidoso em função de fatores intangíveis não contemplados na modelagem, tais como percepção de segurança, qualidade da vizinhança e especificidades locacionais (DANTAS, 2005). A análise dos coeficientes estimados pelo modelo de regressão linear múltipla evidencia que o intercepto foi estimado em R$ 158.294,90, com erro padrão de R$ 13.947,30, sendo altamente significativo (p < $2e^{-16}$) (Tabela 3).

Tabela 3: Modelo C (lm) – Considerando as Distâncias – Parque e Praia

Modelo C (lm) – Considerando as distâncias – Parque e Praia

	Variável Dependente:
	Modelo C
Área do Terreno	253.260***
	(9.922)
Área da Benfeitoria	207.176
	(9.886)
Distância à Praia	1,017.636
	(764.524)
Distância ao Parque	-2,286.149
	(1,893.833)
Zona Fiscal	-52,992.050***
	(3,221.546)
PVG	179.800***
	(11.860)
Renda Per capita	2.457***
	(0.478)
Constant	158,294.900***
	(13,947.300)
Observações	6,692
R^2	0.312
R^2 Ajustado	0.311
Erro Padrão dos Resíduos	202,829.700 (df=6684)
Nota:	*p<0.1; **p<0.05;*** p<0.01

Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

A variável terreno apresentou um coeficiente estimado de R$ 253,26 por metro quadrado adicional, com significância estatística ao nível de 1%, evidenciando um impacto positivo e robusto sobre o valor de transação dos imóveis. De modo semelhante, a variável benfeitoria resultou em um acréscimo médio de R$ 207,18 para cada metro quadrado de área construída, também significativo ao nível de 1%, reforçando sua relevância na composição do valor dos imóveis analisados.

No que se refere às variáveis ambientais, a distância à praia apresentou coeficiente positivo de R$ 1.017,64 por quilômetro adicional; contudo, não se mostrou estatisticamente significativa, sugerindo que a proximidade ou o afastamento da praia não exerceu efeito robusto sobre o valor dos imóveis nesta amostra. De forma semelhante, a distância ao parque resultou em um coeficiente negativo de R$ 2.286,15 por quilômetro, também sem significância estatística, o que indica que, embora o sinal estimado esteja em conformidade com a literatura, segundo a qual a proximidade a áreas verdes tende a valorizar os imóveis, tal relação não se confirmou de forma estatisticamente relevante no presente estudo.

Em contrapartida, as variáveis locacionais e fiscais evidenciaram forte influência sobre o valor de transação dos imóveis. A variável zona fiscal apresentou coeficiente negativo expressivo, estimando-se uma redução de R$ 52.992,05 no valor dos imóveis a cada unidade de incremento, com significância estatística ao nível de 1%, o que reforça a evidência de que imóveis localizados em zonas fiscais de menor qualificação tendem a apresentar menor valor de mercado. Por sua vez, o valor da Planta de Valores Genéricos (PVG) destacou-se como um importante preditor, com coeficiente positivo de R$ 179,80 para cada real adicional no valor de referência por metro quadrado, sendo também significativo ao nível de 1%, corroborando sua relevância na precificação dos imóveis urbanos.

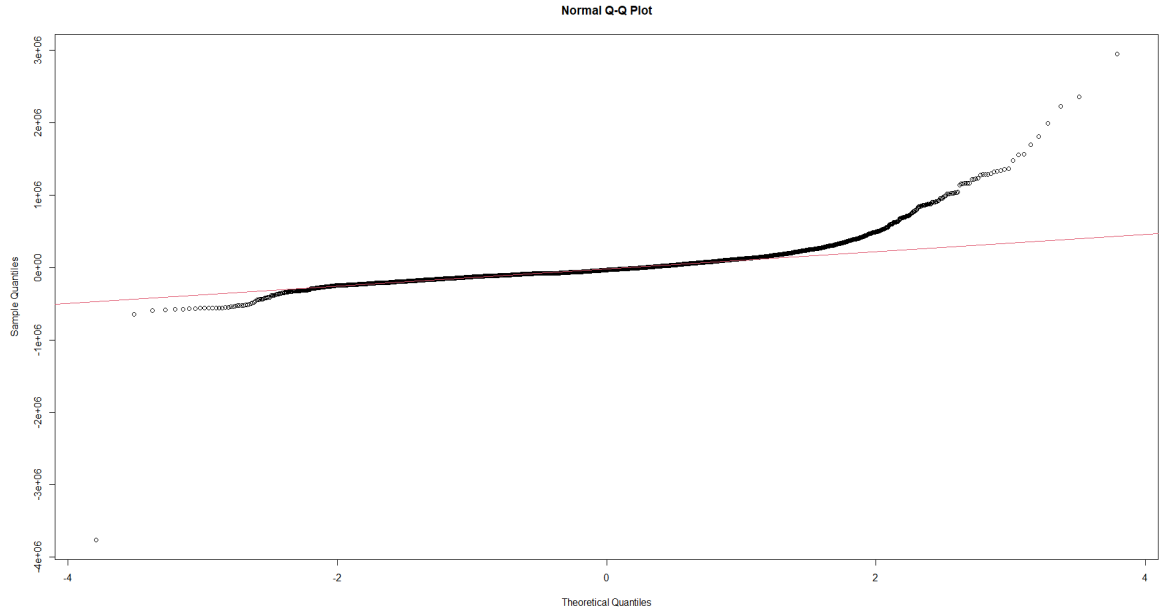
Por fim, a variável renda média per capita das quadras demonstrou um coeficiente positivo de R$ 2,46 por real, com elevada significância estatística, ao nível de 1%, o que reforça a relação entre o poder aquisitivo local e a valorização imobiliária, em consonância com o que é amplamente discutido na literatura sobre mercados habitacionais (BRUECKNER *et al.*, 1999).

A avaliação do modelo ajustado foi realizada de forma sistemática, envolvendo o teste global de ajuste, a análise da função de variância, a verificação da normalidade dos resíduos, o diagnóstico da função de ligação e a verificação da autocorrelação dos resíduos. A normalidade dos resíduos foi avaliada inicialmente pelo teste de Lilliefors (Kolmogorov-Smirnov), cujo resultado indicou a rejeição da hipótese nula de normalidade ($D = 0{,}12865$; $p < 2^{2e\text{-}16}$). A

tentativa de aplicação do teste de Shapiro-Wilk não foi possível em razão do tamanho da amostra.

O gráfico Q-Q Plot dos resíduos (Figura 8) reforça esse resultado, evidenciando desvios sistemáticos dos resíduos em relação à linha de referência, especialmente nas caudas da distribuição. Isso indica a presença de assimetria e valores extremos (*outliers*), confirmando que a distribuição dos resíduos não segue a normalidade, o que exige cautela na interpretação das inferências baseadas em testes paramétricos.

Figura 8: Gráfico Q-Q Plot dos resíduos do Modelo C (lm)



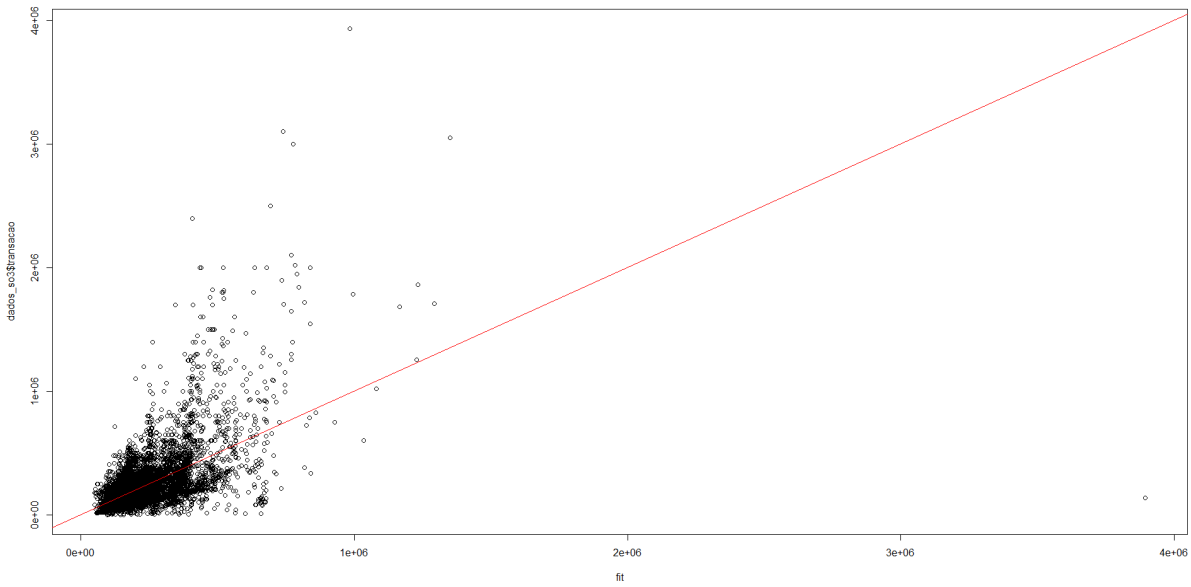
Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

No tocante à homocedasticidade, o teste de Breusch-Pagan revelou forte evidência de heterocedasticidade (BP = 3429,8; $p < 2{,}2e^{-16}$), resultado corroborado pelo teste de White (BP = 3349,2; $p < 2{,}2e^{-16}$).

A inspeção da função de ligação foi realizada por meio do gráfico de valores ajustados *versus* valores observados (Figura 9). Observa-se uma tendência linear compatível com o modelo ajustado, mas com considerável dispersão, principalmente para valores de transação mais elevados.

Este resultado confirma a limitação do modelo em capturar, com precisão, a totalidade das variações presentes no mercado, fenômeno que, conforme destacado por Brueckner *et al.* (1999), é influenciado por múltiplos fatores socioeconômicos e ambientais, nem sempre plenamente observados ou mensuráveis.

Figura 9: Gráfico de valores ajustados versus valores observados



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Em função das evidências de heterocedasticidade, foram aplicadas correções robustas aos erros padrão utilizando o estimador matrizes de variância-covariância heterocedasticidade-consistentes (vcovHC). Após a correção, observou-se alteração relevante na significância de algumas variáveis: enquanto benfeitoria perdeu significância (p = 0,73063), as variáveis fiscal, terreno, PVG e renda mantiveram-se significativas, reforçando sua robustez como determinantes do valor de transação dos imóveis.

A análise do pseudo R^2, calculado a partir da correlação entre os valores ajustados e a inversa das transações, indicou um baixo poder explicativo adicional ($R^2 = 0,01608$), sugerindo que transformações mais sofisticadas ou a utilização de modelos não lineares poderiam capturar melhor as relações subjacentes entre as variáveis.

Por fim, a análise realizada via *bootstrap*, com 5.000 repetições, proporcionou estimativas mais robustas dos intervalos de confiança. As variáveis terreno, PVG e renda mantiveram intervalos de confiança que não incluem zero, reforçando a sua significância estatística. Por outro lado, os intervalos de confiança para as variáveis praia e parque abrangeram a nulidade, reforçando a conclusão de que tais variáveis ambientais, nesta modelagem linear, não apresentaram influência robusta sobre o valor de transação dos imóveis.

