**UNIVERSIDADE FEDERAL DO CEARÁ – UFC**
**FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE – FEAAC**
**PROGRAMA DE ECONOMIA PROFISSIONAL – PEP**

**ANTÔNIO AUGUSTO FERREIRA DE OLIVEIRA**

# AVALIAÇÃO EM MASSA COM MODELOS DE APRENDIZADO DE MÁQUINA APLICADOS AOS TERRENOS URBANOS DO MUNICÍPIO DE FORTALEZA

**FORTALEZA**
**2020**

**ANTÔNIO AUGUSTO FERREIRA DE OLIVEIRA**

# AVALIAÇÃO EM MASSA COM MODELOS DE APRENDIZADO DE MÁQUINA APLICADOS AOS TERRENOS URBANOS DO MUNICÍPIO DE FORTALEZA

Dissertação submetida à Coordenação do Programa de Economia Profissional – PEP, da Universidade Federal do Ceará - UFC, como requisito parcial para a obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Orientador: Prof. Dr. Andrei Gomes Simonassi

**FORTALEZA**

**2020**

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

O45a Oliveira, Antônio Augusto Ferreira de.
Avaliação em massa com modelos de aprendizado de máquina aplicados aos terrenos urbanos do município de Fortaleza / Antônio Augusto Ferreira de Oliveira. – 2020.
80 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Mestrado Profissional em Economia do Setor Público, Fortaleza, 2020.
Orientação: Prof. Dr. Andrei Gomes Simonassi.

1. Avaliação em massa. 2. Modelos de Aprendizado de Máquina. 3. Florestas Aleatórias. 4. XGBoost. 5. Município de Fortaleza. I. Título.

CDD 330

**ANTÔNIO AUGUSTO FERREIRA DE OLIVEIRA**

# AVALIAÇÃO EM MASSA COM MODELOS DE APRENDIZADO DE MÁQUINA APLICADOS AOS TERRENOS URBANOS DO MUNICÍPIO DE FORTALEZA

Dissertação submetida à Coordenação do Programa de Economia Profissional – PEP, da Universidade Federal do Ceará - UFC, como requisito parcial para a obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Aprovada em: **10 de março de 2020**.

**BANCA EXAMINADORA**

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

_____
Prof. Dr. Luiz Ivan de Melo Castelar
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Leandro de Almeida Rocco
Universidade Federal do Ceará (UFC)

## AGRADECIMENTOS

Agradeço à Deus, grato pelo precioso dom da vida.

À minha mãe, pelo imenso amor e dedicação para a minha formação.

À minha esposa, pela inestimada paciência com minha ausência durante essa dissertação e extrema dedicação com os nossos filhos.

À minha irmã Terezinha, pelas contribuições, sugestões e revisão deste trabalho.

Ao Prof. Dr. Andrei Simonassi pela ajuda e orientação.

Aos colaboradores da Secretaria de Finanças de Fortaleza do setor SEPLAN-PGVI, servidores e estagiários, que incansavelmente trabalham com toda dedicação e zelo na coleta de preços do mercado imobiliário. Sem vocês, seria impossível esse trabalho.

Agradeço em modo especial, ao amigo-irmão Sandro Bandeira, cuja parceria de estudo na temática me incentivou a aprofundar meus conhecimentos.

Por fim, dedico esse trabalho a Tarcísio Eduardo Nobre (*in memoriam*).

## RESUMO

O estudo realiza uma avaliação em massa dos terrenos do Município de Fortaleza utilizando modelos de aprendizado de máquina (*machine learning*), a partir de uma amostra com mais de 8.000 informações providas pelo observatório urbano de valores da Secretaria das Finanças de Fortaleza no período de 2015 a 2019. É realizada uma análise exploratória extensiva para a definição e escolha das variáveis explicativas dessa avaliação, tendo como variável resposta o preço unitário dos terrenos. Posteriormente, são avaliados três modelos: regressão linear múltipla, florestas aleatórias e XGBoost. Para cada um destes, verifica-se os pressupostos de aplicação, principalmente para o modelo estimado por Mínimos Quadrados Ordinários, dada sua dificuldade de atendimento de todos os seus pressupostos na avaliação em massa de imóveis de toda uma municipalidade. Este modelo, com o preço unitário na escala logaritmo natural, apresenta as estimativas dos seus coeficientes condizentes com o esperado na prática e observado na análise exploratória prévia. Para os modelos de aprendizado de máquina, florestas aleatórias e XGBoost, são equacionados a relação entre viés-variância, poder de generalização preditiva e o sobreajustamento. O conjunto de atributos mais importantes para explicação do comportamento dos preços unitários dos terrenos obtido com ambos são muito similares. O modelo XGBoost apresentou o melhor desempenho em todos as métricas avaliadas. Ao final, apresenta-se uma proposição de planta genérica de valores (PGV) para todas as parcelas territoriais georrefenciadas do Município de Fortaleza.



**Palavras-chave**: Avaliação em massa. Modelos de Aprendizado de Máquina. Florestas Aleatórias. XGBoost. Município de Fortaleza.

## ABSTRACT

This study concerns the mass appraisal of the market value of land in the city of Fortaleza. It is made by using machine learning models, from a sample with more than 8 thousand observations collected through an urban observatory of market values in the period from 2015 to 2019. An extensive exploratory analysis is carried out for the definition and choice of the explanatory variables of this evaluation, having as response variable the unit price of land. Subsequently, ordinary least squares regression is studied as a preliminary model to be outperformed by machine learning models, random forests and XGBoost. For each of these, the assumptions are assessed, mainly for the ordinary least squares model, due to its difficulty in meeting all its premises in real estate mass appraisal. The estimates of this model, with the unit price on a natural-log scale, are consistent with what is expected in practice and observed in the previous exploratory analysis. For machine learning models, random forests and XGBoost, the relationships among bias-variance trade-off, power of predictive generalization and overfitting are verified. The most important features to explain the unit prize of land are remarkably similar in both models. The XGBoost model outperforms the others in all the performance measures evaluated. At the end, a market value map is proposed for all georeferenced land parcels in Fortaleza.



**Keywords**: Mass appraisal. Machine learning models. Random Forest. XGBoost. Municipality of Fortaleza.

## LISTA DE FIGURAS

## LISTA DE GRÁFICOS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ABNT | Associação Brasileira de Normas Técnicas |
| CART | *Classification and Regression Tree* |
| CEF | Caixa Econômica Federal |
| CF | Constituição da República Federativa do Brasil de 1988 |
| CTM | Cadastro Territorial Multifinalitário |
| CTN | Código Tributário Nacional |
| CV | Coeficiente de Variação Percentual |
| FIV | Fator de Inflação da Variância |
| HABITAFOR | Fundação de Desenvolvimento Habitacional de Fortaleza |
| IDH-B | Índice de Desenvolvimento Humano por Bairro |
| IDHM | Índice de Desenvolvimento Humano Municipal |
| IPLANFOR | Instituto de Planejamento de Fortaleza |
| IPTU | Imposto Predial e Territorial Urbano |
| IAAO | *International Association of Assessing Officers* |
| ITBI | Imposto de Transmissão de Bens Imóveis *Inter Vivos* e cessões de direito real a eles relativos |
| LUOS | Lei de Uso e Ocupação do Solo do Município de Fortaleza (Lei Complementar Municipal 236/2017) |
| OCDE | Organização para a Cooperação e Desenvolvimento Econômico |
| OODC | Outorga Onerosa do Direito de Construir |
| OUC | Operação Urbana Consorciada |
| MQO | Mínimos Quadrados Ordinários |
| NBR | Norma técnica da ABNT |
| PDPFor | Plano Diretor Participativo de Fortaleza (Lei Complementar Municipal 62/2009) |
| PLHIS-FOR | Plano Local de Habitação de Interesse Social de Fortaleza. |
| PGV | Planta Genérica de Valores (ou Planta de Valores Genéricos - PVG) |
| RF | *Random Forest* (florestas aleatórias) |
| RNA | Redes Neurais Artificiais |
| SEFIN | Secretaria das Finanças do Município de Fortaleza |
| SIG | Sistemas de Informações Geográficas |
| UTM | Projeção Universal Transversa de Mercator |

**SUMÁRIO**

# 1 INTRODUÇÃO

O objetivo precípuo de uma avaliação em massa é determinar o valor de mercado para uma grande quantidade de imóveis de uma região. Quando esta é aplicada na tributação imobiliária, é comumente denominada “avaliação em massa para fins fiscais” (DE CESARE; CUNHA, 2012). As estimativas geradas por estas avaliações exigem alta precisão, de tal forma a resultar em avaliações uniformes e equânimes (idem, p. 34), promovendo a justiça fiscal na cobrança dos tributos imobiliários.

As abordagens tradicionais e científicas de avaliação de imóveis se pautam no modelo de preços hedônicos, geralmente com a utilização da regressão linear múltipla pelos métodos dos mínimos quadrados ordinários (MQO), com aplicação da econometria espacial e até mesmo com a utilização de redes neurais artificiais.

Atualmente, são emergentes na literatura internacional, embora ainda muito incipientes no país[1], as modelagens de preços de imóveis baseadas nos algoritmos de aprendizado de máquina (*machine learning*), subcampo da inteligência artificial. Estas técnicas chamam atenção pela sua superior capacidade de predição frente as abordagens clássicas. Nesse diapasão, a demonstração empírica de eficiência e eficácia de aplicação de modelos de aprendizado de máquina nas avaliações de imóveis com o fito de aproximar as bases de cálculos dos tributos patrimoniais ao real valor de mercado pode fortemente melhorar as receitas públicas municipais. É nesse sentido que se encontra a importância e utilidade desta pesquisa.

Por conseguinte, o objetivo geral dessa pesquisa é empregar os modelos de aprendizado de máquina, florestas aleatórias e XGBoost, na avaliação em massa dos terrenos urbanos do Município de Fortaleza para fins fiscais atestando a sua adequabilidade.

Como objetivos específicos, citam-se **i)** o estabelecimento de um conjunto de variáveis explicativas aptas a explicar o comportamento dos preços unitários dos terrenos; **ii)** a estimação de tais valores com aplicação dos algoritmos de florestas aleatórias e XGBoost; **iii)** aferição da performance de suas predições frente ao modelo

[1] A norma brasileira de avaliação de imóveis urbanos na sua parte 2 (ABNT NBR 14653-2, 2011) só trata da regressão paramétrica via modelos lineares e redes neurais artificiais.

de regressão linear múltipla pelo método MQO e **iv**) proposição de uma PGV para todos terrenos urbanos do Município de Fortaleza a partir do conhecimento adquirido.

Esta dissertação foi desenvolvida em seis capítulos, incluindo esse introdutório. No capítulo 2, discorre-se sobre a revisão da literatura relacionada a modelos de aprendizado de máquina aplicado à avaliação em massa de imóveis, bem como as últimas modelagens aplicadas no Município de Fortaleza.

No capítulo 3, caracteriza-se a área de estudo como sendo Município de Fortaleza. Ressalta-se sobre este os aspectos sociais, econômicos, sempre através de uma perspectiva espacial. Comenta-se sobre o cadastro territorial multifinalitário do município e resultados econômicos da tributação do IPTU. Também são descritas as principais variáveis utilizadas para modelagem de aprendizado de máquina.

O capítulo 4 inicia com as técnicas tradicionais de avaliação de imóveis baseadas no modelo de preços hedônicos com uso da regressão linear múltipla pelo MQO. Na sequência, são descritos os conceitos e a formalização dos modelos de aprendizado de máquina.

No capitulo 5, são mostrados os resultados e discussões da análise exploratória de dados, aplicação de cada um dos modelos e suas estimativas de desempenho de predição. Posteriormente, é apresentada a construção de uma PGV para todo o município com o algoritmo XGBoost. Este último tem sido o campeão em acurácia de predições nas comunidade *on line* de competições de *data science*, e, até o presente momento, desconhece-se a sua aplicação em avaliações em massa para terrenos urbanos, dando, desta forma, certo grau de ineditismo dessa pesquisa.

No capitulo 6, são feitas as conclusões e sugestões de trabalhos futuros.

# 2 REVISÃO DA LITERATURA

Como supra comentado, as técnicas clássicas de avaliação de imóveis são baseadas na teoria dos modelos hedônicos. Ela foi desenvolvida formalmente por Lancaster (1966 apud Fávero, 2007, p. 62) e depois aperfeiçoada por Rosen (1974 apud idem, p. 49). Conforme Fávero (ibidem), aquele primeiro autor inovou a teoria clássica do consumidor, ao propor que a sua função utilidade passa a ser condicionada a uma cesta de características do bem (atributos) que condicionarão sua escolha, e não o próprio bem em si, visto como um produto final. No mercado imobiliário a equação hedônica é constituída pela definição do preço observado do imóvel como função da quantidade dos diversos atributos que o compõe.

É vasta a literatura nacional e estrangeira de avaliação em massa de imóveis urbanos baseada na abordagem de preços hedônicos com modelos regressão linear múltipla, regressão espacial com e sem o uso de técnicas de geoestatística[2].

Entretanto, ainda é muito incipiente no país a aplicação de técnicas de aprendizado de máquina nas avaliações em massa de imóveis. Em âmbito internacional, merece destaque o trabalho pioneiro de Antipov e Pokryshevskaya (2012) que utilizaram *boosted trees* e *random forest* para avaliação de 2.848 apartamentos residenciais de dois quartos na cidade de São Petersburgo, Rússia. O trabalho de pesquisa concluiu pelo superior desempenho das florestas aleatórias frente outras técnicas consagradas, tais como regressão linear múltipla, redes neurais artificiais e *boosted trees*[3].

Yoo, Im e Wagner (2012) aplicaram técnicas de aprendizagem de máquina para predição de preços de 4.469 imóveis no condado de Onondaga, Nova Yorque, EEUU. Os autores obtiveram as predições por meio de modelos de preços hedônicos com regressão linear múltipla pelos mínimos quadrados ordinários (MQO) e sugeriram as florestas aleatórias para a seleção das variáveis preditoras mais importantes a serem aplicadas naquele primeiro modelo. Estes autores também utilizaram de variáveis de amenidades formadas com a utilização de raios de distâncias de 100m e

[2] Essa seção deter-se-á apenas na literatura relacionada aos modelos de regressão linear múltipla com econometria espacial e geoestatística, ao aprendizado de máquina aplicado ou últimas modelagens em massa aplicadas no Município de Fortaleza por terem pertinência imediata com o propósito dessa pesquisa.

[3] Apresentar-se-á a definição de *boosting* na seção 4.2.5.

1km (*buffers*) na tentativa de incorporar efeito espacial ao modelo de aprendizado de máquina[4]. Essas variáveis, quando utilizadas no modelo de florestas aleatórias, conseguiram reduzir a autocorrelação espacial dos resíduos em comparação com o modelo de regressão linear (ibidem, p. 305)[5].

Čeh et al. (2018) fizeram uma comparação do desempenho dos modelos das florestas aleatórias frente aos modelos de preços hedônicos para 7.407 apartamentos na cidade de Liubliana, capital da República da Eslovênia no período de 2008 a 2013. Eles se utilizaram de técnicas de SIG para obtenção das variáveis explicativas, incluindo variáveis relativas à acessibilidade (distância ao aeroporto, estação de ônibus, universidade, entrada da *highway*, áreas verdes, florestas etc.) e meio-ambiente (indicação de ruídos por proximidade à linha de trem, *highway*, dentre outros). Para evitar a multicolinearidade, utilizaram a análise de componentes principais. Os autores concluíram que as florestas aleatórias apresentam os melhores resultados.

Carranza et al. (2018) foram os primeiros a utilizarem as florestas aleatórias em avaliação em massa de uma amostra de dados de ofertas, transações, arrematações e vendas declaradas no imposto do selo (espécie de ITBI) de 283 terrenos para a cidade de Río Cuarto, província de Córdoba, Argentina. Eles também interpolaram os resíduos por krigagem ordinária para incorporar os efeitos espaciais no algoritmo de aprendizagem de máquina das florestas aleatórias e melhorar a estimativa final. Em seguida, geraram um mapa de valores do solo para um lote padrão de 10mx30m em toda a cidade. Os resultados obtidos foram compatíveis com o recomendado pelo IAAO, sendo pontuado pelos autores a simplificação da atualização dos valores de solo por essa metodologia, contribuindo para a equidade fiscal do sistema de tributação imobiliária e gestão urbana.

No que tange à literatura relacionada à avaliação em massa para o Município de Fortaleza, os trabalhos que se seguem merecem destaque.

Codes (2018) estimou três redes neurais artificiais (RNA) referentes aos anos 2014, 2015 e 2016 para a determinação dos valores de base de cálculo do ITBI (valor de mercado) das edificações verticais residenciais dos bairros Meireles,

[4] Variáveis compostas com a quantidade de amenidades distantes até o raio predefinido de cada imóvel incorporam no modelo preditivo (inclusive hedônico) a percepção de facilidades disponíveis, as quais refletem as preferencias individuais de cada comprador (YOO; IM; WAGNER, 2012, p. 297).
[5] Foi utilizado o índice Moran's I para cálculo e comparação da autocorrelação espacial.

Aldeota, Cocó e Messejana. Com treze variáveis independentes, os resultados finais das estimativas ficaram dentro da margem de 15% de tolerância admitida pela ABNT NBR 14.653.

Oliveira, Bandeira e Silva (2018) estimaram três modelos de avaliação em massa para valoração do solo no Município de Fortaleza por aprendizado de máquina baseados em árvores de decisão. A amostra utilizada era composta de 18.584 dados de terrenos, compreendendo os anos de 2009 a 2018, abrangendo fontes de dados referente a ofertas, transações e transmissões de ITBI. Verificou-se que o modelo por florestas aleatórias teve melhor desempenho do que os demais modelos estimados, quais sejam, regressão linear múltipla com ajuste de superfície de tendência (com polinômio de 3º grau), *bagging* e GBR-*gradient boosting regression*.

Isto posto, esta pesquisa difere dos trabalhos citados por empregar o algoritmo de aprendizado de máquina XGBoost, bem como testar sua suposta superior performance frente ao modelo de florestas aleatórias nas avaliações em massa de imóveis urbanos. Assim como realizado em Čeh et al. (2018), utilizou-se das coordenadas dos centroides dos terrenos como variáveis explicativas para a modelagem. Impende ainda observar que, como o XGBoost já trabalha com um melhoramento contínuo dos resíduos das predições das diversas árvores que o compõe, não foi necessário nesta pesquisa qualquer tratamento geoestatístico posterior sobre estes, tal como realizado por Carranza et al. (2018). Partiu-se do pressuposto que os erros das predições trazem consigo a autocorrelação espacial, que é ajustada em cada etapa posterior de predição em cada árvore, conforme se verá na seção 4.2.5.

# 3 EVIDÊNCIAS EMPÍRICAS

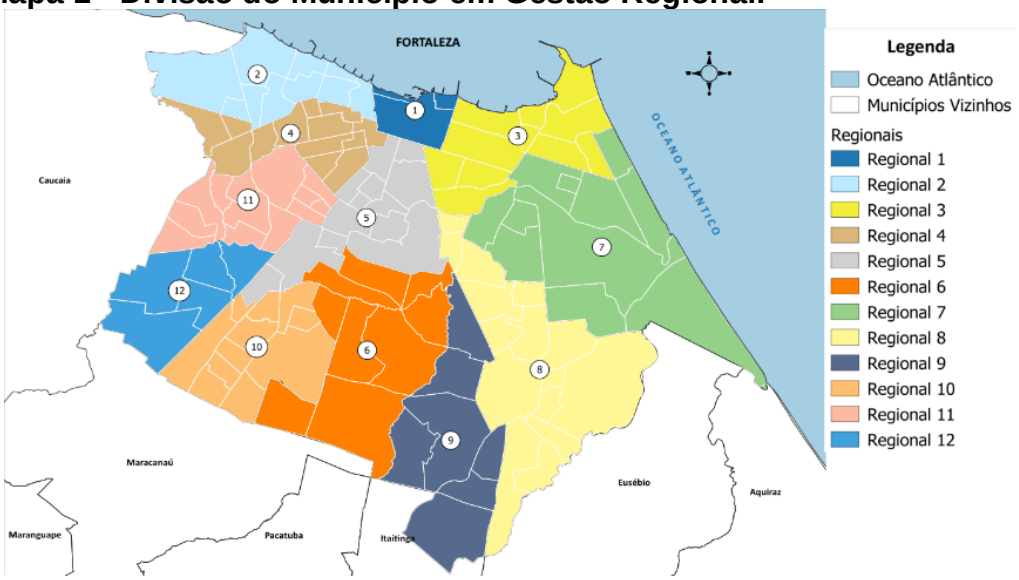
## 3.1 Município de Fortaleza: caracterização socioeconômica e espacial

O Município de Fortaleza[6], capital do estado do Ceará, tem a quinta maior população do país[7], estimada em 2.669.342 pessoas, segundo o IBGE (2019), possui IDHM de 0,754 (2010) e ocupa uma área de 312,44km². É banhada pelo Oceano Atlântico na parte norte e leste, possui 34km de litoral, tendo como limites ao norte a foz do rio Ceará e ao sul a foz do rio Pacoti. Sua condição geográfica propícia à defesa, por dividir o litoral do Ceará "em dois" (LEMENHE apud BRUNO; FARIAS, 2015, p.14), influenciou sua gênese histórica relacionada a fortalezas e fortificações. Por ter condição geográfica singular, que a torna a capital brasileira mais próxima da Europa e da América do Norte, é considerada como "principal ponto de transferência (entrada e saída) de dados de alto tráfego de informação do Brasil para os demais quatro continentes" (FORTALEZA, 2020).

A cidade é composta por 121 bairros. Por determinação da recente Lei Complementar Municipal 278/2019, a cidade foi dividida administrativamente em 12 regionais e 39 territórios (uma regional pode ser composta por até cinco territórios). Os territórios (e consequentemente as regionais) agrupam os bairros segundo critérios de "quantidade de habitantes (entre 200 mil e 300 mil por território), aproximação cultural e a utilização de equipamentos públicos" (BARROS, 2019). Essa divisão administrativa está representada no Mapa 1:

[6] Fortaleza foi fundada em 13 de abril de 1726.
[7] Atrás de São Paulo, Rio de Janeiro, Brasília e Salvador.

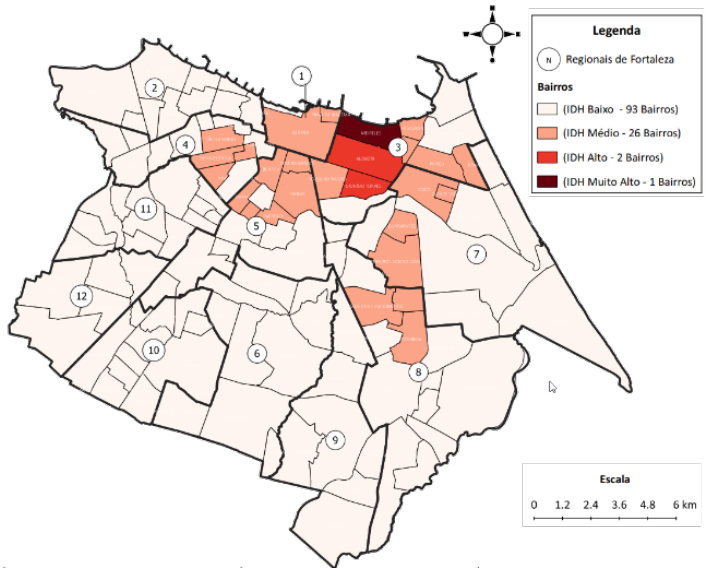
**Mapa 1 - Divisão do Município em Gestão Regional.**




Fonte: Prefeitura Municipal de Fortaleza.

A desigualdade entre os bairros da cidade pode ser refletida na disparidade do índice de desenvolvimento humano por bairro (IDH-B): o bairro Meireles possui o mais alto IDH-B com 0,953, enquanto o bairro Conjunto Palmeiras possui o menor, com IDH-B de apenas 0,119 (FORTALEZA SDE, 2014, p. 7). Insta, ainda, observar que 93 bairros têm IDH baixo (inferior a 0,50), 26 tem IDH médio (até 0,799), apenas dois alto (Aldeota e Dionísio Torres com valores até 0,899) e um muito alto (Meireles com valor acima de 0,90), conforme apresentado no Mapa 2:

**Mapa 2 - IDH dos bairros de Fortaleza.**

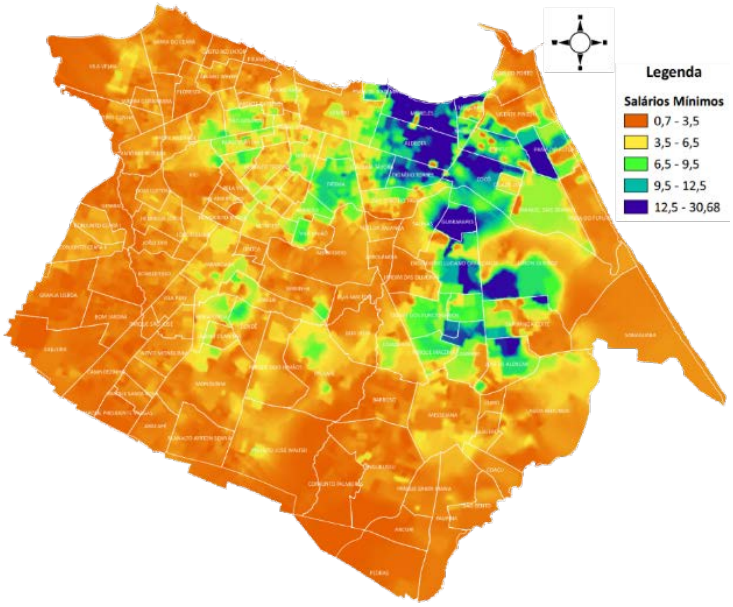



Fonte: elaboração própria a partir dos dados da SDE (FORTALEZA, 2014).