Os resultados indicam que os principais determinantes do valor de transação dos imóveis são as características físicas (área do terreno e benfeitoria), as condições locacionais (fiscal e PVG) e os fatores socioeconômicos (renda média da quadra). De outro modo, a ausência de significância das variáveis ambientais sugere a necessidade de investigações

complementares, possivelmente com modelos não lineares ou de aprendizagem de máquina, que podem capturar relações complexas não detectadas pela modelagem linear tradicional.

Além disso, a violação das hipóteses de homocedasticidade e normalidade dos resíduos reforça a necessidade de correções metodológicas ou a adoção de modelos alternativos que sejam mais robustos às limitações dos dados observados.

5.2.2 - Modelo D (*lm*) - Considerando as *Dummys* - Parque e Praia

Neste quarto modelo, estimado por meio de regressão linear múltipla, a variável dependente permanece a transação imobiliária e, como variáveis independentes, a área do terreno, a área da benfeitoria, as *dummies* associadas às características ambientais (*dummycp*, *dummycm*, *dummypp* e *dummypm*), bem como as variáveis fiscais (zona fiscal), a Planta de Valores Genéricos (PVG) e a renda per capita das quadras.

Os resultados indicaram um coeficiente de determinação (R^2) de 0,3131 e um R^2 ajustado de 0,3122, o que significa que aproximadamente 31% da variação observada nos preços de transação dos imóveis foi explicada pelas variáveis incorporadas ao modelo. Apesar de moderado, esse nível de explicação é compatível com o encontrado na literatura especializada em avaliação de imóveis urbanos, notadamente quando se trabalha com atributos físicos e ambientais que, por definição, não abarcam integralmente a complexidade subjetiva inerente ao processo de formação de preços.

Conforme destacam Jim e Chen (2006), análises hedônicas de imóveis urbanos usualmente apresentam valores de R^2 situados entre 0,20 e 0,50, resultado que reflete a diversidade e a heterogeneidade presentes nos mercados imobiliários.

O teste F apresentou-se altamente significativo ($F = 338,4$; p-valor $< 2e^{-16}$), evidenciando a validade estatística do modelo como um todo, ou seja, indicando que, pelo menos, uma das variáveis independentes contribui efetivamente para a explicação da variável dependente. O erro padrão residual foi estimado em R$ 202.661,00, um valor que, considerando-se a magnitude das transações imobiliárias analisadas, revela uma dispersão importante, mas usual em estudos sobre preços de mercado, os quais são frequentemente influenciados por aspectos intangíveis não modelados, como percepção de segurança, qualidade ambiental local e especificidades locacionais (DANTAS, 2005).

A análise dos coeficientes estimados revelou que o intercepto foi de R$ 135.734,80, com erro padrão de R$ 15.612,14, sendo estatisticamente significativo ao nível de 1%, o que indica a presença de um valor base elevado para os imóveis considerados. A variável terreno apresentou coeficiente positivo de R$ 256,14 por metro quadrado adicional, com significância

estatística ao nível de 1%, evidenciando um impacto robusto sobre o valor de transação dos imóveis. De forma semelhante, a variável benfeitoria resultou em um acréscimo médio de R$ 204,73 por metro quadrado de área construída, também significativa ao nível de 1%, reforçando sua relevância na composição do preço final dos imóveis urbanos analisados. Os resultados encontram-se sistematizados na Tabela 4.

Tabela 4: Modelo D (lm) - Considerando as Dummys - Parque e Praia

Modelo D (lm) – Considerando as *Dummy* – Parque e Praia

	Variável Dependente:
	Modelo D
Área do Terreno	256.139***
	(9.907)
Área da Benfeitoria	204.725***
	(9.887)
Dummy-Curta Distância à Praia	5,166.690
	(11,505.750))
Dummy-Média Distância à Praia	-14,485.060
	(10,394.500)
Dummy-Curta Distância ao Parque	9,758.836
	(8735.047)
Dummy-Média Distância ao Parque	16,208.090**
	(8,120.013)
Zona Fiscal	-48,470.170***
	(3,465.615)
PVG	180.696***
	(11,844)
Renda per capita	2,520***
	(0.506)
Constant	135,734.800***
	(15,612.140)
Observações	6,692
R^2	0.313
R^2 Ajustado	0.312
Erro Padrão dos Resíduos	212,661.000 (df=6682)
Nota:	*p<0.1; **p<0.05;*** p<0.01

Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Com relação às variáveis ambientais, observa-se que, dentre as quatro *dummies* analisadas (*dummycp, dummycm, dummypp* e *dummypm*), apenas a *dummypm* que representa a média distância ao Parque Cesamar apresentou significância estatística ao nível de 5%, com coeficiente positivo de R$ 16.208,09, sugerindo que a localização intermediária do imóvel em

relação à amenidade ambiental pode resultar em um acréscimo relevante no seu valor de mercado. Esse resultado indica uma possível valorização associada à presença do ativo ambiental considerado, ainda que com evidência estatística moderada. As demais variáveis ambientais não demonstraram significância estatística, evidenciando a limitação dos modelos lineares tradicionais em capturar adequadamente os efeitos complexos e, por vezes, não lineares que atributos ambientais exercem sobre os preços dos imóveis.

Este resultado corrobora a constatação de que modelos lineares, embora amplamente empregados na avaliação imobiliária, possuem limitações importantes para detectar interações mais sutis e efeitos não lineares associados às variáveis ambientais.

Nesse contexto, destaca-se a relevância do emprego de modelos de aprendizagem de máquina, que se mostram metodologicamente mais adequados para lidar com padrões complexos e identificar interações ocultas entre variáveis, ampliando a capacidade de compreensão sobre a influência dos atributos ambientais na formação dos preços dos imóveis.

As variáveis relacionadas à localização e aos aspectos fiscais demonstraram impacto expressivo sobre o valor de transação dos imóveis. A variável fiscal (zona fiscal) apresentou coeficiente negativo de R$ 48.470,17 por unidade de incremento, com significância estatística ao nível de 1%, indicando que imóveis situados em quadras com menor índice fiscal tendem a ser mais valorizados, o que reforça a preferência do mercado por áreas dotadas de melhor infraestrutura e qualidade urbanística. Já a variável PVG (Planta de Valores Genéricos) destacou-se como um importante preditor, com coeficiente positivo de R$ 180,70 para cada real adicional no valor de referência por metro quadrado, sendo também significativa ao nível de 1%, o que evidencia a existência de uma relação direta entre a valorização oficial dos imóveis e os preços efetivamente praticados no mercado.

Por fim, a renda média per capita das quadras apresentou coeficiente positivo de R$ 2,52 por real, com significância estatística robusta, o que confirma a relação direta entre poder aquisitivo local e valorização imobiliária, em consonância com o que é amplamente descrito na literatura especializada sobre mercados habitacionais (BRUECKNER *et al*., 1999).

O modelo ajustado por regressão linear múltipla, com variáveis independentes que incorporam atributos físicos, ambientais e socioeconômicos, apresentou resultados que reforçam as tendências já observadas nas modelagens anteriores. O modelo foi especificado com a variável dependente transação e as variáveis explicativas: área do terreno, área da benfeitoria, variáveis *dummies* associadas à proximidade ambiental (*dummycp, dummycm, dummypp, dummypm*), além das variáveis fiscais e socioeconômicas (fiscal, PVG e renda).

Os resultados indicam que as variáveis físicas (terreno e benfeitoria) possuem efeito positivo e altamente significativo sobre o valor de transação dos imóveis, confirmando a importância dessas características na precificação imobiliária em Palmas-TO, tal como discutido por Dantas (2005).

A variável fiscal apresentou coeficiente negativo e altamente significativo, o que corrobora a influência das zonas fiscais na formação dos preços, refletindo que imóveis localizados em zonas com maior carga tributária tendem a apresentar valores de transação reduzidos.

As variáveis ambientais representadas pelas *dummies* indicativas da presença ou proximidade aos ativos naturais mostraram um comportamento mais discreto. Apenas a variável *dummypm* apresentou significância estatística marginal, sugerindo um efeito positivo, embora moderado, da proximidade em relação a esse componente ambiental específico. As demais variáveis ambientais — *dummycp, dummycm e dummypp* — não apresentaram significância estatística, indicando que, sob a modelagem linear utilizada, a influência direta dessas amenidades ambientais não foi robustamente detectada.

O modelo apresentou um coeficiente de determinação (R^2) de aproximadamente 31%, indicando que cerca de um terço da variabilidade dos preços de transação dos imóveis é explicada pelas variáveis incluídas no modelo. Embora este valor seja moderado, é considerado adequado para estudos empíricos em avaliação imobiliária que trabalham com dados heterogêneos e atributos multifacetados.

Entretanto, a análise residual evidenciou importantes limitações. O teste de normalidade de Lilliefors rejeitou a hipótese nula de normalidade dos resíduos ($D = 0{,}1275$; $p < 2{,}2e^{-16}$), e o teste de Breusch-Pagan apontou forte evidência de heterocedasticidade ($BP = 3395{,}4$; $p < 2{,}2e^{-16}$), corroborado pelo teste de White ($BP = 3325$; $p < 2{,}2e^{-16}$). Tais resultados indicam que a variância dos resíduos não é constante, violando um dos pressupostos fundamentais do modelo linear clássico e comprometendo a validade das inferências baseadas nos erros padrão tradicionais.

Em razão dessas violações, foram aplicadas correções robustas aos erros padrão por meio do estimador heterocedasticity-consistent (HC). Após a correção, algumas alterações na significância das variáveis foram observadas. Notadamente, a variável benfeitoria perdeu significância estatística, o que sugere que seu impacto positivo sobre o valor de transação pode estar associado a efeitos espúrios relacionados à heterocedasticidade.

O cálculo do pseudo R^2, obtido pela correlação entre os valores ajustados e a inversa das transações, resultou em um valor extremamente baixo ($R^2 = 0{,}01645$), reforçando a

limitação do modelo linear em captar toda a complexidade das relações entre as variáveis analisadas. A dispersão gráfica entre valores ajustados e observados reforçou essa constatação, evidenciando que, especialmente para os valores mais elevados de transação, o modelo falha em produzir estimativas suficientemente precisas.