A distribuição da renda no Município de Fortaleza pode ser visualizada pelo censo do IBGE 2010, por meio da divulgação nos setores censitários da planilha “Renda da Pessoa Responsável”. Há uma conhecida correlação positiva entre renda e preços dos imóveis. Isto já foi objeto de extenso estudo por Lucena (1981), que ressalta, conforme citação de Lopes Filho (2004, p.103), “[...] há suposições de que o indivíduo pondera com bastante relevância os aspectos de vizinhança, dispondo-se a pagar mais para se localizar em um local com padrão de renda mais elevado”.

Dantas (2014, p. 8-9), a partir dos dados do último censo, elaborou uma superfície de renda em salários mínimos do rendimento médio das pessoas responsáveis no setor censitário (**Mapa 3**). Essa superfície foi obtida por técnica de geoestatística já consagradas, quais sejam, ajuste de um semivariograma teórico e interpolação dos valores renda média em toda a superfície do município por krigagem. A utilização de superfícies como fornecedoras de variáveis explicativas do tipo “*proxies* espaciais” é de grande valia nos modelos econométricos de preços hedônicos, pois estas passam a ter continuidade e suavidade numérica em toda a região de estudo, evitando-se os “degraus” de valores quando se passa de uma região para outra.

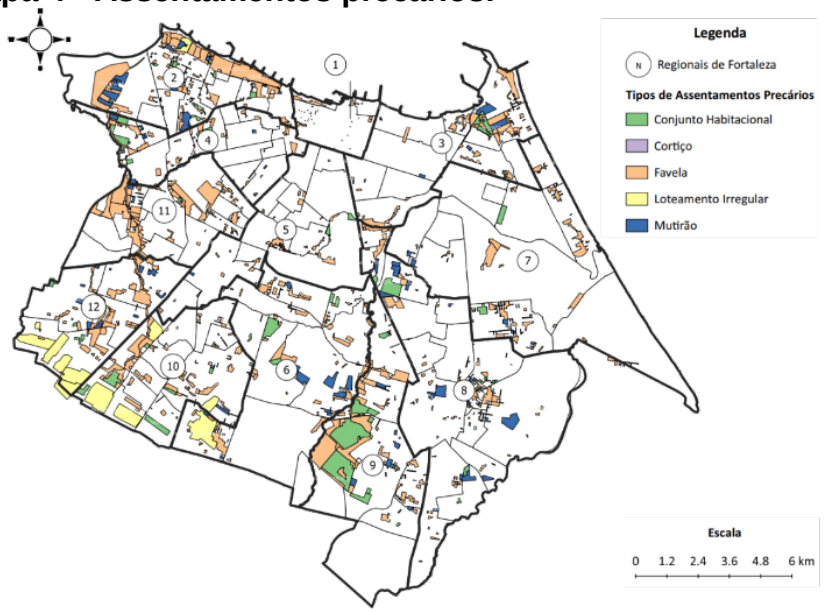
**Mapa 3 - Mapa da superfície de rendas em salários mínimos obtido por krigagem.**




Fonte: Dantas (2014, p. 8).

Segundo levantamento realizado pela HABITAFOR (FORTALEZA, 2010), o município tem 842 assentamentos precários, dos quais 621 são favelas[8]. A área total dos assentamentos precários importa em 39,83km², aproximadamente 12,75% da área total do município.

**Mapa 4 - Assentamentos precários.**




Fonte: PLHIS-FOR. (FORTALEZA, 2010).

Como se observada do mapa acima, todas as regionais têm assentamentos precários. Utilizando-se técnicas de geoprocessamento, verifica-se que as regionais 2 e 12 tem o maior percentual de suas áreas ocupadas por assentamentos precários, nas proporções de 29% e 23%. As regionais 1 e 5 tem os menores percentuais, ambas com 3%. Em relação aos bairros, os que apresentam assentamentos precários em maior número são Pirambu (89%), Cristo Redentor e Genibaú (ambos com 64%), Parque São José (61%), Conjunto Palmeiras (52%) e Canindezinho (51%). Todos demais têm menos que 50%. Somente 15 bairros não os têm: Benfica, Cidade 2000, Cocó, Conjunto Ceará I, Conjunto Ceará II, De Lourdes, Guararapes, Jardim Guanabara, Pan Americano, Parque Araxá, Parquelândia, Parreão, Pedras, Varjota e Vila Ellery.

Todas as desigualdades apresentadas importam num enorme contingente de pessoas vivendo em moradias inadequadas, demandando por políticas públicas de aluguel social, construção de moradias, desapropriações, execução de programas

[8] 621 favelas, 128 mutirões, 48 conjuntos habitacionais, 29 cortiços e 16 loteamentos irregulares.

de remoção de áreas de riscos e de regularização fundiária, dentre outros. Sob outra ótica, impõem ainda ao Poder Público um controle mais efetivo sobre o direito da propriedade urbana, exigindo que esta seja usada em “prol do bem coletivo , da segurança e do bem-estar dos cidadãos, bem como do equilíbrio ambiental”, conforme reza o Estatuto da Cidade no seu art. 1º, parágrafo único. Cai a lanço notar que esse controle deve ser por meio do conhecimento dos vazios urbanos com retenção especulativa, da subutilização da propriedade, bem como pela implantação de instrumentos tributários e financeiros de política urbana para coibir tais condutas. Todos perpassam, de alguma forma, pelo conhecimento do valor do solo urbano.

## 3.2 Cadastro territorial e tributação imobiliária no Município de Fortaleza

Dentre os vários fatores que fazem uma administração tributária imobiliária eficiente e eficaz se destacam a manutenção de uma base cadastral (de imóveis e contribuintes) continuamente atualizada e a manutenção dos valores venais de base de cálculo dos impostos próximos aos reais valores de mercado através de avaliações técnicas realizadas periodicamente.

Chama-se atenção para a correta determinação dos valores venais da parcela territorial no lançamento do IPTU, principal imposto da tributação imobiliária. Como a legislação tributária do Município de Fortaleza estabelece a base de cálculo sendo o valor venal do imóvel e composto pela soma das parcelas territorial e predial de cada inscrição imobiliária, sempre é necessário o cálculo daquela, mesmo nos imóveis prediais. Destarte, métodos mais rigorosos de apuração desse valor devem ser propostos.

De acordo com o cadastro imobiliário da SEFIN, Fortaleza possui 778.927 inscrições municipais ativas localizadas em 385.024 lotes[9] (um lote pode ter mais de uma inscrição). Apenas 9,5% dessas inscrições são territoriais, ou seja, sem parcela edificada no lote.

Em 2017, a arrecadação do IPTU como percentual do PIB correspondeu a 0,69%[10], o que se assemelha ao percentual dos países do leste europeu, mas abaixo

[9] O conceito de lote utilizado é meramente fiscal, não se confundindo com o conceito urbanístico de modalidade de parcelamento do solo.
[10] O PIB do Município de Fortaleza em 2017 foi de R$ 61.579.403,17 mil e a arrecadação de IPTU naquele ano foi de R$ 427.275.398,04.

de 1,15%, referente ao alcançado pelos 32 países da OCDE (BID, 2013; BAHL; MARTÍNEZ-VÁZQUEZ, 2008 apud BONET; MUÑOZ; MANNHEIM, 2016, p. IX).

Em 2018, a arrecadação do IPTU correspondeu a R$ 510,69 milhões, o que correspondeu a um IPTU *per capita* de R$ 193,20 (FRENTE NACIONAL DE PREFEITOS, 2020). O valor arrecadado deixou Fortaleza na décima posição entre os municípios do Brasil e na segunda posição entre as capitais nordestinas, atrás apenas de Salvador (ibidem). Entretanto, ao se comparar a arrecadação *per* capita entre as capitais do Nordeste, Fortaleza cai para quarta posição atrás de Aracaju, Salvador e Recife[11][12].

Em 2019, o arrecadado do IPTU foi de 501,46 milhões[13], o que corresponde a 22,96% da arrecadação própria[14] (SEFIN). Para o lançamento do IPTU no exercício de 2020, 615.870 inscrições foram efetivamente tributadas (79,07%)[15], importando num valor total de R$ 735,19 milhões (ibidem). Ao somar-se todos os valores venais de base de cálculo desse exercício, tem-se um valor global para os imóveis do município no valor de R$ 85,42 bilhões. Dividindo-se este valor pelo valor total lançado em 2020, chega-se a uma alíquota efetiva de tributação correspondente a 0,86%[16]. O valor médio de IPTU lançado por inscrição foi de R$ 943,85 em 2020 (ibidem).

## 3.3 Base de dados

As fontes principais de dados para compor a amostra de 8.209 dados[17] representativos dos preços praticados no mercado imobiliário de terrenos em Fortaleza foram: **i)** anúncios de ofertas de imóveis de imobiliárias, "portais de imóveis" na internet, anúncios de classificados de imóveis, informações colhidas de corretores e imobiliárias por meio de ligação telefônica etc., **ii)** informação dada pelo adquirente

[11] O Município de Aracaju foi o pioneiro no Brasil a se utilizar de uma PGV elaborada por meio de modelos econométricos espaciais para aplicação no IPTU e ITBI no ano de 2015.
[12] Segundo a Frente Nacional de Prefeitos (2020, p. 106/108), Fortaleza teve uma arrecadação de ITBI de R$ 128,43 milhões, a segunda maior arrecadação entre as capitais do NE, atrás apenas de Salvador (R$ 164,88 milhões), ocupando a nona posição entre os municípios brasileiros.
[13] Neste valor não está incluso valores de dívida ativa, por isso ser menor que o valor de 2018. Se utilizarmos o mesmo critério de 2019, o valor arrecadado nominal no exercício de 2018 foi de 461,52 milhões.
[14] A arrecadação própria em 2019 foi de 2.183,78 milhões. Se considerarmos somente a arrecadação própria de impostos a participação do IPTU foi de 33,33% e a do ITBI de 8,17%.
[15] Nesse número foram considerados todos os valores devidos de IPTU no exercício de 2020, incluindo os valores das isenções parciais.
[16] As alíquotas legais de IPTU variam de 0,6% a 2,0%.
[17] Considerando o número de lotes da cidade (374.095), essa amostra corresponde a 2,19% do total.

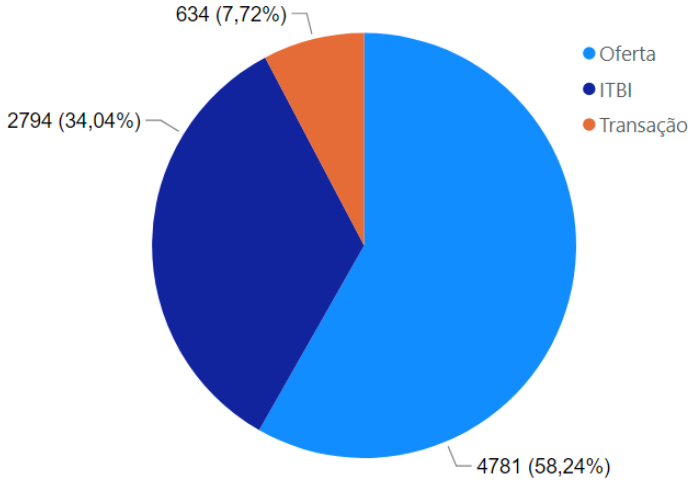
na guia de ITBI da SEFIN, desde que esta guardasse pertinência com o valor de mercado e **iii)** valores de avaliação de ITBI realizada por auditores no lançamento do imposto[18]. Os dados compreenderam o período de 01/01/2015 a 31/12/2019, 5 (cinco) anos.

Esses dados estavam organizados numa base de dados denominada observatório urbano de valores (OUV) e mantida por servidores da SEFIN. Um OUV tem como objetivos “capturar, armazenar e entregar informação econômica predial referenciada cartograficamente” e “servir de apoio aos estudos sobre valorização econômica da propriedade e o comportamento do mercado imobiliário” (ERBA, 2008, p. 141). Pazolini (2019, p. 23) define OUV como “um sistema de informação destinado a coleta periódica e sistemática dos dados do mercado imobiliário, capaz de disponibilizar informações sobre o comportamento do mercado imobiliário”. No âmbito municipal, as observações do OUV devem ser necessariamente relacionadas com o cadastro imobiliário no nível mínimo da parcela territorial. Assim, para cada observação inserida no banco de dados, deve-se determinar sua parcela territorial correspondente, que é a menor a menor unidade cadastral do CTM. Desta feita, o dado passa a ter sua localização geográfica precisa (georreferenciamento), permitindo se realizar cruzamentos espaciais com outras camadas temáticas do CTM e a criação de novas variáveis explicativas não disponíveis na informação colhida, enriquecendo de sobremaneira o poder de explicação dos preços observados na futura modelagem. Nesses cruzamentos de informações é indispensável a utilização de técnicas de geoprocessamento e o que o banco de dados do OUV seja um banco espacial. Na seção 3.4 verificar-se-á a relação de variáveis utilizadas nos modelos, onde muitas delas foram obtidas com as técnicas ora descritas e com auxílio do banco de dados *open source* PostgreSQL e sua extensão espacial PostGIS.