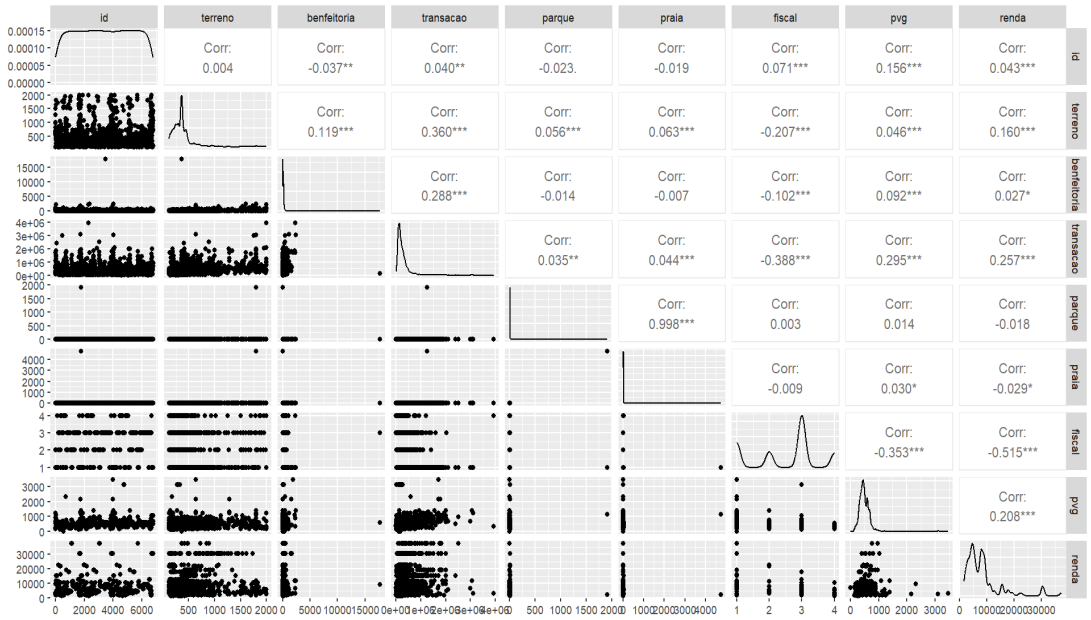
Adicionalmente, a análise de intervalos de confiança via *bootstrap* com 5000 repetições forneceu estimativas mais robustas, demonstrando que, para as variáveis ambientais, os intervalos frequentemente incluem a nulidade, reforçando a ausência de efeitos estatisticamente robustos. Por outro lado, variáveis como fiscal, PVG e renda mantiveram intervalos de confiança que não incluem o zero, reiterando sua importância na determinação do valor de mercado dos imóveis.

Portanto, os resultados obtidos apontam para a robustez das variáveis físicas e fiscais como principais determinantes dos valores de transação, enquanto a influência das variáveis ambientais, modeladas por *dummies*, não se apresentou estatisticamente consistente. Esse conjunto de evidências sugere a necessidade de explorar modelagens mais flexíveis, capazes de lidar com não linearidades e heterogeneidades presentes no mercado imobiliário, como os modelos de aprendizagem de máquina, conforme destacado na literatura contemporânea.

5.3 – Resultados dos Modelos de Aprendizagem de Máquina (AM)

Com base na análise da matriz de correlação apresentada na Figura 10, é possível extrair considerações relevantes sobre as relações entre as variáveis explicativas incluídas no estudo e o valor de transação dos imóveis urbanos no município de Palmas-TO. A matriz contempla, além dos coeficientes de correlação de Pearson, os gráficos de dispersão bivariada e as distribuições univariadas, permitindo uma compreensão mais ampla sobre a estrutura dos dados.

Figura 10: Matriz de Correlação das variáveis utilizadas no modelo



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Inicialmente, observa-se que as variáveis físicas do imóvel (área do terreno e área da benfeitoria) apresentam correlação positiva e estatisticamente significativa com o valor de transação, com coeficientes de 0,360 e 0,288, respectivamente. Esses resultados corroboram os fundamentos da Engenharia de Avaliações e da literatura empírica aplicável, como apontado por González (2003) e Hoffman e Vieira (1977), que destacam o papel central dessas variáveis na formação do valor de mercado dos imóveis.

As variáveis relacionadas à infraestrutura urbana e aspectos fiscais também mostraram correlações relevantes. A variável "zona fiscal", que representa a classificação fiscal das quadras conforme a Planta de Valores Genéricos (PVG), apresentou uma correlação negativa acentuada com o valor da transação (r = -0,388), indicando que imóveis situados em regiões com pior classificação fiscal tendem a apresentar menor valor de mercado. Já o valor do metro quadrado da PVG, por sua vez, revelou uma correlação positiva (r = 0,295), reforçando sua aderência ao comportamento real do mercado imobiliário. Além disso, a variável de renda média familiar também se correlacionou positivamente com o valor dos imóveis (r = 0,257), o que é consistente com os achados de Brueckner *et al*. (1999) e Paixão (2015), os quais apontam a renda como proxy da qualidade urbana e da atratividade locacional.

No que tange às variáveis ambientais, representadas pelas distâncias ao Parque Cesamar e à Praia da Graciosa, as correlações com o valor da transação foram estatisticamente significativas, porém de baixa magnitude (r = 0,056 e r = 0,063, respectivamente). Embora os

sinais positivos não confirmem, à primeira vista, a hipótese de que a proximidade a tais ativos ambientais resulta em valorização imobiliária, esse comportamento pode ser interpretado como indício de uma relação não linear ou com efeitos condicionados por outras variáveis, como zona fiscal e renda.

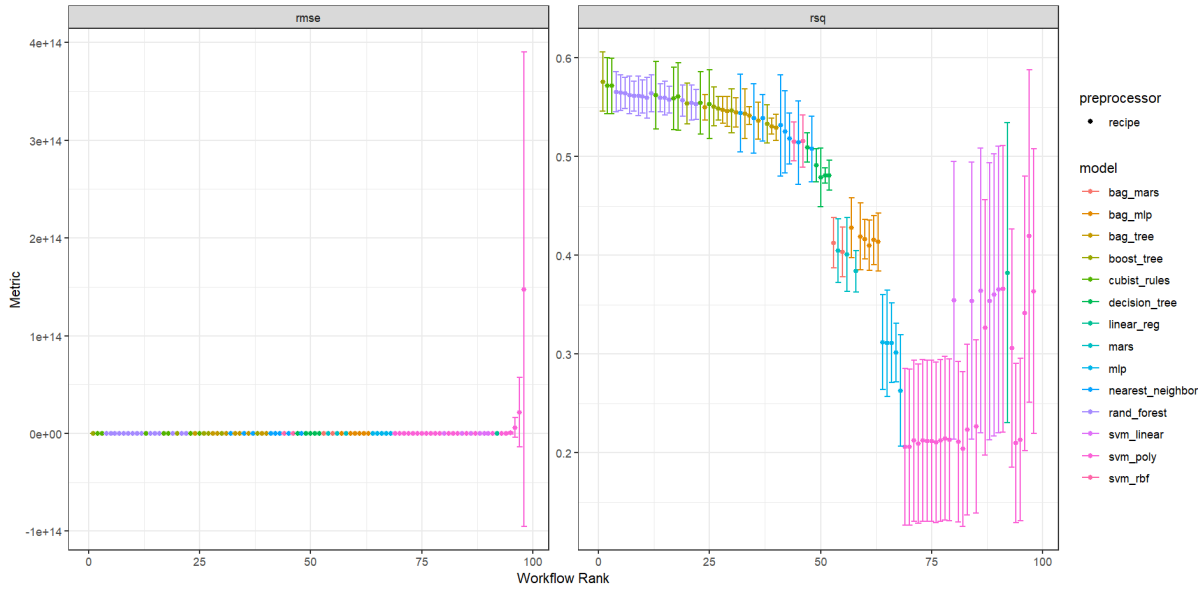
Nessa perspectiva, o uso de algoritmos de aprendizagem de máquina justifica-se amplamente, dado seu potencial para identificar padrões complexos e não lineares, como destacado por Hansen (2021) e James *et al*. (2023).

Outro ponto que merece destaque é a elevada correlação entre as variáveis "parque" e "praia" (r = 0,998), indicando colinearidade quase perfeita. Essa alta dependência linear sugere que essas variáveis medem praticamente a mesma dimensão geográfica, o que exige cautela em modelos lineares, como o Modelo Linear Generalizado (MLG), sob pena de multicolinearidade e perda de eficiência dos estimadores. A recomendação, conforme James *et al*., 2023; Morettin e Singer, 2022, é aplicar procedimentos de seleção de variáveis ou sintetizá-las em um único indicador de acessibilidade ambiental para evitar redundâncias.

Em síntese, os resultados da matriz de correlação confirmam que as variáveis físicas, fiscais e socioeconômicas são relevantes para explicar o valor dos imóveis urbanos em Palmas. As variáveis ambientais, embora apresentem correlações diretas fracas, revelam potencial explicativo adicional a ser captado por modelos não paramétricos e de aprendizagem de máquina. Esses achados sustentam a abordagem metodológica adotada nesta tese, que combina modelagem estatística tradicional com técnicas avançadas de previsão, visando a uma avaliação mais abrangente e acurada da influência de atributos ambientais na precificação dos imóveis urbanos.

As Figuras 11, 12, 13 e 14 apresentam os resultados obtidos na fase de teste dos modelos de aprendizagem de máquina, a partir da técnica de validação cruzada com reamostragem. Os gráficos expressam, respectivamente, as métricas de desempenho (RMSE e R^2) ao longo dos modelos testados, com seus respectivos intervalos de confiança, permitindo avaliar não apenas a média das estimativas, mas também sua variabilidade.

Figura 11: Desempenho Geral dos Modelos

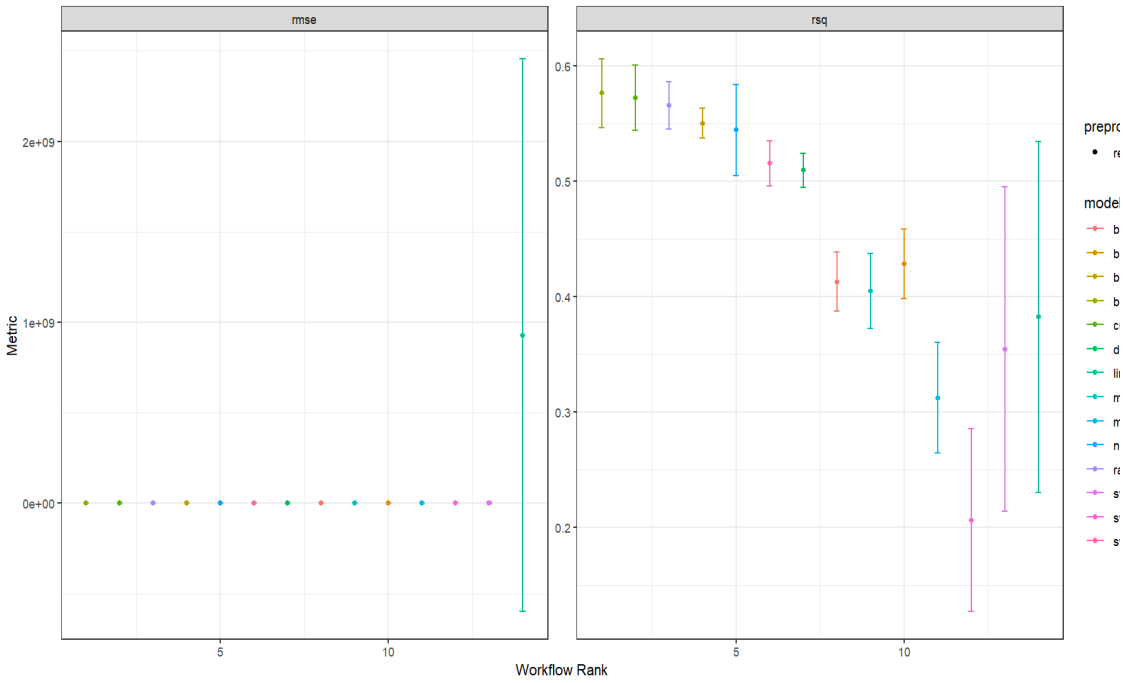


Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

A Figura 11 apresenta a dispersão dos resultados de todos os modelos avaliados ao longo de diferentes configurações (workflows), ordenados conforme o desempenho de cada métrica. À esquerda, observa-se que os modelos *Bagged Trees*, *Random Forest* e *Boosted Trees* concentram os menores valores de RMSE, indicando menor erro médio de predição dos preços dos imóveis. À direita, no gráfico de R^2, esses mesmos modelos mantêm os melhores desempenhos.

Esses resultados reforçam a robustez dos modelos de ensamble baseados em árvores de decisão, especialmente em cenários onde há não linearidade nas relações entre os atributos (como proximidade a amenidades ambientais, renda, padrão construtivo etc.), conforme discutido por James *et al*. (2023) e Zhang *et al*. (2021).

Figura 12: Melhores Modelos por Desempenho Agregado

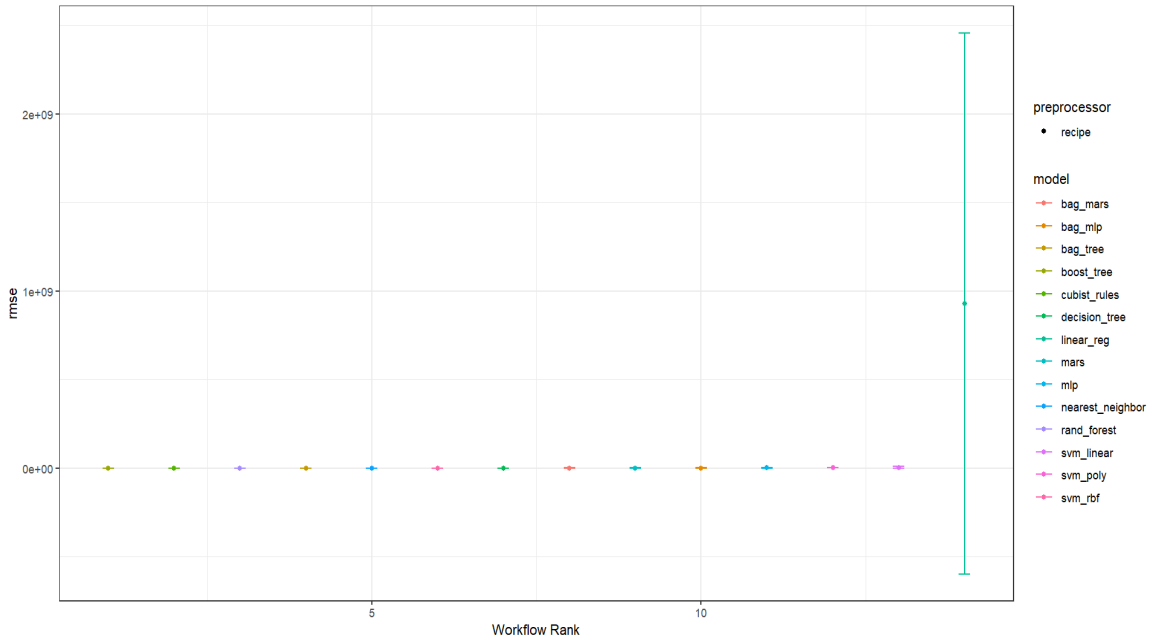


Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

A Figura 12 sintetiza os modelos com os melhores desempenhos agregados em termos de R^2 e RMSE. Nela, observa-se a dominância dos modelos *Bagged Trees*, *Random Forest* e *Boosted Trees* no grupo superior. Os intervalos de confiança dessas estimativas, além de estreitos, indicam estabilidade nos resultados, o que é desejável em modelos aplicados à avaliação imobiliária em bases urbanas heterogêneas.

Em contrapartida, modelos como o SVM com kernel polinomial e o *Elastic Net* apresentaram os piores desempenhos, com valores de R^2 próximos de 0,2 e RMSE elevados, o que sugere dificuldades em capturar as complexas relações presentes na base.

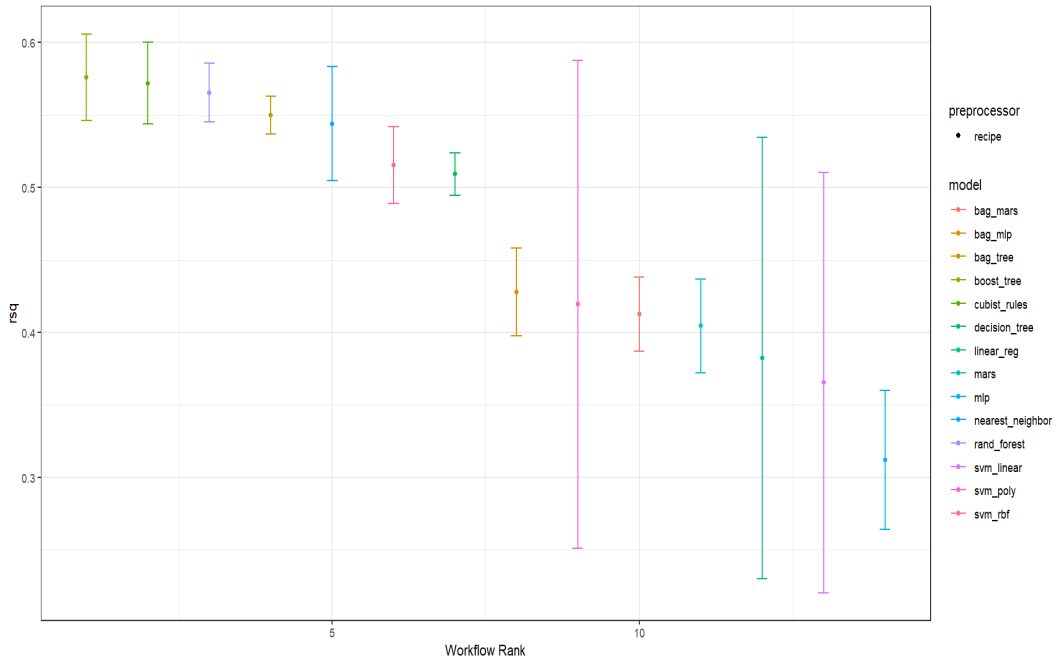
Figura 13: Avaliação Específica do RMSE



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Ao focar exclusivamente na métrica RMSE (Figura 13), percebe-se com maior nitidez que os modelos de árvore (*ensamble*) mantêm o menor erro médio de predição. O bom desempenho dos modelos *Bagged Trees* e *Random Forests* nessa métrica está em consonância com os achados de estudos aplicados ao mercado imobiliário como os de Root, Strader e Huang (2023) e Rodriguez-Serrano (2024), os quais destacam que tais métodos são mais apropriados para situações com forte heterogeneidade espacial e múltiplos atributos interativos.

Figura 14: Avaliação Específica do R^2



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Na última imagem (Figura 14), observa-se o ranqueamento dos modelos com base no coeficiente de determinação R^2. Mais uma vez, os modelos *Bagged Trees, Random Forest e Boosted Trees* figuram entre os mais eficientes em termos de explicação da variabilidade do valor de mercado dos imóveis. O modelo MLP (rede neural de múltiplas camadas) apresentou desempenho intermediário, o que pode estar relacionado à sensibilidade do modelo à normalização dos dados e ao número de amostras disponíveis, conforme evidenciado por Morettin e Singer (2022).

O modelo de regressão linear (Linear Regression), apesar de sua interpretabilidade, obteve desempenho inferior em R^2, situando-se entre 0,35 e 0,45, o que confirma as limitações dos modelos paramétricos em cenários com variáveis correlacionadas, como a distância aos ativos ambientais (parque e praia), cujo coeficiente de correlação, conforme previamente analisado, é de 0,998.

Já na fase de testes, a análise dos resultados obtidos nesta pesquisa evidencia que, entre os quatorze modelos de aprendizagem de máquina testados, cinco se destacaram em termos de desempenho preditivo, sobretudo quando se considera simultaneamente o coeficiente de determinação (R^2) e o erro quadrático médio (RMSE) (Tabela 5).

Esses dois indicadores são centrais na avaliação de modelos de regressão: o R^2 mensura a proporção da variabilidade da variável resposta explicada pelo modelo, enquanto o RMSE quantifica o erro médio das predições em valores absolutos.

Tabela 5: Desempenho dos modelos de aprendizagem de máquina com base em diferentes métricas de avaliação

Modelo	RMSE	MAE	MAPE	SMAPE	RSQ
Bagged Trees	163767.7	87546.53	71.95977	32.09382	0.5718944
Random Forest	166660.4	87483.95	72.66966	31.43243	0.5552101
Boosted Trees	170421.5	91137.24	75.59076	33.57981	0.5412897
Gaussian Kernel SVM	171007.1	84455.95	67.48392	31.11887	0.5377775
K-NN	177077.6	91179.93	72.45597	32.36496	0.5023987
Regression Tree	183406.4	98224.23	80.22300	36.25457	0.4819303
Bagged MLP	192122.5	104108.94	90.12895	38.59656	0.4501968
MARS	188232.7	111621.90	94.31211	42.34459	0.4324630
Cubist Rules	192581.3	89336.11	75.16964	32.21054	0.4319501
MLP	200804.1	119867.63	100.50801	45.19510	0.3605429
Linear SVM	207244.5	104543.33	80.73384	37.60360	0.3287224
Bagged MARS	206527.7	110872.73	96.31787	40.18823	0.3262206
Polynomial Kernel SVM	259447.9	145782.51	104.26709	55.85292	0.2140047
Elastic Net	279723.0	110242.05	91.09790	43.47377	0.1824059

Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

O modelo com melhor desempenho global foi o *Bagged Trees*, que alcançou o maior valor de R^2 dentre todos os modelos testados: 0,5719. Além disso, obteve o menor RMSE, de 163.767,7, o que indica menor erro absoluto médio na predição dos valores dos imóveis urbanos. Esse resultado reforça a robustez do modelo quando aplicado a contextos com variáveis altamente correlacionadas e heterogeneidade espacial, como é o caso da cidade de Palmas-TO. Resultados semelhantes foram encontrados por Paixão (2025), que utilizou *Bagging* em problemas de classificação social com múltiplos atributos, demonstrando sua capacidade de reduzir variância preditiva sem comprometer a flexibilidade do modelo.

Na segunda posição, observa-se o *Random Forest*, com um R^2 de 0,5552 e RMSE de 166.660,4. A leve queda em relação ao *Bagged Trees* pode ser atribuída ao controle mais agressivo sobre a complexidade das árvores, promovido pela aleatoriedade na escolha das variáveis em cada divisão. Ainda assim, seu desempenho foi superior à maioria dos modelos, o

que corrobora a literatura recente, como em Zhang *et al*. (2021), que identificou o Random Forest como um dos métodos mais confiáveis para predição de valores imobiliários.