O Gráfico 1 mostra a distribuição dos dados por fonte de informação:

[18] No Município de Fortaleza, toda declaração de ITBI passa por um processo de avaliação interno para aferir a realidade do preço declarado pelo contribuinte.

**Gráfico 1 - Distribuição dos 8.209 dados de terrenos por fonte da informação.**



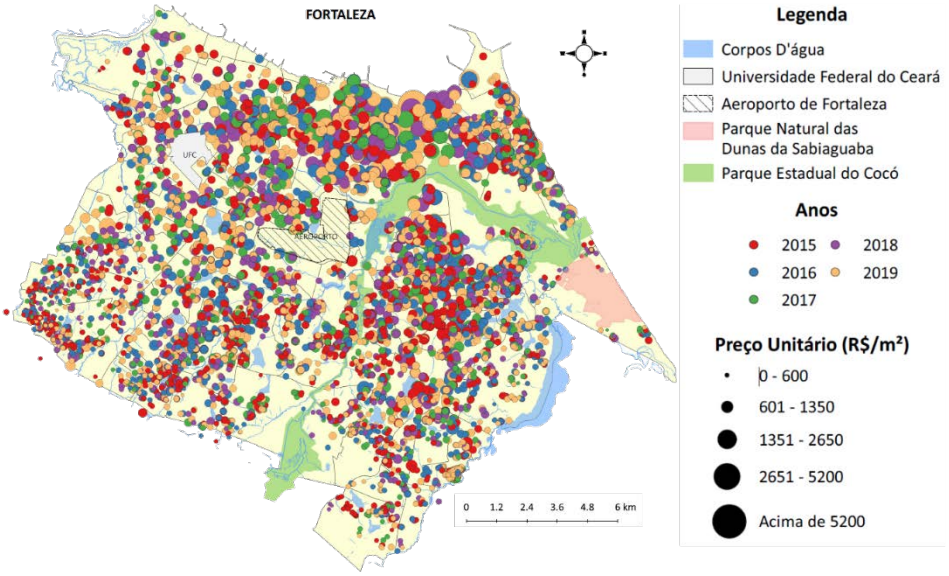

Fonte: elaboração própria a partir dos dados da pesquisa.

Considerou-se como representativos de mercado, na modalidade de transação, os preços declarados pelo contribuinte em suas guias de ITBI, que ficaram dentro da margem de mais ou menos 15% do avaliado pela Administração Tributária[19]. Também estão nessas categorias os preços informados pelas pessoas quando inquiridas sobre imóvel anunciado à venda e que já foram efetivamente transacionados. Como se observa do Gráfico 1, as maiores fonte de informações foram os anúncios de ofertas, seguidos pelas declarações de ITBI e, por último, os dados de transação.

O Mapa 5 mostra como se distribuem espacialmente os dados colhidos no município. Observa-se que os “espaços” de ausência de informação se referem às seguintes a elementos naturais e equipamentos públicos: foz do Rio Ceará no bairro Vila Velha, Parque Estadual do Cocó, Parque Natural das Dunas da Sabiaguaba (ZPA3), cais do Porto do Mucuripe e lagoas. Os seguintes bairros e regiões também tiveram poucas informações: Pirambu, Colônia, Cristo Redentor e grandes áreas no sul da cidade. As regionais 1 e 2 foram as que tiveram o menor número de dados, com 135 e 119 observações de terrenos, respectivamente.

[19] Não existe um critério técnico-científico para estabelecer esse limite de 15%. Sabe-se que as guias de ITBI contêm muitos valores subdeclarados. Como todo lançamento de ITBI precede de uma avaliação administrativa realizada pela SEFIN, esta foi o parâmetro sobre o qual se compararia o declarado pelo contribuinte, podendo este valor ser considerado como um valor de “transação efetiva” desde que estivesse nos intervalos mencionados.

**Mapa 5 - Distribuição espacial dos terrenos da amostra.**




Fonte: elaboração própria a partir dos dados da pesquisa.
Nota: amostra antes da divisão entre dados de treinamento e de teste.

Feita essas ressalvas, considera-se atendido o requisito de representatividade espacial da amostra.

## 3.4 Variáveis utilizadas no modelo

Apresenta-se a seguir uma descrição de todas as variáveis relevantes para a definição dos modelos de avaliação em massa de terrenos urbanos. Muitas delas são variáveis comumente usadas nos modelos clássicos de preços hedônicos em diversos trabalhos já publicados no campo da engenharia de avaliações[20]. Outras, em certa medida, são novidades e só foram possíveis de serem obtidas por técnicas de geoprocessamento, as quais permitiram o cruzamento espacial entre diversas bases de dados do cadastro territorial multifinalitário do Município de Fortaleza. A ideia foi ter um conjunto de variáveis que incorporem o efeito de localização, funcionando como espécies de "*proxies* espaciais", mas, sempre que possíveis, suavizadas com alguma técnica de geoestatística (superfícies interpoladas ou de densidade *kernel*).

- **regional_nn**: variáveis *dummies* agrupadas, relacionadas à localização, indicando o pertencimento do terreno à n-ésima região administrativa do município (dentre as 12 existentes e que englobam os 121 bairros).

[20] Subcampo da engenharia aplicada à avaliação de bens.

- **loteamento_condominio**: variável *dummy* isolada que indica se o terreno se encontra em condomínio fechado, quando assume o valor 1 (um)[21]. Terrenos localizados nestes têm maior valor de mercado dada às facilidades do próprio condomínio (segurança, piscina, área comum, privacidade etc.).
- **tipo_gleba**, **tipo_incorporacao** e **tipo_lote**: variáveis *dummies* agrupadas que assumem o valor 1 (um), no caso de a) terrenos com mais de 10.000m², b) mais de 750m² e até 10.000m² e c) até 750 m², respectivamente. Essas variáveis estão relacionadas à vocação do terreno a partir de sua própria dimensão. Observa-se que, pelo princípio da utilidade marginal decrescente, o valor unitário do terreno tende a diminuir com o aumento de sua dimensão. Entretanto, para terrenos propícios à incorporação de edifícios verticais e condomínios horizontais, acontece justamente o contrário.
- **zona_incorporacao_vertical**: variável *dummy* isolada que assume o valor 1 (um) no caso de o imóvel estar situado em região de alta concentração de edifícios verticais com elevador. A determinação dessas zonas na superfície do município foi obtida por mapas de densidade kernel, onde se estabeleceu uma função quártica, com raio de vizinhança de 200m e célula de grid de 25m para o *raster*[22] gerado. A partir desse mapa inicial, estabeleceu-se polígonos de delimitação dessas zonas por meio dos eixos de logradouros. Terrenos nessas zonas tendem a ter valor mais elevado.
- **zona_incorporacao_horizontal**: variável *dummy* isolada que assume o valor 1 (um) no caso de o imóvel estar situado em região de alta concentração de condomínios horizontais. A determinação dessas zonas seguiu a mesma metodologia acima. Terrenos nessas zonas tendem a ter valor mais elevado.
- **avenida**: variável *dummy* isolada que assume o valor 1 (um) no caso de o imóvel estar situado em avenida, considerando sua testada principal. Terrenos em avenida tendem a ser mais valorizados que aqueles situados em ruas, vilas etc.
- **classificação viária (via_01_expressa, via_02_arterial_1, via_03_arterial_2, via_04_coletora, via_05_paisagistica, via_06_comercial e via_07_local)**: variáveis

[21] As *dummies* foram codificadas seguindo a notação: a) 1 (um) para "com", "pertencente à categoria" etc.; 0 (zero) para "sem", "não pertencente à categoria" etc.

[22] *Raster* é o método matricial de representação dos fenômenos geográficos nas aplicações SIG. O outro método é o vetorial, que corresponde a representação por ponto, linha ou polígono. No *raster*, o espaço físico (mapa) é dividido em matrizes de células, onde cada qual tem um atributo associado (LONGLEY et al., 2013, p. 87). No caso em tela, cada célula de 25m x 25m tem a densidade kernel de probabilidade de se encontrar um condomínio vertical com elevador.

*dummies* agrupadas que indicam a classificação viária de localização do imóvel[23]. Terrenos em vias expressas, arteriais, coletoras e comerciais tendem a ser mais valorizados.

- **numero_frentes**: variável quantitativa que indica o número de frentes do terreno. A amostra apresentou terrenos com muitas frentes, mais de 10 (dez) em alguns casos, sugerindo erro de cadastramento. Por conta disso, agrupou-se todos os terrenos com mais de 5 (cinco) frentes nesse valor. O comportamento esperado do mercado imobiliário é a maior valorização de terrenos com mais frentes disponíveis.
- **esquina**: variável *dummy* isolada que assume 1 (um) no caso do imóvel ser de esquina. O comportamento esperado para essa variável é controverso, pois, para imóveis comerciais é um ponto positivo, mas para residenciais, negativo (por questões de segurança).
- **renda**: variável de macrolocalização (*proxy* espacial), representada pela renda média do chefe da família no setor censitário, segundo censo do IBGE 2010, em salários mínimos, ajustada a uma superfície de tendência construída pela krigagem ordinária (DANTAS, 2014, p. 9)[24]. Nessa interpolação foi ajustado uma variograma teórico, tendo como resultado da interpolação um *raster* de valores. Com auxílio de consultas espaciais disponíveis no PostGIS, recuperou-se o valor de renda correspondente ao centroide de cada terreno da amostra. Essa variável tem correlação positiva forte com os valores observados, conforme já comentado na seção 3.1.
- **testada**: variável quantitativa que indica o comprimento da frente do imóvel em metros. Usualmente, imóveis comerciais com grande testada têm maior valor de mercado, dada a maior visibilidade para exposição de produtos e facilidade de acesso. Essa variável é bastante correlacionada com a área do terreno, o que causa problemas de multicolinearidade em modelos por regressão linear múltipla.
- **profundidade_equivalente**: variável quantitativa que representa uma profundidade "fictícia", em metros, resultante da divisão entre a área do terreno e o comprimento de sua testada. Para terrenos de mesma área, os de maior profundidade tendem a ser mais desvalorizados.

---

[23] Essa classificação viária provém da LUOS e é determinada pela sua função, caixa carroçável e capacidade de fluxo de veículos no sistema viário urbano.

[24] O autor ajustou um semivariograma esférico, isotrópico, com alcance de 1.185m, efeito pepita de 0,22646, patamar parcial (patamar menos o efeito pepita) de 2,0925 e passo de 100m no total de 12.

- **area_terreno**: variável quantitativa correspondendo a área do terreno em m². Seu comportamento com a variável dependente "valor unitário" é, em regra, negativo, ou seja, terrenos com grande área tendem a ter o valor por m² menor. Isso devido, conforme mencionamos acima, ao princípio da utilidade marginal decrescente. Entretanto, em áreas de incorporação, para terrenos até 10.000m² (área de uma quadra em zona adensada), evidências nos mostram efeito inverso.
- **percentual_area_preservacao**: variável quantitativa que indica o percentual de área do terreno com limites administrativos de preservação ambiental permanente dos recursos hídricos (zona ZPA1), segundo o PDPFor.
- **agua, esgoto, galeria_pluvial, sarjeta, pavimentacao_asfalto_concreto, iluminacao_publica**: *dummies* isoladas de indicação do respectivo serviço público disponível para o terreno, onde nesse caso assumem o valor 1 (um). Todas têm efeito de valorizar o terreno.
- **indice_aproveitamento_basico_equivalente**: variável quantitativa que representa o índice de aproveitamento básico do terreno. Este é um parâmetro urbanístico determinado pelo PDPFor e LUOS que estabeleceram um índice único para cada zona. Este índice representa o potencial construtivo do terreno. Desta feita, se o terreno estiver em uma zona de índice igual a 3 (três), significa que, de maneira simplificada, pode-se construir uma área edificada equivalente a 3 (três) vezes a sua área territorial. Assim, terrenos com altos índices têm maior valor de mercado, devido a possibilidade de se incorporar mais edificação ao solo. Deve-se observar que terrenos de grandes dimensões podem estar contidos em mais de uma zona, cada qual com valores diferentes desse parâmetro urbanístico. Portanto, essa variável foi ponderada pelas áreas de interseção com as respectivas zonas, por meio de técnicas de geoprocessamento.
- **indice_aproveitamento_maximo_equivalente**: variável quantitativa que representa o índice de aproveitamento máximo do terreno, representado pela soma do índice de aproveitamento básico e as áreas de construção acrescidas a partir da transferência do direito de construir e/ou da outorga onerosa[25]. Também foi ponderado pelas as áreas de interseção com as várias zonas que porventura existam no terreno.

[25] Outorga Onerosa do Direito de Construir (OODC), segundo o PDPFor e LUOS, é a autorização que o município dá para construção acima do índice de aproveitamento básico até o índice de aproveitamento máximo, mediante o pagamento de contrapartida financeira (preço público) pelo beneficiário.

- **influencia_distancia_beiramar**: variável quantitativa que é o cálculo do inverso da distância do terreno (considerada pelo seu centroide) até à Av. Beira Mar (considerada pela geometria de sua extensão). Essa variável foi calculada apenas para os terrenos nos bairros Aldeota, Meireles, Mucuripe, Praia de Iracema e Cais do Porto. Para os demais bairros, foi arbitrada como zero. Espera-se que terrenos mais próximos dessa avenida sejam mais valorizados.
- **densidade_comercializacao_trecho**: variável quantitativa de densidade de comercialização no trecho de logradouro onde está situado o imóvel. Representa o percentual de imóveis comerciais em relação ao total de imóveis naquele trecho[26]. Essa variável é uma proposta de substituição a uma variável *dummy* comumente chamada "vocação comercial" usada nos modelos clássicos de preços hedônicos. Espera-se que terrenos localizados em trechos de densidade elevada sejam mais valorizados.
- **densidade_verticalizacao_kernel_200**: variável *proxy* espacial que representa a densidade de probabilidade de se encontrar um terreno com condomínio vertical possuindo elevador. A determinação desta probabilidade na superfície do município foi realizada a partir de mapas de densidade kernel, onde se estabeleceu uma função quártica, com raio de vizinhança de 200m e célula de grid de 25m para o *raster* gerado. Quanto maior a densidade, maior o valor unitário do terreno esperado, pois há indicativo de que ele está em zona onde os terrenos são propícios à incorporação de edifícios verticais. Essa variável traduz de maneira contínua a variável *dummy* **zona_incorporacao_vertical.**
- **distancia_via_principal**: variável quantitativa que indica a distância em metros à via principal mais próxima. As vias mais importantes do município foram determinadas pelos critérios de maior lançamento de IPTU, classificação viária, fluxo de veículos e pela experiência do autor. Foram selecionadas 260 vias principais no município. A partir das técnicas de geoprocessamento, calculou-se a distância do centroide de cada lote da amostra a todas às vias principais. Em seguida, selecionou-se apenas a distância correspondente à via mais próxima.

[26] Deve-se observar que essa variável considera a totalidade de imóveis que tenham testadas para o mesmo trecho, considerando-se todas as faces de quadras que tenham frente para aquele trecho de logradouro.

- **valor_m2_terreno_face_quadra_iptu_2014**: variável *proxy* indicando o valor unitário (R$/m²) base do terreno para o lançamento do IPTU, referente ao ano 2014[27].
- **assentamento_precario_area_percentual**: variável quantitativa que representa o percentual da área do terreno abrangida por assentamento precário, representado pelo conjunto de assentamentos urbanos inadequados ocupados por moradores de baixa renda, conforme Mapa 4 (PLHIS-FOR, FORTALEZA, 2010).
- **interacao_incorporacao_vertical:** é uma variável do tipo "fator interação" (DANTAS, 1998, p. 167), com objetivo de se diferenciar o efeito da área do terreno para as zonas de incorporação, com alta concentração de edifícios verticais e alto índice de aproveitamento. Ela é formada pelo produto das variáveis "zona_incorporacao_vertical", "area_terreno", "indice_aproveitamento_maximo_equivalente" e densidadeverticalizacaokernel_200.
- **interacao_incorporacao_horizontal**: também é variável de interação com a anterior. A única diferença é que se utilizou da variável "zona_incorporacao_horizontal" no lugar da variável "zona_incorporacao_vertical".
- **idhm_2010_bairro**: índice de desenvolvimento humano por bairro (IDH-B), segundo Fortaleza, SDE (2014), onde se situa o terreno.
- **idh_educ**: variável *proxy* do índice de desenvolvimento humano, dimensão educação, segundo Fortaleza, SDE (2014), referente ao bairro de localização do terreno.
- **ano_2015, ano_2016, ano_2017, ano_2018 e ano_2019**: variáveis *dummies* que indicam o ano em que a informação de preço foi obtida.
- **origem_itbi, origem_oferta e origem_transacao**: variáveis *dummies* agrupadas que indicam a origem da informação do preço coletado, conforme definido na seção 3.3.
- **preco_unitario:** variável dependente indicando o preço unitário observado (R$/m²).

Destaca-se que a relevância das variáveis citadas acima não implica na imediata inclusão no modelo de avaliação MQO, haja vista possível multicolinearidade entre várias delas.

[27] Esse é um valor unitário base antes de se aplicarem os percentuais de aumento linear introduzidos pelas Leis Complementares Municipais 73/2009 e 155/2013. Portanto, é único para todos os imóveis que estão na mesma face de quadra. No caso de imóveis com frentes para mais de uma face de quadra (e, portanto, a mais de um logradouro), considerou-se o valor correspondente à face de quadra mais valorizada.

# 4 ASPECTOS METODOLÓGICOS

## 4.1 Regressão linear múltipla

O modelo de regressão linear múltipla clássica pelos mínimos quadrados ordinários (MQO) para preços hedônicos, na sua forma matricial, pode ser definido como:

$$\boldsymbol{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \qquad (4.1)$$

onde:

$$\boldsymbol{Y}=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}_{(n\ x\ 1)}, \mathbf{X}=\begin{bmatrix}1 & X_{12} & \cdots & X_{1k}\\1 & X_{22} & \cdots & X_{2k}\\\vdots & \vdots & \ddots & \vdots\\1 & X_{n2} & \cdots & X_{nk}\end{bmatrix}_{(n\ x\ k+1)}, \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{bmatrix}_{(k+1\ x\ 1)} \text{ e } \boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}_{(n\ x\ 1)}$$

sendo $\boldsymbol{y}$ é o vetor das variáveis dependentes, **X** é a matriz das variáveis independentes (regressores exógenos não estocásticos), $\boldsymbol{\beta}$ é o vetor de coeficientes da regressão (ou parâmetros) a serem estimados e $\boldsymbol{\varepsilon}$ o vetor de resíduos estocásticos (também conhecido como erro aleatório não observável ou termo de perturbação).

Para a correta aplicação do modelo MQO, se exige os seguintes pressupostos derivados das hipóteses de Gauss-Markov[28] (WOOLDRIDGE, 2016, p. 87 e ss.):

1) o modelo é linear nos parâmetros representados pelo vetor $\beta$ e $\varepsilon$ é um erro aleatório não observável; esse pressuposto representa a própria forma da equação *(4.1)* (linearidade dos parâmetros);
2) a amostragem é aleatória e representa a população da hipótese anterior (amostragem aleatória);
3) na amostra e na população nenhuma das variáveis independentes é constante e não há colinearidade perfeita entre elas (colinearidade imperfeita ou não multicolinearidade);
4) o erro aleatório não observável $\boldsymbol{\varepsilon}$ tem valor esperado zero dados quaisquer valores das variáveis independentes:

$$\mathrm{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}] = 0 \qquad (4.2)$$

[28] O atendimento dos pressupostos garante que o estimador MQO "é o melhor, no sentido de eficiência, entre todos os estimadores não enviesados" (ALMEIDA, 2012, p. 19).

5) o erro aleatório não observável $\boldsymbol{\varepsilon}$ tem a mesma variância $\sigma^2$, dados quaisquer valores das variáveis explicativas (os erros são independentes e não heterocedásticos, ou seja, homocedásticos):

$$\text{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \sigma^2 \qquad (4.3)$$

ou ainda,

$$\text{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}} \mid \mathbf{X}] = \sigma^2\mathbf{I} \qquad (4.4)$$

6) os erros não são correlacionados (inclusive espacialmente autocorrelacionados) (ANSELIN; REY, 2014, p.97):

$$\text{E}[\ \boldsymbol{\varepsilon_i}\boldsymbol{\varepsilon_j}\ ] = 0 \qquad (4.5)$$

7) os erros não dependerem das variáveis explicativas (condição de exogeneidade) (ARBIA, 2014 p. 2 e ANSELIN; REY, 2014, p.97):

$$\text{E}[\ \boldsymbol{x_i}\boldsymbol{\varepsilon_j}\ ] = 0 \qquad (4.6)$$

Sob as hipóteses de Gauss-Markov (até a hipótese 4), o estimador de $\boldsymbol{\beta}$ pelo MQO é o **m**elhor **e**stimador **l**inear **n**ão **t**endencioso (MELNT, ou *BLUE - Best Linear Unbiased Estimator*) (ALMEIDA, 2012, p. 16).

Florêncio (2010, p. 4) pontua que o modelo normal de regressão linear clássico tem sido muito utilizado no mercado imobiliário, mas na maioria das vezes, "o pesquisador não toma os cuidados necessários na modelagem em relação aos pressupostos básicos". Ressalta ainda a dificuldade de "aplicação de metodologias econométricas que resultem em modelos simultaneamente parcimoniosos, abrangentes e fidedignos ao mercado" (idem, p. 84). Almeida (2012, p. 20) comenta que "se o pesquisador estimar sua regressão e obtiver resultados que atendam a todos estes pressupostos, ele pode ser considerado uma pessoa de muita sorte".

De todo o exposto, vê-se que os pressupostos do modelo de regressão clássico de regressão linear são muitos e de difícil alcance, principalmente na utilização de modelos de avaliação em massa de imóveis, onde, na maioria das vezes, é necessária a utilização de muitas variáveis explicativas. Como se observou no capítulo 2, a literatura internacional e nacional já propõe novos modelos baseados em aprendizado de máquina para a superação de tais limites. É o que se verá na próxima seção.

## 4.2 Modelos de aprendizado de máquina

Aprendizado de máquina (*machine learning*) é uma subárea da ciência de dados (*data science*[29]), que por sua vez faz parte da inteligência artificial. Tem como objetivo o desenvolvimento de técnicas computacionais capazes de adquirir conhecimento baseados em experiências adquiridas a partir de exemplos e/ou soluções/problemas anteriores.

Os modelos de aprendizado de máquina ganharam ultimamente grande relevância na modelagem de dados com o avanço do poder de processamento de *hardware* e sofisticação dos próprios algoritmos (*software*). Hastie, Tibshirani e Friedman (2008, p. 295)[30] ressaltam a utilidade dos modelos de aprendizado de máquina na solução dos problemas da vida real e criticam o uso de modelagem tradicional, pois na maioria das vezes, as soluções para estes, a partir de modelos lineares, são inadequadas. De fato, viu-se na seção anterior, que vários são os pressupostos a serem obedecidos na regressão linear múltipla pelo método dos MQO, o que é difícil de obter em uma avaliação em massa de imóveis em escala municipal.

Os algoritmos de aprendizado de máquina apresentados nessa dissertação fazem parte da categoria de aprendizado supervisionado e podem ser definidos como o treinamento de dados onde se sabe o valor da variável resposta (*label*)[31]. O treinamento, a partir dos atributos fornecidos (variáveis dependentes), tenta estimar o valor de resposta e minimizar o erro entre a predição e o valor observado. Na categoria de aprendizado supervisionado, podemos citar os seguintes algoritmos: vizinhos KNN, regressão linear, regressão logística, máquinas de vetores de suporte (SVM[32]), redes neurais artificiais, árvores de decisão, florestas aleatórias, XGBoost, dentre muitos outros.

---

[29] Provost e Fawcett (2016, p.4) definem *data science* como o envolvimento de "princípios, processos e técnicas para compreender fenômenos por meio da análise (automatizada) de dados". E ainda, "envolvem extrair conhecimento a partir dos dados ou tomar decisões orientadas a eles" (idem, p. 8).

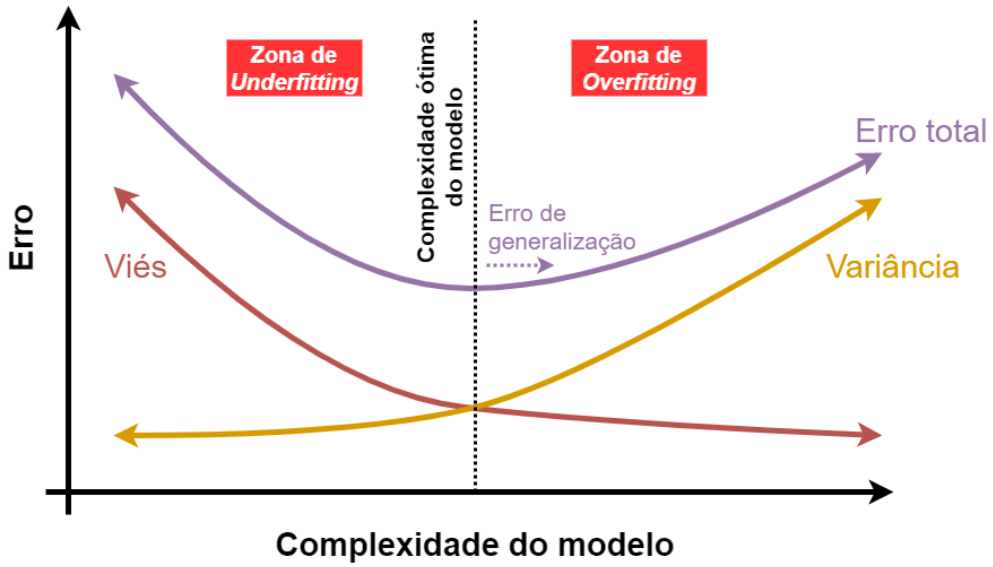
[30] "*Although attractively simple, the traditional linear model often fails in these situations: in real life, effects are often not linear*" (HASTIE; TIBSHIRANI; FRIEDMAN, 2008, p. 295).