O *Boosted Trees* aparece na terceira colocação, com $R^2 = 0,5413$ e RMSE de 170.421,5. Embora tenha apresentado o terceiro melhor coeficiente de determinação, seu RMSE foi ligeiramente mais alto do que os dois anteriores, sugerindo que, apesar da boa explicação da variabilidade, os erros absolutos de predição foram mais expressivos. Isso pode estar relacionado à sensibilidade do *Boosting* a outliers e à necessidade de ajuste fino dos hiperparâmetros, como já discutido por Friedman (2001) e confirmado em análises de Morettin e Singer (2022).

Em quarto lugar, destaca-se o modelo *Support Vector Machine* com Kernel Gaussiano, que atingiu um $R^2 = 0,5378$ e RMSE de 171.007,1. Embora seu RMSE esteja próximo ao do *Boosted Trees*, sua performance sugere um bom equilíbrio entre viés e variância, desde que os hiperparâmetros do kernel sejam adequadamente ajustados. É notável, porém, que o desempenho dos modelos baseados em SVM tendem a variar mais com o pré-processamento dos dados e com a escolha dos parâmetros de suavização, conforme discutido por James *et al*. (2023).

Finalizando a lista dos cinco melhores, encontra-se o modelo *K-Nearest Neighbors* (K-NN), com $R^2 = 0,5024$ e RMSE de 177.077,6. Embora mais simples e intuitivo, o K-NN mostrou-se competitivo, principalmente em regiões mais densas da cidade, onde a similaridade entre observações é mais evidente. No entanto, como também apontado por Lantz (2015), o desempenho do K-NN tende a se degradar em espaços de alta dimensionalidade e com variáveis em escalas distintas, o que limita sua eficácia em contextos mais heterogêneos.

Na sequência dos modelos com desempenho intermediário, observa-se a árvore de regressão (*Regression Tree*), que obteve um R^2 de 0,4819. Este resultado era esperado, tendo em vista a simplicidade estrutural do modelo, que não incorpora técnicas de regularização nem mecanismos de correção de viés.

Por outro lado, o modelo *Cubist Rules* apresentou um R^2 de 0,4319, com desempenho semelhante ao MARS (0,4324), indicando que a abordagem baseada em regras e modelos lineares locais, não alcançou os melhores resultados preditivos.

Entre os modelos baseados em redes neurais, tanto o MLP quanto o *Bagged* MLP, ficaram abaixo das expectativas, com R^2 de 0,3605 e 0,4502, respectivamente. A complexidade computacional, a sensibilidade à normalização das variáveis e a necessidade de um número elevado de observações para treinamento podem ter comprometido sua performance neste estudo, como já evidenciado por Morettin e Singer (2022).

Na parte inferior do ranking estão o Linear SVM (R^2 = 0,3287), *Bagged* MARS (0,3262), *Polynomial Kernel* SVM (0,2140) e *Elastic* Net (0,1824). Esses modelos apresentaram dificuldades em capturar as interações não lineares e as heterogeneidades espaciais presentes na base de dados. Como evidenciado por Marchetti e Teixeira (2018), modelos lineares penalizados como o *Elastic Net*, apesar de eficientes para seleção de variáveis, tendem a subestimar relações complexas quando aplicados em contextos urbanos.

Sob a ótica da influência das variáveis ambientais nos resultados, é possível afirmar que os modelos que melhor captaram a relação entre proximidade a áreas verdes e corpos d'água, como o Parque Cesamar e a Praia da Graciosa, foram justamente os modelos baseados em árvores. A capacidade desses algoritmos de identificar divisões não lineares no espaço amostral foi crucial para mapear padrões de valorização fundiária associados à presença de amenidades ambientais. Esse comportamento reforça o pressuposto da teoria do valor hedônico, segundo a qual atributos ambientais, mesmo não transacionáveis de forma direta, são capitalizados no valor dos imóveis (ROSEN, 1974).

Esse resultado também dialoga com estudos realizados em outras localidades. Batalhone *et al.* (2002) e Dantas (2005) demonstraram que a proximidade a recursos naturais urbanos, como praças e lagos, influencia positivamente o preço dos imóveis. Da mesma forma, estudos internacionais, como o de Gibbons, Mourato e Resende (2014), confiram que compradores estão dispostos a pagar mais para viver próximo de áreas verdes urbanas.

Portanto, a interpretação conjunta dos resultados empíricos e da literatura permitem afirmar que a presença de variáveis ambientais na base de dados contribuiu significativamente para o desempenho dos modelos mais complexos. Esses modelos captaram não apenas o efeito isolado da proximidade a parques e áreas de água, mas também suas interações com outras variáveis urbanas, como padrão construtivo, localização dentro do plano diretor e nível de urbanização do entorno.

Esses achados estão em consonância com o estudo de Paixão (2025), que observou desempenho superior dos modelos ensemble em comparação às abordagens tradicionais em um contexto aplicado à identificação de condição de pobreza. A convergência desses estudos evidencia que, para problemas de precificação fundiária urbana com múltiplas variáveis interativas e relações não lineares, os modelos baseados em árvores, especialmente em sua forma agregada, são mais indicados.

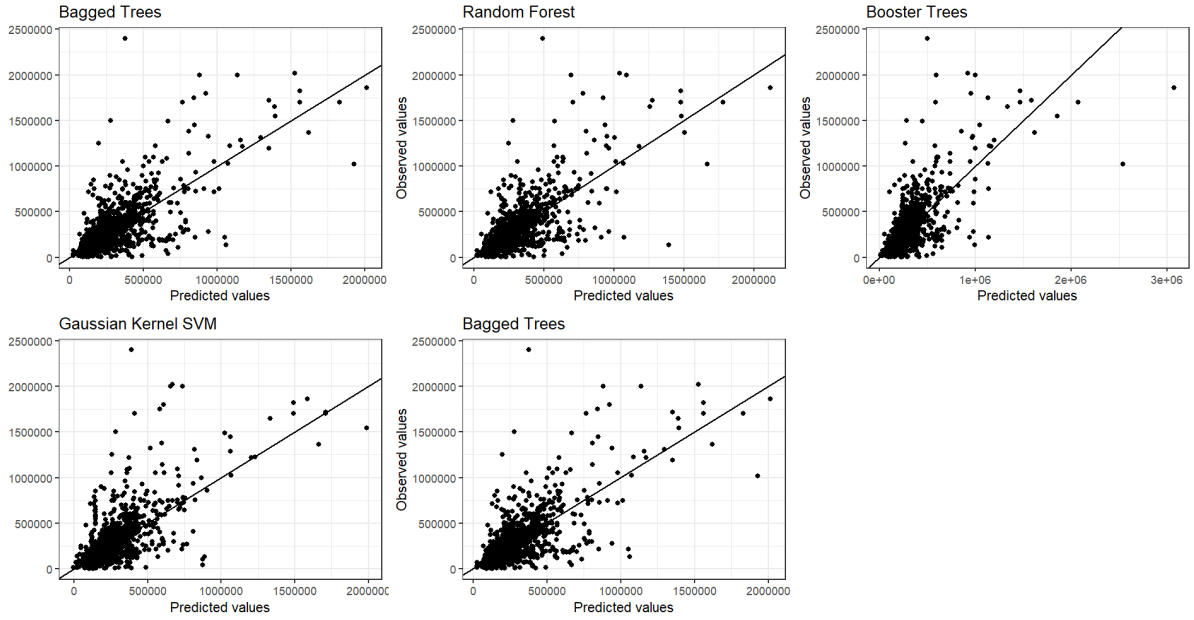
A inclusão das variáveis de distância aos ativos ambientais (Parque Cesamar e Praia da Graciosa) como variáveis contínuas e categorizadas (*dummies* de proximidade), mostrou-se determinante para a performance dos modelos com melhor desempenho. Isso valida

empiricamente a hipótese central deste estudo. Ou seja, a presença de ativos ambientais é internalizada no preço dos imóveis urbanos, como já defendido por Rosen (1974) e operacionalizado no Método dos Preços Hedônicos.

Os modelos baseados em árvores, como o *Bagged Trees* e o *Random Forest*, foram particularmente eficazes em captar os efeitos não lineares dessas variáveis ambientais e suas interações com atributos fiscais, físicos e socioeconômicos dos imóveis. Tal comportamento também foi verificado nos estudos de Batalhone *et al*. (2002) e Dantas (2005), que aplicaram o MPH em contextos urbanos brasileiros, evidenciando a capitalização de amenidades ambientais nos preços de mercado.

A Figura 15 apresenta os gráficos de dispersão entre os valores observados e os valores preditos pelos modelos com melhor desempenho na fase de teste: Bagged Trees, *Random Forest*, *Boosted Trees* e *Support Vector Machine* com Kernel Gaussiano (SVM-RBF).

Figura 15: Gráficos de dispersão entre os valores observados e os valores preditos



Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

A linha diagonal em cada gráfico representa a situação ideal de predição perfeita, onde os valores preditos coincidem exatamente com os valores observados. Assim, quanto mais próximos os pontos estiverem dessa linha, maior é a acurácia do modelo no processo de ajuste.

A visualização do *Bagged Trees* revela uma forte concentração dos pontos ao longo da linha de identidade, com dispersão relativamente baixa mesmo para valores elevados de imóveis. Este comportamento indica excelente capacidade de ajuste do modelo aos dados de

treinamento, refletindo seu elevado R^2 (> 0,57) e menor RMSE (< 165 mil), conforme discutido anteriormente. Isso reforça o potencial do *bagging* para reduzir a variância do modelo sem incorrer em *overfitting*, conforme demonstrado por James *et al*. (2023).

O modelo *Random Forest* também apresenta uma distribuição dos pontos bastante próxima da diagonal, embora com leve aumento na dispersão em imóveis de alto valor. A boa performance nesse estágio está de acordo com sua elevada estabilidade e eficiência na captura de interações entre variáveis ambientais, físicas e fiscais. Este desempenho durante o treino, como indicado pela concentração dos pontos, reforça sua capacidade de modelar estruturas complexas, característica amplamente descrita por Zhang *et al*. (2021).

O *Boosted Trees* mostra alta concentração dos pontos até aproximadamente R$ 1.000.000,00, porém com alguma dispersão à medida que os valores observados aumentam. Essa característica está relacionada ao potencial de *overfitting* do algoritmo, especialmente quando os hiperparâmetros não são devidamente regulados. Mesmo assim, o modelo mostrou excelente desempenho na fase de treino, sendo um dos três melhores em R^2 e RMSE.

O gráfico do SVM com kernel gaussiano evidencia desempenho competitivo, especialmente para imóveis até R$ 800.000,00, com boa aderência à linha de identidade. Contudo, a maior dispersão em valores elevados sugere certa limitação na generalização do modelo para imóveis fora do padrão médio, comportamento coerente com a natureza do SVM, que requer cuidadoso ajuste de hiperparâmetros (C e γ), como destacado por Cortes e Vapnik (1995).