[31] Há ainda o aprendizado não supervisionado (*v.g.*, K-Means, análise de componentes principais, DBSCAN etc.), aprendizado semi-supervisionado, e aprendizado por reforço. Maiores detalhes podem ser vistos em Géron (2019, p. 7-14).

[32] Do inglês: *support vector machine*.

Ao se trabalhar com aprendizado de máquina, o pesquisador se depara com o dilema viés-variância (*bias-variance trade-off*), que pode ser melhor explicado através da Figura 1:

**Figura 1 - Dilema viés-variância (*bias-variance trade-off*).**




Fonte: elaboração própria a partir de Goodfellow, Bengio e Couville (2016, p. 127).

Pela figura acima, o dilema se apresenta na forma de melhorar o modelo aumentando sua complexidade, diminuindo o viés, mas aumentando o erro de predição nos novos dados (aumento da variância). Este último caso exemplifica o sobreajustamento (*overfitting*) onde o modelo praticamente "fotografa" o comportamento dos dados, tendo poder de predição pífio nos dados de teste, ou seja, não se presta a generalizar ("erro de generalização")[33]. Por outro lado, a utilização de modelos de baixa complexidade acarretando erros altos nos dados de treinamento leva a situação de *underfitting*. A situação ideal é um "meio-termo" onde temos um modelo de complexidade média, com bom resultado de predição nos dados de treinamento, preservando seu poder de generalização, ou seja, apresentando baixa variância do erro nas predições dos dados de teste.

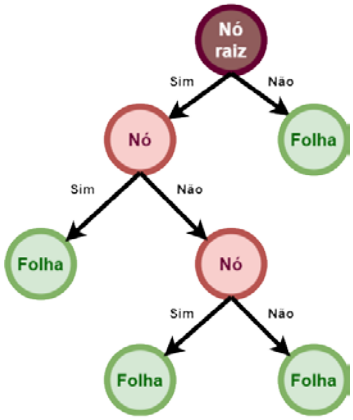
[33] Como Provost e Fawcett (2016, p. 15) advertem: "se olhar muito para um conjunto de dado, encontrará alguma coisa --, mas isto pode não ser generalizável além dos dados para os quais se está olhando".

### *4.2.1 Árvores de decisão (decision trees)*

Os algoritmos das florestas aleatórias e XGBoost trabalham com árvores de decisão, sendo necessário, portanto, nessa seção, abordar alguns conceitos fundamentais.

Um dos modelos mais simples em aprendizado de máquina é o de árvore de decisão (*decision trees*). Este consiste num método de aprendizado supervisionado, não paramétrico, que particiona um conjunto de dados em nós, ramos e folhas para predição dos valores.

**Figura 2 - Esquema de nós, ramos e folhas de uma árvore de decisão.**




Fonte: elaboração própria

Este particionamento se dá por uma estrutura condicional (se/senão) que se realiza em cada nó, comparando determinado atributo com um "valor de corte" dentro do domínio daquele [atributo]. A questão que se coloca é exatamente como determinar o melhor atributo e o melhor ponto de corte, em cada nó, para que a árvore tenha a melhor performance de predição da variável alvo.

A seguir apresentamos um pseudocódigo do método[34], baseado no algoritmo CART proposto por Breiman (1984) (GÉRON, 2019, p. 179).

1. Selecione um atributo $X_k$ e uma medida limite dentro do domínio de k ($t_k$) de tal forma que minimize a função de perda "J" assim definida[35]:

[34] Os termos "*max_depth*" e "*min_samples_split*" são os hiperparâmetros da biblioteca scikit-learn (PEDREGOSA et al., 2011).
[35] Utilizou-se o erro quadrático médio como critério. Poderia ter sido usado também o erro absoluto médio (MAE).

$$J(k, t_k) = \frac{m_{esq}}{m} MSE_{esq} + \frac{m_{dir}}{m} MSE_{dir} \qquad (4.7)$$

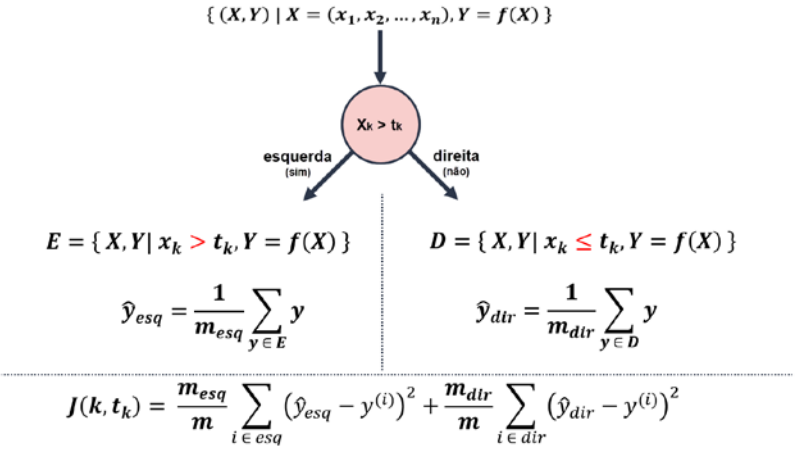
onde,

$$\begin{cases} MSE_{nó} = \sum_{i \in nó} \left(\hat{y}_{nó} - y^{(i)}\right)^2 \\ \hat{y}_{nó} = \frac{1}{m_{nó}} \sum y^{(i)} \\ \mathrm{m}_{\mathrm{esq/dir}} = \text{nº de elementos no lado esquerdo/direito da árvore abaixo do nó} \end{cases}$$

2. Divida o nó pelo melhor par ($X_k$, $t_k$) em 2 (dois) nós filhos, conforme critério acima.
3. Repita os passos 1 e 2 até que as seguintes condições sejam satisfeitas[36]:
    a. atingir a profundidade máxima da árvore (*max_depth*) e/ou
    b. atingir o número mínimo de elementos no nó requerido para continuar a divisão em nós filhos (*min_samples_split*).

O pseudocódigo acima também pode ser representado de maneira esquemática pela Figura 3:

**Figura 3 – Escolha do melhor atributo e melhor "ponto de corte" em cada nó na árvore de decisão.**




Fonte: elaboração própria a partir de Géron (2019, p. 179) e Čeh, (2018, p. 6).

[36] Existem outras opções como, por exemplo, o número mínimo de elementos para o nó ser folha (*min_samples_leaf*). Citam-se apenas os 2 (dois) principais mais usados.

A visualização de uma árvore decisão facilita a compreensão do problema, inclusive podendo ser uma forma de identificar as variáveis mais relevantes e a relação entre elas. Esta técnica é menos suscetível a *outliers* e a valores faltantes (*missing values*). Nesse último caso, é possível utilizar uma categoria nova para os dados faltantes (no caso de variáveis categóricas) ou deixar que o próprio algoritmo só se utilize de observações onde a informação é disponível (opção para o caso de variáveis contínuas onde há *missing values*) (HASTIE; TIBSHIRANI; FRIEDMAN, 2008, p. 311)[37].

Apesar dessas vantagens, a maior dificuldade da técnica está na instabilidade, uma vez que um erro na divisão superior é propagado em todas as demais divisões (alta variância). Esta também apresenta baixo poder de previsão quando a região em estudo não é retangular, pois as decisões são ortogonais, ou seja, as divisões são perpendiculares a um eixo (GÉRON, 2019, p. 185). Géron sugere, para contornar essa limitação, a utilização da técnica de análise de componentes principais (ibidem), por esta permitir uma transformação/rotação dos eixos originais.

Como bem ressalta GÉRON (2019, p. 181), as árvores de decisão devem ser controladas através de hiperparâmetros para se evitar o sobreajustamento (*overfitting*) através de técnicas denominadas de “regularização”. A regularização das árvores pode ser feita através da “poda” (*prune*), onde se limita a sua “profundidade”, ou seja, a quantidade de níveis de nós a partir do nó raiz. Pode-se também estipular um número mínimo de dados na folha (para evitar nova divisão), dentre outros.

### *4.2.2* Aprendizado *ensemble*

Aprendizado *ensemble* ou *ensemble learning* é a técnica de se agrupar um conjunto de preditores individuais fracos (*weak learners*) para se obter um preditor final (*stronger learner*) e melhorar a previsão final. É baseado no princípio da “sabedoria das multidões”, no qual o “conhecimento coletivo de um conjunto de pessoas geralmente excede o conhecimento de um simples indivíduo” (SUROWIECKI, 2004 *apud* HASTIE; TIBSHIRANI; FRIEDMAN, 2008, p. 286).

[37] Os autores sugerem estas técnicas a utilizar a eliminação de dados faltantes e a imputação de valores.

Segundo Géron (2019, p.191), o aprendizado *ensemble* tem melhores resultados quando o grupo de preditores seja o mais independente entre si e que utilizem diferentes algoritmos.

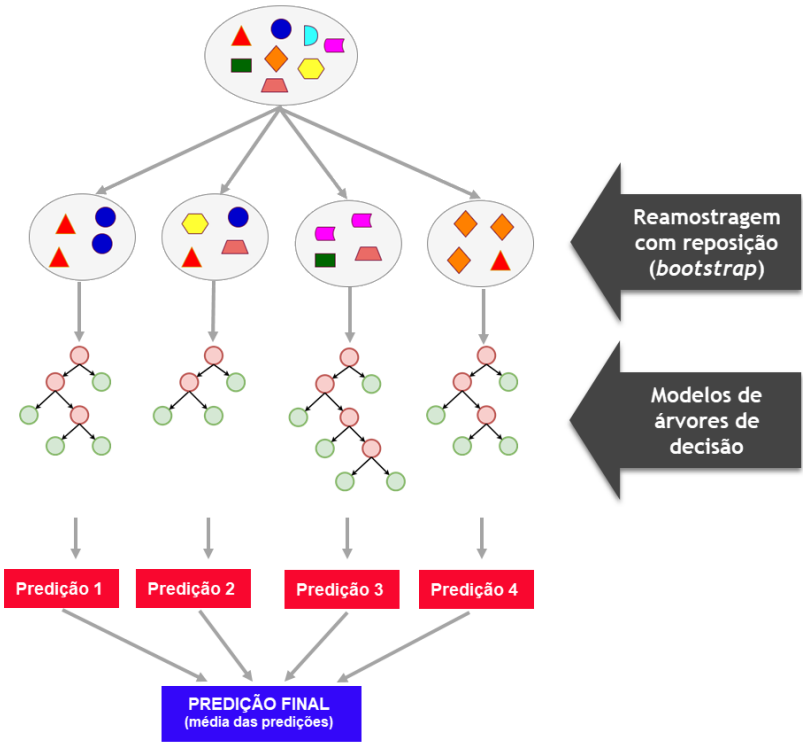
Discutir-se-á nas próximas seções os principais algoritmos de aprendizado ensemble: *bagging*, florestas aleatórias e *XGBoost*.

### *4.2.3 Bagging*

O *bagging* (***b**ootstrap **agg**regat**ing***) é um dos métodos chamados de *ensembles* porque combinam diferentes formas de árvores para aumentar o poder preditivo de decisão.

Este método gera um conjunto de dados por amostragem *bootstrap*, a partir dos dados originais, para construir uma coleção de árvores de decisão que são combinadas, ao final, em uma única predição. As árvores de decisão são treinadas de forma independente por diferentes conjuntos de treinamento.

**Figura 4 - Ilustração esquemática de bagging.**




Fonte: elaboração própria a partir de Géron (2019, p. 193).

De acordo com a Figura 4, observa-se que a seleção de dados “para pôr na sacola” (*bag*) se dá com reposição (*boostsrap*). Caso essa reamostragem seja feita sem reposição, o método passa a ser chamado de *pasting* (GÉRON, 2019, p. 193). Depois do modelo treinado, a predição de uma novo elemento se dá pela média de todos os preditores (no caso da Figura 4, através da média das predições 1,2,3 e 4).

O pseudocódigo do método é a generalização do apresentado na seção 4.2.1, conforme esquema da Figura 3.

### *4.2.4 Florestas aleatórias*

Florestas aleatórias (*random forest* - RF) também é um método *ensemble*. É uma modificação do *bagging*, visto na seção anterior, introduzido por Breiman (2001, apud LIAW; WIENER, 2002, p. 18). A principal diferença para *bagging* é que o algoritmo de florestas aleatórias utiliza somente um subconjunto dos atributos para divisão do nó da árvore. Isso introduz mais aleatoriedade no “cultivo das árvores”, aumentando-se o viés para se obter uma menor variância de predição (GÉRON, 2019, p. 197). Desta feita, a coleção de árvores cultivadas (“floresta”) passa a ter árvores não correlacionadas entre si com sensível diminuição do sobreajustamento (*overfitting*).