Os gráficos apresentados confirmam a conclusão estatística obtida com base nas métricas R^2 e RMSE: os modelos *Bagged Trees*, *Random Forest* e *Boosted Trees* foram os que melhor generalizaram os padrões de preços dos imóveis urbanos na cidade de Palmas-TO. Esses algoritmos demonstraram elevada capacidade de prever, com razoável precisão, os valores de imóveis com base em seus atributos físicos, locacionais, fiscais e, em especial, ambientais, tais como a proximidade ao Parque Cesamar e à Praia da Graciosa.

Esses resultados visuais reforçam a validade da hipótese deste estudo, no sentido de que atributos ambientais, mesmo não sendo diretamente precificados pelo mercado, são internalizados nos valores das transações e podem ser captados por modelos de aprendizagem de máquina, conforme defendido por Rosen (1974) e Jim e Chen (2006).

Os resultados demonstram, de forma clara, a superioridade dos modelos de Aprendizagem de Máquina (AM) em relação aos Modelos Lineares (ML) e Lineares Generalizados (MLG), conforme apresentado na Tabela 6.

Tabela 6: Tabela Comparativa dos Modelos

Modelo	**Tipo**	**R^2/Pseudo R^2**	**RMSE**	**Testes Aplicados**
Modelo A - GLM Distâncias	Generalizado	0.0161	≈ 202.800	Breusch-Pagan, White, Lilliefors, Correção robusta
Modelo B - GLM Dummies	Generalizado	0.0164	≈ 202.700	Breusch-Pagan, White, Lilliefors, Correção robusta
Modelo C - LM Distâncias	Linear	0.311	≈ 202.800	Breusch-Pagan, White, Lilliefors, Correção robusta
Modelo D - LM Dummies	Linear	0.312	≈ 202.700	Breusch-Pagan, White, Lilliefors, Correção robusta
Bagged Trees	Aprendizagem de Máquina	0.5719	163767.7	Validação cruzada, tuning de hiperparâmetros
Random Forest	Aprendizagem de Máquina	0.5552	166660.4	Validação cruzada, tuning de hiperparâmetros
Boosted Trees	Aprendizagem de Máquina	0.5413	170421.5	Validação cruzada, tuning de hiperparâmetros
SVM - Kernel Gaussiano	Aprendizagem de Máquina	0.5378	171007.1	Validação cruzada, tuning de hiperparâmetros
K-Nearest Neighbors	Aprendizagem de Máquina	0.5024	177077.6	Validação cruzada, tuning de hiperparâmetros

Fonte: Elaborado pelo autor a partir dos resultados da pesquisa.

Enquanto os MLGs (modelos A e B) apresentaram *pseudo* R^2 muito baixos (≈ 0,016) e elevados erros (RMSE ≈ 202.700), os MLs (Modelos C e D) mostraram desempenho moderado, com R2em torno de 0,31 e RMSE, também elevado.

Em contraste, os cinco melhores modelos de AM apresentaram desempenhos superiores, com R^2 variando de 0,50 a 0,57 e RMSE significativamente menores, com destaque para o *Bagged Trees* (R^2 = 0,5719; RMSE ≈ 163.767), seguido por *Random Forest* e *Boosted Trees*.

Esses resultados demonstram que os modelos de AM capturam com maior eficácia padrões não lineares e interações complexas entre variáveis, sendo mais adequados para a previsão do valor de imóveis em mercados urbanos heterogêneos como Palmas-TO.

Além disso, esses modelos se mostraram particularmente eficazes em captar a influência das variáveis ambientais (proximidade ao Parque Cesamar e à Praia da Graciosa) na formação dos valores dos imóveis, superando as limitações observadas nos modelos lineares, que não conseguiram representar de forma robusta tais efeitos. A robustez das técnicas baseadas em árvores de decisão foi particularmente evidente, confirmando tendências recentes na literatura correlacionada.

6 CONSIDERAÇÕES FINAIS

6. 1 Conclusões

Os resultados deste estudo confirmaram, de forma robusta, que as variáveis físicas e fiscais são os principais determinantes do valor de transação dos imóveis urbanos na cidade de Palmas-TO, reforçando a validade empírica do Método dos Preços Hedônicos (MPH) como ferramenta de explicação para os fatores tradicionais que influenciam a formação de preços no mercado imobiliário.

Por outro lado, os Modelos Lineares (ML) e Lineares Generalizados (MLG) apresentaram desempenhos moderados, com R^2 em torno de 0,31 e pseudo R^2 inferior a 0,02, além de elevados erros de predição (RMSE ≈ 202.700), evidenciando limitações importantes na capacidade de captura de relações complexas e não lineares, especialmente no que se refere à influência das variáveis ambientais sobre os valores dos imóveis

Em contraste, os algoritmos de Aprendizagem de Máquina (AM) demonstraram desempenho superior, com destaque para o *Bagged Trees* (R^2 = 0,5719; RMSE = 163.767,7), seguido por *Random Forest* e *Boosted Trees*, comprovando maior eficácia na modelagem de padrões não lineares e interações complexas entre variáveis, bem como na identificação dos efeitos das variáveis ambientais na valorização imobiliária, superando as limitações dos modelos tradicionais.

Ainda que as variáveis ambientais apresentem correlações diretas de baixa magnitude, os modelos de AM foram capazes de identificar que a proximidade aos ativos ambientais (especialmente ao Parque Cesamar e à Praia da Graciosa) exerce influência significativa na formação dos preços dos imóveis, validando a hipótese central da tese e reforçando a importância das amenidades ambientais como componentes valorativos no mercado imobiliário urbano.

No que se refere ao atendimento dos objetivos específicos propostos, constatou-se que todos foram plenamente alcançados. Primeiramente, foi possível identificar a influência dos atributos ambientais (Parque Cesamar e Praia da Graciosa) na precificação dos imóveis transacionados e localizados no Plano Diretor de Palmas-TO. Os resultados obtidos evidenciaram a existência de uma relação significativa, entre a proximidade dos imóveis a tais ativos ambientais e a valorização imobiliária, corroborando com achados similares na literatura.

Em seguida, o segundo objetivo específico foi igualmente cumprido ao se desenvolver um modelo de valoração ambiental a partir da aplicação do Modelo Linear Generalizado (MLG) e do Modelo Linear (ML), os quais permitiram mensurar, ainda que com limitações, o impacto da proximidade aos ativos ambientais sobre os preços dos imóveis. As análises indicaram que, embora os modelos lineares consigam capturar parcialmente as influências diretas dos atributos físicos e fiscais, apresentaram importantes restrições na detecção de efeitos não lineares e complexos decorrentes das variáveis ambientais.

Do mesmo modo, o terceiro objetivo específico foi plenamente atendido com a implementação de um modelo de valoração ambiental utilizando técnicas de Aprendizagem de Máquina, o qual se revelou mais eficiente na identificação da influência dos ativos ambientais nos valores imobiliários urbanos. Destaca-se, nesse sentido, o desempenho superior dos modelos *Bagged Trees*, *Random Forest* e *Boosted Trees*, que, ao superarem os modelos tradicionais, demonstraram capacidade mais robusta para modelar padrões complexos e captar interações múltiplas entre as variáveis analisadas.

O quarto objetivo específico também foi contemplado, ao se proceder à avaliação e comparação da relevância das variáveis ambientais, físicas, fiscais e de renda na determinação dos preços dos imóveis urbanos, a partir das abordagens modeladas. As análises demonstraram que as variáveis físicas e fiscais, como área do terreno e índice fiscal, mantém-se como principais determinantes do valor dos imóveis. Contudo, evidenciou-se que a variável ambiental (proximidade ao Parque Cesamar e à Praia da Graciosa), embora com menor peso relativo, apresenta influência estatisticamente significativa, sobretudo quando analisada por meio dos modelos de Aprendizagem de Máquina.

Já o quinto objetivo foi igualmente atingido ao se comparar o desempenho preditivo entre o Modelo Linear Generalizado, Modelo Linear e os Modelos de Aprendizagem de Máquina, apontando-se as vantagens e limitações de cada metodologia no contexto analisado. Os modelos de AM destacaram-se pelo elevado desempenho preditivo, maior capacidade de captura de relações não lineares e menor erro de predição, ao passo que os modelos tradicionais, embora mais parcimoniosos e de interpretação direta, mostraram-se limitados frente à complexidade das relações existentes no mercado imobiliário urbano de Palmas-TO.

Nesse cenário, os resultados obtidos também trazem reflexões importantes para o campo do planejamento urbano. Fica evidente que a valorização imobiliária não pode ser entendida apenas pelos atributos físicos ou fiscais, mas está fortemente ligada à localização, às condições de acessibilidade e à presença de ativos ambientais e de infraestrutura. A análise mostra que a inclusão dessas variáveis ambientais na modelagem dos preços dos imóveis oferece subsídios

técnicos valiosos para orientar políticas de ordenamento territorial. Isso abre caminho para que a expansão urbana de Palmas seja conduzida de forma mais equilibrada e sustentável. Desse modo, a tese ressalta a importância de articular a avaliação imobiliária com instrumentos de política tributária e de planejamento urbano, de modo a alinhar a valorização do solo urbano com objetivos mais amplos de justiça socioespacial e preservação da qualidade ambiental.

Por fim, os achados indicam a necessidade de ampliar a utilização de modelos não paramétricos e de aprendizagem de máquina na avaliação imobiliária, especialmente em contextos urbanos caracterizados por alta heterogeneidade espacial e múltiplas interações entre variáveis. Recomendando-se ainda, a incorporação de variáveis adicionais relacionadas à percepção de qualidade ambiental, segurança e infraestrutura urbana, bem como a replicação da metodologia proposta em outras cidades brasileiras, visando a generalização e comparação dos resultados.

6. 2 Limitações do Estudo

Apesar dos avanços alcançados nesta pesquisa, algumas limitações metodológicas e operacionais merecem ser destacadas. Primeiramente, destaca-se a limitação relacionada à base de dados utilizada, que compreende exclusivamente os registros do Imposto sobre Transmissão de Bens Imóveis (ITBI) referentes aos anos de 2019, 2020 e 2021. Tal base, embora representativa, pode conter distorções oriundas de subdeclarações ou de lacunas quanto à caracterização completa dos imóveis, comprometendo a acurácia dos modelos ajustados.

Ademais, a mensuração das variáveis ambientais, distância ao Parque Cesamar e à Praia da Graciosa, foi realizada com base em distâncias euclidianas entre os centróides das quadras e os referidos ativos ambientais. Essa abordagem, embora usual em trabalhos similares, desconsidera barreiras físicas e aspectos da malha urbana que influenciam a percepção de proximidade e acessibilidade real aos ativos ambientais.

Também se reconhece que o modelo de valoração foi estimado com base em informações agregadas por quadra, não sendo possível capturar as especificidades de cada unidade imobiliária transacionada. Além disso, embora tenham sido testados diversos modelos de aprendizagem de máquina, com destaque para o bom desempenho dos algoritmos *Bagged Tree, Random Forest* e *SVM com Kernel Polinomial*, a ausência de dados complementares, como características construtivas detalhadas e aspectos subjetivos da vizinhança, pode ter limitado a capacidade preditiva dos modelos.