O pseudocódigo do método é o que se segue[38] (LIAW; WIENER, 2002, p. 18):

1. Selecione “*n_estimators*” amostras *bootstrap* dos dados originais que corresponderão ao número de árvores a serem treinadas.
2. Cultive uma floresta de árvores de decisão onde o critério de divisão do nó conforme a Figura 3, mas com a diferença de que somente um subconjunto aleatório de atributos estará disponível para o “ponto de corte ótimo” (*max_features*).
3. Para a predição de novo dado, basta seguir as regras de se-senão em todas as árvores da floresta e tirar a média das predições de cada uma delas.

[38] Os termos “*n_estimators*” e “*max_features”* utilizados no pseudocódigo são os hiperparâmetros da biblioteca sckit-learn (PEDREGOSA et al., 2011).

As florestas aleatórias também uma importante informação que pode ser utilizada na seleção de atributos para outras modelagens tradicionais, tais como a regressão linear múltipla: “importância relativa do atributo” (*feature importance*). Para esse cálculo, segundo Ronaghan (2018) e Géron (2019, p. 198), o algoritmo mede a importância pela ponderação do número de nós em que o atributo foi utilizado para diminuir o erro quadrático médio e a probabilidade de se alcançá-lo.

Uma das causas do sucesso das florestas aleatórias, além de sua alta taxa de precisão em comparação com os métodos tradicionais de análise de dados é a simplicidade de seu uso. Antipov e Pokryshevskaya (2012, p. 1.773) citaram as principais vantagens das florestas aleatórias, (além da melhor performance): **i)** trabalha bem com as variáveis categóricas, independentemente da sua quantidade, pois podem ser transformadas em *dummies*, sem que isso acarrete em *overfitting* (pelo uso de muitas delas); **ii)** trabalha bem com *missing data*, sem a necessidade de imputação; **iii)** é robusta a outliers; **iv)** é superior ao algoritmo de árvores de decisão simples, com sua predição sendo calculada para um valor não vinculado às observações de entrada; **v)** permite o adequado tratamento das relações não lineares entre a variável dependente e suas explicativas, nos mais diversos segmentos de dados; **vi)** o algoritmo não requer uma especificação detalhada e se ajusta bem às diferenças de preços observados em diversas áreas (segmentação de preços a depender da área[39]); **vii)** as predições dos preços para novas observações não geram extrapolações, ou seja, estão sempre dentro do intervalo de preços dos dados de treinamento e, por fim[40], **viii)** permite se estudar a importância de cada variável explicativa na formação dos preços (efeito marginal da variável).

Yoo, Im e Wagner (2012, p. 295) ressaltam o problema da multicolinearidade na utilização de muitas variáveis nos modelos de regressão linear. Observaram que as florestas aleatórias podem eliminar as variáveis menos importantes e mantendo um número suficiente delas para atender ao princípio da parcimônia para boas predições.

[39] Os autores observaram que os efeitos marginais de cada característica do imóvel são diferentes entre os diversos segmentos, o que deixaria controverso as predições por uma regressão linear múltipla para toda a área pesquisada.

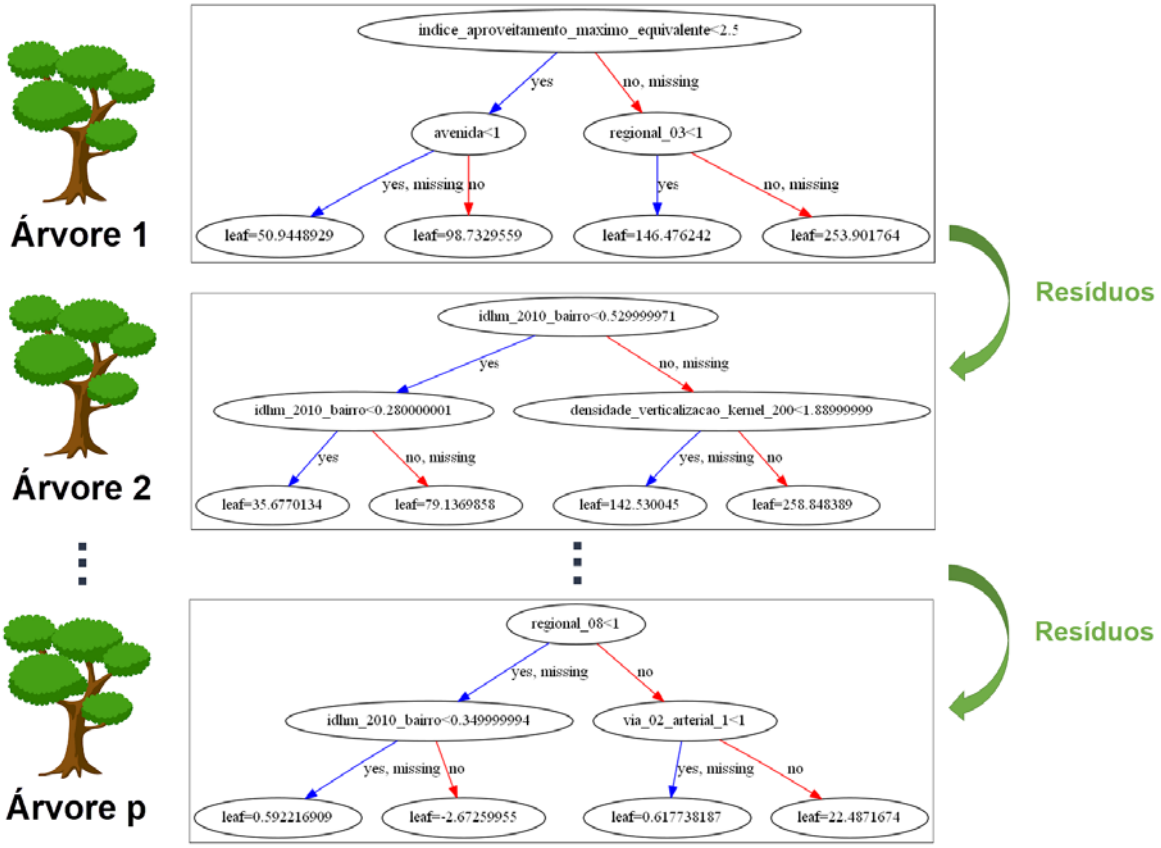
[40] Foi comentado na seção 3.3 a dificuldade de composição de amostra com elementos que representem toda a amplitude observada em cada variável na população.

### *4.2.5 XGBoost*

Antes de se comentar sobre o XGBoost, necessário se faz abordar o algoritmo de sua gênese, o *boosting*. O método *boosting* é um tipo de aprendizado *ensemble*, conforme visto na seção 4.2.2, a partir da ideia de se melhorar o poder preditivo final de um algoritmo, a partir de aglutinação de diversos preditores fracos para formar um único preditor forte. Nos modelos *boosting*, onde os preditores fracos são árvores de decisão, o procedimento consiste em utilizar diversas árvores sequencialmente cada qual tentando melhorar a performance de sua antecessora. (HASTIE; TIBSHIRANI; FRIEDMAN, 2008, p. 337; GÉRON, 2019, p. 199-200). Como afirmado as árvores utilizadas devem ser preditoras fracas, no sentido que possuam poucos nós ou seja aplicada sobre elas algum tipo de regularização, pois a ideia é ir "apreendendo aos poucos" (AMORIM, 2019, p. 70).

Como principais modelos de *boosting* citam-se o pioneiro *adaboost* e o *gradient boosting*. *Adaboost* se utiliza de um modelo de árvore de decisão sobre a base de treinamento. Realizado a primeira predição, o algoritmo realizar novo treinamento na árvore subsequente com ponderação maior (ajustes de pesos) sobre os dados de treinamento que provocaram maior erro na etapa anterior. E assim sucessivamente (GÉRON, 2019, p. 200). Já o *gradient boosting* trabalha com ajustes de pesos para os resíduos (e não os dados) da etapa anterior. A Figura 5 mostra um exemplo esquemático com a primeira, segunda e última árvore (do total de 200) para o algoritmo XGBoost realizado sobre os dados da pesquisa. Observa-se que cada árvore tem uma escolha aleatória de atributos para a divisão dos nós. Para melhor entendimento, definiu-se o nível de profundidade da árvore igual a 2. O valor no balão *leaf* (folha) representa a predição daquele ramo para a respectiva árvore.

**Figura 5 - Exemplo esquemático do *gradient boosting* com árvores de nível de profundidade igual a 2.**




Fonte: elaboração própria com simulação do algoritmo XGBoost sobre dados da pesquisa.

Segundo Brownlee (2019, p.11), o *gradient boosting* é um algoritmo “guloso”, pois o método de correção sucessivo de erros leva rapidamente ao *overfitting*, sendo necessário algumas técnicas de regularização para diminuir o erro de generalização, tais como as já apresentadas na seção 4.2.1.

O XGBoost (ou *eXtreme Gradient Boosting*) é uma biblioteca de código aberto de extensão para *gradient boosting*, desenvolvida inicialmente por Tianqi Chen e o brasileiro Carlos Guestrin. Atualmente, é mantida por diversos colaboradores da *Distributed Machine Learning Common* (DMLC). Ele procura melhorar a otimização utilizando os recursos de software e hardware. No lado do software, o método incorpora nativamente a regularização para evitar modelos muito complexos, que tendem a gerar *overfitting*. Esse problema também é controlado pela validação cruzada nativa. A técnica ainda aceita variáveis esparsas para treinamento, o que significa que ele pode trabalhar com valores ausentes (*missing values)* de forma eficiente, sem que o usuário tenha que escolher a estratégia de pré-processamento.

Do lado do hardware, o método usa paralelização e memória cache o que torna o aprendizado mais rápido (CHEN; GUESTRIN, 2016).

A escolha desse algoritmo para essa pesquisa se deveu a sua superior performance para dados estruturado e tabulados dentre vários outros algoritmos de aprendizado de máquina, inclusive sobre florestas aleatórias e redes neurais artificiais, reconhecida nas competições de cientista de dados na comunidade *kaggle*[41].

### *4.2.6 Intepretação dos modelos com o gráfico de dependência parcial*

Os modelos de aprendizado de máquina, tais como florestas aleatórias e XGBoost, não permitem avaliar diretamente como cada preditor está associado com a variável resposta, em uma equação única, tal qual se observa nos modelos econométricos, sendo muitas vezes, por isso, chamados de modelos "*black box"*. Entretanto, com o gráfico de dependência parcial é possível avaliar esta associação visualmente. Esse gráfico mostra o efeito da marginal de um preditor no valor predito pelo modelo. A partir dele, é possível investigar qual a forma e o sentido da relação entre cada preditor e a variável resposta. Estes gráficos podem mostrar se essa relação é linear, monotônica ou mais complexa. (HASTIE; TIBSHIRANI; FRIEDMAN, 2008, p. 369; AMORIM, 2019, p.7).

## 4.3 Técnicas estatísticas, medidas de desempenho e performance dos modelos

### *4.3.1 Técnicas estatísticas*

Para a análise exploratória dos dados, a ser apresentada na seção 5.1, utilizou-se de estatísticas não-paramétricas para avaliar as variáveis contínuas. Com uso das referidas técnicas, não é necessário especificar certas condições sobre os parâmetros da população da qual a amostra foi obtida, dispensando a pressuposição de normalidade dos dados, ou ainda a existência de *outliers*. Isto pois, se observou

[41] É uma comunidade na *web* de cientista de dados subsidiária da Google que promove competições de aprendizado de máquina. Pode ser acessada pelo endereço eletrônico https://www.kaggle.com/.

dados *outliers* (os quais foram verificados e validados, impossibilitando sua remoção) e grande assimetria da variável de resposta (preço unitário dos terrenos)[42], bem como em várias variáveis independentes.

Como alternativa para o teste t de *Student* para amostras independentes, foi utilizado o teste não paramétrico U de Mann-Whitney para testar se a posição das medianas são iguais. Para analisar as correlações entre as variáveis explicativas e entres estas e a variável resposta foi utilizado coeficiente de correlação de postos de Spearman ($\rho_S$). Para avaliar a associação entre as variáveis categóricas, ou seja, saber se as diferenças observadas entre as variáveis são significativas, foi aplicado o teste qui-quadrado ($\chi^2$) (PAULINO; SINGER, 2006 e SIEGEL; CASTELLAN JR., 2006). Utilizou-se também a técnica de análise de correspondência (HOFFMAN; FRANKE, 1986) para examinar as relações entre algumas variáveis nominais (regional e faixa de área de terreno) com o preço unitário do terreno. Por fim, para testar a diferença mediana entre os resíduos dos três modelos testados, utilizou-se do teste de postos sinalizados de Wilcoxon[43].