Por fim, cumpre ressaltar que a pandemia da COVID-19, ocorrida no período de abrangência da base de dados, pode ter introduzido distorções nos preços de mercado e nas dinâmicas de valoração imobiliária, implicando potenciais efeitos sazonais que não puderam ser integralmente controlados.

6.3 Recomendações para Pesquisas Futuras

Considerando as limitações evidenciadas nesta pesquisa e os avanços recentes no campo da modelagem estatística e da ciência de dados, diversas direções podem ser propostas para o aprofundamento de estudos futuros sobre avaliação imobiliária com ênfase em variáveis ambientais.

Inicialmente, recomenda-se a ampliação e diversificação das fontes de dados, com a inclusão de registros provenientes do Cadastro Técnico Multifinalitário (CTM), do banco de dados da Caixa Econômica Federal utilizada para avaliação de imóveis em processos de financiamento habitacional, da base de dados da Receita Federal do Brasil, especialmente no que se refere à Declaração de Informações sobre Atividades Imobiliárias (DIMOB), e de cadastros administrativos georreferenciados. A integração de dados de sensoriamento remoto, imagens de satélite de alta resolução e dados de mobilidade urbana pode enriquecer substancialmente a base empírica, permitindo análises mais granulares sobre a configuração espacial, uso do solo e padrões de urbanização.

No que se refere às variáveis ambientais, futuras investigações poderiam incorporar indicadores mais abrangentes de qualidade ambiental, tais como a cobertura vegetal, os índices de temperatura de superfície, a poluição sonora e a proximidade a áreas de preservação permanente, utilizando técnicas de geoprocessamento e análise espacial. A utilização de métricas baseadas em tempo de deslocamento, como o tempo médio de caminhada ou de transporte público até os ativos ambientais, pode substituir as medidas euclidianas utilizadas nesta pesquisa, refletindo de forma mais realista a percepção de acessibilidade.

Do ponto de vista metodológico, sugere-se o aprofundamento na utilização de algoritmos de aprendizagem profundo (*Deep Learning*), incluindo Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN), especialmente em estudos que integrem imagens urbanas e séries temporais. Adicionalmente, a incorporação de técnicas de ensemble learning mais robustas, como o *XGBoost e o LightGBM*, poderá contribuir para o aumento da acurácia preditiva e a redução da variância nos modelos.

Um aspecto que também merece atenção está ligado ao planejamento urbano. Os achados desta tese mostram que considerar variáveis ambientais na avaliação imobiliária pode gerar informações valiosas para orientar estratégias de ordenamento territorial. Pesquisas futuras podem aprofundar a análise da relação entre valorização imobiliária, acessibilidade e infraestrutura, reforçando o elo entre avaliação de imóveis, política tributária e planejamento urbano. Essa abordagem tem potencial não apenas para aprimorar a gestão do solo urbano em Palmas, mas também para servir de referência em estudos comparativos em outras cidades brasileiras, conectando valorização fundiária, justiça socioespacial e sustentabilidade ambiental.

Por fim, reforça-se a importância de replicar este estudo em diferentes contextos urbanos, considerando cidades com distintas escalas populacionais, regimes fundiários e dinâmicas de crescimento. A comparação entre distintas realidades regionais pode contribuir para a formulação de políticas públicas mais eficientes, orientadas por evidências empíricas robustas, e para o aprimoramento dos modelos de valoração utilizados na gestão territorial e na administração pública.

REFERÊNCIAS

ADAMOWICZ, W.; LOUVIERE, J.; WILLIAMS, M. Combining revealed and stated preference methods for valuing environmental amenities. **Journal of Environmental Economics and Management**, v. 26, p. 291–292, 1994.

ARRAES, R. A.; FILHOII, E. S. Economia Aplicada - Externalidades e formação de preços no mercado imobiliário urbano brasileiro: um estudo de caso. **Economia Aplicada**, Ribeirão Preto, v.12, n. 2, 2008.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14.653-2**: Avaliação de Imóveis Urbanos. Referência: Elaboração. Rio de Janeiro: ABNT, 2011, 54 p.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14.653-1**: Avaliação de Imóveis Urbanos. Referência: Elaboração. Rio de Janeiro: ABNT, 2019, 10 p.

AGUIRRE, A.; FARIA, D. V. A utilização de preços hedônicos na avaliação social de projetos. **Revista Brasileira de Economia**, v. 51, n. 3, p. 391–411, 1997.

ALBUQUERQUE, E.; MELO, A.; SOUZA, H. Ativo ambiental e preço de imóvel em Recife: um estudo exploratório a partir da utilização do método dos preços hedônicos. *In*: VII ENCONTRO DA SOCIEDADE BRASILEIRA DE ECONOMIA ECOLÓGICA, 2007, Fortaleza. **Anais** [...]. Fortaleza, Ceará, 2007.

AMREIN, C. **Capital humano e capital urbano: o impacto das escolas nos preços dos imóveis no município de São Paulo**. 2010. Dissertação (Mestrado em Ciências Econômicas) - Universidade de São Paulo, Faculdade de Economia, Administração e Contabilidade, São Paulo, 2010.

BAUMOL, W. J.; OATES, W. E. **The theory of environmental policy**. Second edition (Reprinted). London: Cambridge University Press, 1998. p.299.

BAUR, D. G.; PORTMANN, D.; ELMINGER, A. Automated real estate valuation with machine learning models using property descriptions. **Preprint**, 2022. Disponível em: https://www.researchgate.net/publication/365347461_Automated_real_estate_valuation_with_machine_learning_models_using_property_descriptions . Acesso em: 19 maio 2025.

BATALHONE, M. *et al*. Avaliação econômica da vegetação urbana: aplicação do método de preços hedônicos na cidade de Londrina. **Revista de Economia e Sociologia Rural**, v. 40, n. 1, p. 107–126, 2002.

BATALHONE, S.; NOGUEIRA, J.; MUELLER, B. **Economics of air pollution: hedonic price model and smell consequences of sewage treatment plants in urban areas**. Brasília, DF: Universidade de Brasília, 2002.

BATEMAN, I. J. *et al*. 2002. **Economic Valuation with Stated Preference Techniques: A Manual**. Edward Elgar, Cheltenham, UK, 2002. p.458.

BELLIA, V**. Introdução à Economia do Meio Ambiente. Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis**. Brasília, 1996. p. 262.

BENAKOUCHE, R.; CRUZ, R. S. **Avaliação Monetária do Meio Ambiente**. Ed. Makron Books do Brasil Ltda. São Paulo, 1994.

BONDUKI, Nabil. **Política habitacional e inclusão social no Brasil: revisão histórica e novas perspectivas no governo Lula**. *Revista Eletrônica de Arquitetura e Urbanismo*, São Paulo, n. 4, p. 70-104, 2010.

BORBA, M. R. **Modelo de avaliação da propriedade imobiliária referenciado à qualidade ambiental, por meio do método dos valores hedônicos, como instrumento para estudos de impacto ambiental**. São Paulo: Universidade de São Paulo, 1992.

BRASIL. **Estatuto da Cidade**. Lei nº 10.257, de 10 de julho de 2001.

BRESSER-PEREIRA, L. C. Desenvolvimento, progresso e crescimento econômico. **Lua Nova**, n. 93, p.33-60, 2014.

BRUECKNER, J.; THISSE, J.; ZENOU, Y. Why is central Paris rich and downtown Detroit poor? An amenity based theory. **European Economic Review**, v. 43, p. 91–107, 1999.

CLARK, D. S. Externality effects on residential property values: the example of noise disamenites. **Growth and Change**, v. 37, n. 3, p.460-488, 2006.

CHOY, M.; HO, C. Forecasting housing prices using machine learning models. **Land, Basel**, v. 12, n. 4, 740, 2023. Disponível em: https://www.mdpi.com/2073-445X/12/4/740 . Acesso em: 19 maio 2025.

COMISSÃO MUNDIAL SOBRE MEIO AMBIENTE E DESENVOLVIMENTO. **Nosso futuro comum**. Rio de Janeiro: FGV, 1988.

CORDEIRO, G. M.; DEMÉTRIO, C. G. **Modelos lineares generalizados e extensões**. São Paulo, 2008.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995.

DANTAS, M. L. C. **Composto mercadológico de imóveis residenciais: uma análise do ponto de vista do incorporador e do cliente**. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis, 2000.

DANTAS, R. A. **Engenharia de Avaliações: uma introdução à metodologia científica**. São Paulo: Pini, 2005.

DANTAS, R. F. **Valoração econômica de recursos naturais por meio do método de preços hedônicos: uma aplicação no município de Lavras/MG**. 2005. Dissertação (Mestrado em Administração) – Universidade Federal de Lavras, Lavras, 2005.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. 2002.

DENG, X. Enhanced AVM based on multi-source image fusion and deep learning. PLoS ONE, San Francisco, v. 20, n. 1, e0321951, 2025. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0321951 . Acesso em: 19 maio 2025.

DIAMANTOUDI, E.; SARTZETAKIS, E. S. Stable international environmental agreements: an analytical approach. *In*: CONFERÊNCIA CREE, 2001, Montreal. **Anais** [...] Montreal: CREE, 2001.

DIEWERT, W. E. **The Paris OECD-IMF Workshop on Real Estate Price Indexes: conclusions and future directions**. 2009. *In*: DIEWERT, E. Erwin *et al*. (Ed.). Price and Productivity Measurement: volume 1 – housing. Trafford Press, 2009. Disponível em: http://faculty.arts.ubc.ca/ediewert/dp0701.pdf . Acesso em: set. 2019.

FARIA, R. E. A. Uma aplicação do método de preços hedônicos no setor de saneamento: o projeto de São Bento do Sul-SC. **Planejamento e Políticas Públicas**, v.31, p.116–127, 2008.

FAVERO, L. P. L.; BELFIORE, P. P.; LIMA, G. A. S. F. Modelos de precificação hedônica de imóveis residenciais na região metropolitana de São Paulo: uma abordagem sob as perspectivas da demanda e da oferta. **Estudos Economicos**, v. 38, n. 1, p. 73-96, 2008.

FIKER, José. **Manual de avaliações e perícias em imóveis urbanos**. 3. Ed. São Paulo: Pini, 2008.

FLORENCIO, L. D. A. **Engenharia de avaliações com base em modelos GAMLSS**. Dissertação (Mestrado em Estatística) – Universidade Federal de Pernambuco, 2010.

FREEMAN, A. Hedonic prices, property values and measuring environmental benefits: a survey of the issues. **The Scandinavian Journal of Economics**, v.81, n. 2, p.154-173, 1979.

FOSTER, J.; GREER, J.; THORBECKE, E. A class of decomposable poverty measures. **Econometrica**, v. 52, n. 3, p. 761–768, 1984.

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

GAZOLA, S. **Construção de um modelo de regressão para avaliação de imóveis**. Dissertação (Mestrado em Engenharia de Produção) Universidade Federal de Santa Catarina, Florianópolis, 2002.

GIBBONS, S.; MACHIN, S. Valuing English primary schools. **Journal of Urban Economics**, v. 53, n. 2, p. 197-219, 2003.

GIBBONS, S.; MOURATO, S.; RESENDE, G. The amenity value of English nature: a hedonic price approach. **Environmental and Resource Economics**, v. 57, n. 2, p. 175–196, 2014.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro: Alta Books, 2019.

GOMIDE, Tito Lívio Ferreira. **Engenharia Legal: novos estudos**. São Paulo: Liv. e Ed. Universitária de Direito, 2008.

GONZÁLEZ, Marco Aurélio Stumpf. **Metodologia de avaliação de imóveis** – Novo Hamburgo: SGE, 2003.

GONZÁLEZ, M. A. S. Fonte alternativa de informação para estudos infraurbanos. *In*: ITBI. ENCONTRO NACIONAL DA ANPUR, 1997 a, Recife. **Anais** [...]. 1997. p. 129-147.

HANSEN, B. E. **Econometrics**. [S.l.]: Mimeo, 2021.

HANLEY, N., SHOGREN, J. F.;WHITE, B. **Introduction to Environmental Economics**. Oxford University Press, 2001.

HERMANN, B. M.; HADDAD, E. A. Mercado imobiliário e amenidades urbanas: a view through the window. **Estudos Econômicos**, v. 35, n. 2, 2005.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009.

HOCHHEIM, Norberto *et al*. **Metodologia de valoração ambiental para diferentes tipos de uso e ocupação do solo**. Relatório do Projeto Integrado de Pesquisa financiado pelo CNPq. Florianópolis, 2001.

HOFFMAN, R.; VIEIRA, S. **Análise de regressão: Uma introdução à econometria**. São Paulo: Hucitec, 1977.

IZBICKI, R.; SANTOS, T. M. **Machine Learning sob a Ótica Estatística: Uma abordagem preditivista para a estatística com exemplos em R**. [S.l.: s.n.], 2019.

JAMES, G. *et al*. **An Introduction to Statistical Learning – with Applications in R**. New York: Springer, 2013.

JAMES, G. *et al*. **An Introduction to Statistical Learning: with Applications in R**. 2. ed. Nova York: Springer, 2023.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2007.

JIANG, H. Machine learning fundamentals: A concise introduction. 1. ed. [S.l.]: Cambridge University Press, 2021. ISBN 978-1108837040.

JIM, C.Y.; CHEN, W. Y. Impacts of urban environmental elements on residential housing prices in Guangzhou. **Landscape and Urban Planning**, v. 78, p. 422-434, 2006.

KWAK, S. J.; YOO, S. H.; HAN, S.Y. Estimating the public's value for urban forest in the Seoul metropolitan area of Korea: a contingent valuation study. **Urban Stud**, v. 40, p. 2207–2221, 2003.

LANTZ, B. **Machine Learning with R**. 2. ed. [S.l.]: Packt Publishing, 396 p, 2013.

LANTZ, B. **Machine Learning with R.** 2. ed. Birmingham: Packt Publishing, 2015.

LARSON, R.; FARBER, B. **Estatística aplicada**. Tradução de Luciane Ferreira Pauleti Viana. 4. ed. São Paulo: Pearson Prentice Hall, 2010.

LORUSSO, Marco. **Qualidade ambiental urbana: critérios e indicadores**. 1992. Tese (Doutorado em Arquitetura e Urbanismo) - Faculdade de Arquitetura e Urbanismo, Universidade de São Paulo, São Paulo, 1992.

MARCHETTI, R.; TEIXEIRA, J. P. Modelagem estatística do preço de imóveis urbanos com regressão penalizada. **Revista Brasileira de Estudos Regionais e Urbanos**, v. 12, n. 1, p. 41–59, 2018.

MARICATO, Ermínia. **As ideias fora do lugar e o lugar fora das ideias. A cidade do pensamento único: desmanchando consensos**. Tradução . Petrópolis: Vozes, 2013. Acesso em: 11 set. 2025.

MARGULIS, Sergio. **Meio Ambiente: Aspectos Técnicos e Econômicos**. 2· ed. Brasília: IPEA, 1996. p. 246.

MARQUES, J. F.; COMUNE, A. E. A. Teoria neoclássica e a valoração ambiental. *In*: ROMERO, A. R. *et al*. **Economia do meio ambiente: teoria, políticas e a gestão de espaços regionais**. Campinas, UNICAMP, IE, 1997.

MCCULLAGH, P. E.; NELDER, J. A. **Generalized Linear Models**, 2. ed. Chapman & Hall/CRC, 1999.

MCCULLOCH, W. S.; PITTS, W. **A Logical Calculus of the Ideas Immanent in Nervous Activity**, 1943.

MENDONÇA, S. A. de. **Avaliação de imóveis urbanos e rurais**. 3. ed. São Paulo: Pini, 1998.

MERICO, Luiz Fernando Krieger. **Introdução à economia ecológica**. Blumenau: FURB, 1996.

MORETTIN, P. A.; SINGER, J. M. **Estatística e Ciência de Dados**. Rio de Janeiro: LTC, 2022.

MORO, M.; BRERETON F.; FERREIRA S.; CLINCH J. P. Methods Ranking quality of life using subjective well-being data. **University College**, Dublin, 2008.

MOTTA, Ronaldo Seroa da. **Manual para Valoração Econômica de Recursos Ambientais**. IPEA/MMA/PNUQ/CNPq. Rio de Janeiro, 1997.

MOUNA, A. *et al*. A comparative study of urban house price prediction using machine learning algorithms. **Preprint**, 2023. Disponível em: https://www.researchgate.net/publication/373200743_A_Comparative_Study_of_Urban_House_Price_Prediction_using_Machine_Learning_Algorithms. Acesso em: 19 maio 2025.

NADAL, A. C.; JULIANO, K. A.; RATTON, E. Testes estatísticos utilizados para a validação de regressões múltiplas aplicadas na avaliação de imóveis urbanos. **Boletim de Ciências Geodésicas**, v. 9, n. 2, p.243-262, 2003.

NIELSEN, A. **Análise Prática de Séries Temporais: Predição com Estatística e Aprendizado de Máquina**. 1. ed. Rio de Janeiro: Alta Books, 2021.

OPENAI. ChatGPT. [S.l.]: OpenAI, 2025. Disponível em: https://chat.openai.com. Acesso em: 22 maio 2025.

PAIXÃO, A. N. **Avaliação Contingente de Serviços de Saneamento Básico em Palmas-TO**. 2008. Tese (Doutorado em Economia Aplicada) – Universidade Federal de Viçosa, 2008.

PAIXÃO, A. N. **Modelos de Aprendizagem de Máquina aplicados à classificação da pobreza na Paraíba**. João Pessoa: UFPB, 2025.

PAIXÃO, L. A. R. Índice de preços hedônicos para imóveis: uma análise para o município de belo horizonte. **Economia Aplicada**, v. 19, n. 1, p. 5-29, 2015.

PAREDES, M. **Estimación del valor económico de los servicios ambientales en la cuenca del río Rimac, Lima-Perú**. Pontificia Universidad Católica del Perú, 2005.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP. São Paulo, 2004.

RADEGAZ, Násser Júnior. **Avaliação de bens: princípios básicos e aplicações**. São Paulo: Liv. e Ed. Universitária de Direito, 2011.

PINDYCK, R. S.; RUBINFELD, D. L. **Microeconomia**. São Paulo: Pearson, 2014.

ROCHA, Raquel Resende. **Técnicas de geoprocessamentos aplicadas à avaliação de imóveis. Estudo de Caso: Região Central de Ibirité**. 2005. Monografia (Especialização) - Universidade Federal de Minas Gerais, Instituto de Geociências. Belo Horizonte, 2005.

RODRIGUEZ-SERRANO, D. **Explainable property valuation via prototype learning: a practical application of XAI in real estate**. Annals of Operations Research, Dordrecht, 2024. Disponível em: https://link.springer.com/article/10.1007/s10479-024-06273-1. Acesso em: 19 maio 2025.

ROLNIK, Raquel. **Guerra dos lugares: a colonização da terra e da moradia na era das finanças**. 2015. Tese (Livre Docência) – Universidade de São Paulo, São Paulo, 2015. Acesso em: 11 set. 2025.

ROSENBLATT, F. **The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain**, 1958.

ROSEN, S. Hedonic prices and implicit market: product differentiation in pure competition. **Journal of Political Economy**, v. 82, p. 34- 55, 1974.

ROOT, R.; STRADER, T. J.; HUANG, Y. A review of machine learning in residential property valuation**. Journal of the Midwest Association for Information Systems**, v. 2023, n. 2, p. 21–42, 2023. Disponível em: https://jmwais.org/wp-content/uploads/sites/8/2023/07/V2023.I2.A2.pdf . Acesso em: 19 maio 2025.

SHEPPARD, S. **Measuring the impact of culture using hedonic analysis**. Technical report, Williamstown: Center for Creative Community Development, 2010. Disponível em: http://web.williams.edu/Economics/ArtsEcon/library/pdfs/UsingHedonicAnalysis2010.pdf

TEIXEIRA, E.; SERRA, M. O impacto da criminalidade no valor de locação de imóveis: o caso de Curitiba. **Economia e Sociedade**, v. 15, n. 1, p. 175–207, 2006.

TITA, G. E., PETRAS, T. L.; GREENBAUM, R. T. Crime and residential choice: a neighborhood level analysis of the impact of crime on housing prices. **Journal of Quantitative Criminology**, v. 22, n. 299-317, 2006.

TYRVAINEN, L.; VÄÄNÄNEN, H. The economic value of urban forest amenities: an application of the contingent valuation method. **Landscape Urban Plann**. v. 43, p. 105–118, 1998.

THOFEHRN, Ragnar. **Avaliação em massa de imóveis urbanos: para cálculo de IPTU e ITBI**. São Paulo: Pini, 2010.

TRIVELLONI, L. R.; HOCHHEIM, N. Avaliação de imóveis urbanos: uma abordagem prática. Pini, 1998.

UBERTI, Marlene Salete. **Valoração ambiental no uso do solo urbano: Aplicação do método dos valores hedônicos** – Universidade Federal de Santa Catarina (UFSC): Estudo de caso no centro de Florianópolis, 2000.

VARIAN, H. **Microeconomic Analysis**. 3. ed. New York: W.W. Norton and Co. 1992. 506p.

VILLAÇA, Flávio. **Espaço intra-urbano no Brasil**. São Paulo: Studio Nobel, 1998.

WONNACOT, R. J.; WONNACOT, T. H. Econometria. Rio de Janeiro: Livros Técnicos e Científicos, 1978.

ZHANG, Y.; ZHAO, J.; LI, X. Modeling house price prediction using ensemble learning algorithms: Evidence from urban China. **Journal of Property Research**, v. 38, n. 2, p. 162–180, 2021.

ZHANG, Y.; ZHAO, J.; LI, X. Modeling house price prediction using ensemble learning algorithms: Evidence from urban China. **Journal of Property Research**, v. 38, n. 2, p. 162–180, 2021.