### *4.3.2 Medidas de desempenho e performance dos modelos*

Como medidas de desempenho e acurácia para comparação entre os diversos modelos, foram definidas as seguintes métricas com base nos trabalhos de Antipov e Pokryshevskaya (2012), Čeh et al. (2018) e Oliveira, Bandeira e Silva (2018), bem como nos padrões estabelecidos pelo IAAO (2010):

- **nível de avaliação** (*sales ratio*) mediano (***SR$_m$***):

$$SR_m = mediana\ de \frac{\hat{y}_i}{y_i} \qquad (4.8)$$

onde $\hat{y}_i$ é o valor predito pelo modelo e $y_i$ é o valor observado;

- **coeficiente de dispersão** (**COD**):

[42] Sabe-se que, mesmo para grandes amostras, quando as distribuições são muito assimétricas, a média é muito afetada por mudanças distantes na cauda da distribuição.
[43] Maiores detalhes sobre o teste podem ser vistos em Doane e Seward (2014, p. 689).

$$COD = \frac{100}{SR_m} \times \left( \frac{\sum_{i=1}^{n} |SR_i - SR_m|}{n} \right) \quad (4.9)$$

onde $SR_i$ é o nível de avaliação de cada terreno individualmente considerado e *n* é o número total de dados da amostra;

- **média percentual absoluta do erro (MAPE)**:

$$MAPE = \frac{100}{n} \times \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4.10)$$

- **raiz do erro quadrático médio** (**RMSE**):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (4.11)$$

Segundo o art. 30, §4º da Portaria 511/2009 do Ministério das Cidades, que normatizou as diretrizes para a criação, instituição e atualização do Cadastro Territorial Multifinalitário (CTM) nos municípios brasileiros, o nível de avaliação é calculado pela média (e não mediana) do quociente $\frac{\hat{y}_i}{y_i}$. Entretanto, preferiu-se utilizar a metodologia do IAAO, já que a mediana é menos sujeita à influência de valores extremos. O referido dispositivo normativo recomenda atualização dos valores de mercado cadastrais sempre que o nível de avaliação seja menor que 70%. A medida de verificação de uniformidade definida pela portaria se assemelha ao MAPE da equação (4.10) e é estabelecida no valor máximo de 30%.

# 5 RESULTADOS

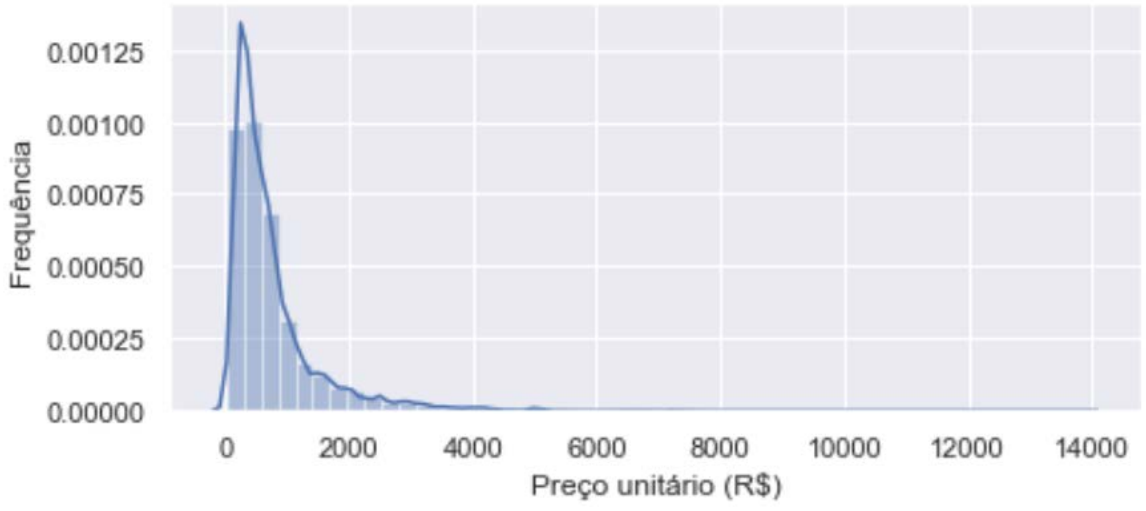
## 5.1 Análise exploratória dos dados

Nesta pesquisa, a análise exploratória consistiu na utilização de técnicas estatísticas para resumir e organizar os dados coletados, por meio de tabelas, gráficos ou medidas numéricas e teste de algumas hipóteses para se tentar procurar alguma regularidade ou padrão para a variável dependente preço unitário do imóvel em relação a algumas variáveis qualitativas regional, bairro, origem da informação e quantitativas, ano da informação e faixas de área de terreno.

### *5.1.1 Preço unitário*

A distribuição dos preços unitários dos terrenos é muito assimétrica com coeficiente de assimetria positivo igual a 3,60 como mostra o Gráfico 2**.**

**Gráfico 2 - Histograma com a densidade do preço unitário dos terrenos analisados.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

A estatística descritiva do preço unitário por regionais encontra-se na Tabela 1. Como se observa, as maiores variações ocorreram nas regionais 7 e 6 com coeficientes de variação percentual (CV%) respectivamente iguais a 87,74% e

87,55%[44]. E a menor variação dos preços na regional 1 com CV% igual a 38,37%[45]. A regional 1 apresenta tanto os preços médios como medianos mais altos e a regional 12 apresenta os menores preços médios e medianos.

Como já mencionado, para testar a hipótese de igualdade da posição das medianas, utilizou-se o teste de U de Mann-Whitney. A Tabela 1 apresenta o resultado deste teste, com as letras iguais na coluna indicando que as posições das medianas não diferem significativamente com $\alpha$=5%. Assim, são consideradas na mesma posição as medianas das regionais 2, 7, 11 e 3 e 5.

**Tabela 1 - Estatísticas descritivas do preço unitário dos terrenos por regional.**

| Regional | N. | Média | CV% | Mediana | Mínimo | Máximo |
|---|---|---|---|---|---|---|
| 1 | 135 | 2212,99 | 38,37 | 2035,62$_a$ | 808,02 | 5131,13 |
| 2 | 119 | 783,28 | 61,2 | 666,67$_b$ | 40,7 | 3129,35 |
| 3 | 567 | 1946,74 | 73,95 | 1644,74$_c$ | 129,44 | 13857,32 |
| 4 | 313 | 1138,73 | 67,8 | 1000,00$_d$ | 166,67 | 5175,98 |
| 5 | 370 | 1211,84 | 56,03 | 1025,76$_c$ | 100 | 4140,79 |
| 6 | 1076 | 559,74 | 49,26 | 503,00$_e$ | 39,41 | 2090,91 |
| 7 | 1236 | 875,41 | 87,74 | 629,66$_b$ | 32,9 | 6500 |
| 8 | 1579 | 723,13 | 87,55 | 524,48$_f$ | 53,03 | 7456,14 |
| 9 | 804 | 382,61 | 72,02 | 292,87$_g$ | 45,27 | 2862,54 |
| 10 | 948 | 464,5 | 54,98 | 397,73$_h$ | 91,83 | 2919,88 |
| 11 | 241 | 754,02 | 59,4 | 662,88$_b$ | 125 | 4139,43 |
| 12 | 821 | 243,98 | 73,86 | 205,71$_i$ | 30,95 | 1998,47 |
| Total | 8209 | 762,17 | 102,15 | 518,90 | 30,95 | 13857,32 |

Fonte: elaboração do autor a partir dos dados da pesquisa.
Nota: Letras iguais nas linhas indicam que as posições das medianas não diferem significativamente, $\alpha$=5%.

A Tabela 2 exibe as estatísticas descritivas do preço unitário dos terrenos por bairro com a mediana do preço unitário superior a R$ 2.000/m². Estes 438 terrenos correspondem a 5,34% do total dos terrenos analisados. O bairro do Meireles apresenta o maior coeficiente de variação de 67,00% e a maior média e mediana dos preços nesses bairros, a qual tem posição diferente das demais.

[44] A regional 7 engloba bairros bem distintos em matéria de preços como Guararapes (altos preços, renda elevada, zona de incorporação imobiliária etc.) e Sabiaguaba. O mesmo pode ser observado na regional 6 quando se comparam, por exemplo, os bairros Passaré e Planalto Ayrton Senna.
[45] A regional 1 é a de menor dimensão geográfica, composta por 3 (três) bairros apenas. O bairro Centro e Praia de Iracema guardam semelhanças, enquanto o bairro Moura Brasil difere daqueles em termos de preços, mas não foi capaz de influenciar significativamente o CV.

**Tabela 2 - Estatísticas descritivas do preço unitário por bairros com mediana superior a R$ 2.000/m².**

| Bairro | N | Média (R$) | Mediana (R$) | CV |
|---|---|---|---|---|
| Aldeota | 107 | 3013,54 | $2864,73_{a}$ | 37 |
| Centro | 110 | 2208,59 | $2021,46_{b}$ | 38 |
| Cocó | 34 | 2403,36 | $2352,63_{b,c}$ | 36 |
| Dionísio Torres | 28 | 2436,2 | $2128,80_{c}$ | 33 |
| Guararapes | 17 | 2467,43 | $2651,52_{a,b,c}$ | 34 |
| Meireles | 28 | 4356,23 | $3596,16_{d}$ | 67 |
| Mucuripe | 12 | 3100,63 | $2951,26_{a,c}$ | 40 |
| Papicu | 70 | 2227,35 | $2010,10_{b,c}$ | 53 |
| Praia de Iracema | 22 | 2332,55 | $2297,69_{b,c}$ | 39 |
| Varjota | 10 | 2409,58 | $2356,92_{a,b,c}$ | 36 |

Fonte: elaboração do autor a partir dos dados da pesquisa.
Nota:
Apenas terrenos com a mediana do preço unitário superior a R$ 2.000/m².
Letras iguais nas linhas indicam que as posições das medianas não diferem significativamente, $\alpha$=5%.

A Tabela 3 exibe as estatísticas descritivas do preço unitário dos terrenos por bairro, cuja mediana daquele seja inferior ou igual a R$ 300/m². Estes 2.134 terrenos correspondem a 26,00% do total dos terrenos analisados. Os bairros que apresentam maiores coeficiente de variação são Granja Lisboa e Granja Portugal. O bairro de Pedras, extremo sul da cidade, apresentou a menor média e mediana dos terrenos estudados.

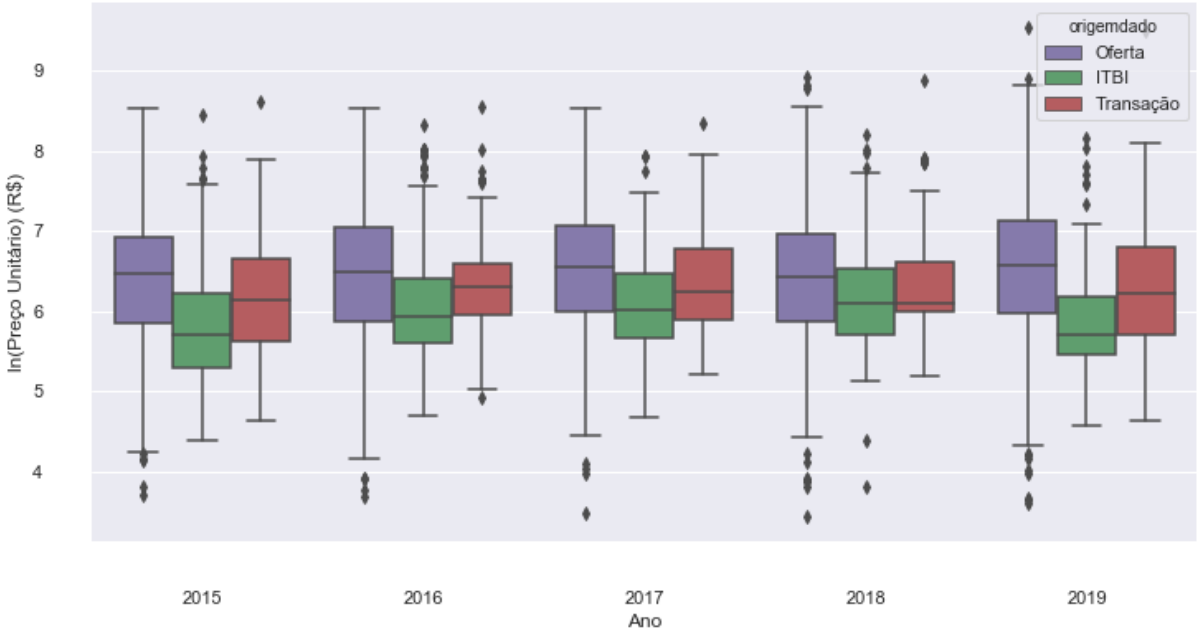
**Tabela 3 - Estatísticas descritivas do preço unitário por bairros com mediana inferior a R$ 300/m²**

| Bairro | N | Média | Mediana | CV% |
|---|---|---|---|---|
| Ancuri | 49 | 250,87 | $218,56_{a}$ | 40,67 |
| Aracapé | 125 | 299,04 | $277,78_{b}$ | 37,96 |
| Bom jardim | 50 | 292,96 | $245,79_{a,b}$ | 52,23 |
| Canindezinho | 61 | 311,62 | $270,00_{b}$ | 59,87 |
| Conjunto Palmeiras | 24 | 253,7 | $245,51_{c}$ | 52,42 |
| Granja Lisboa | 250 | 242,43 | $197,11_{d}$ | 84,49 |
| Granja Portugal | 60 | 316,84 | 258,79a,b | 77,32 |
| Jangurussu | 226 | 350,99 | $291,97_{b}$ | 74,32 |
| Parque Presidente Vargas | 60 | 217,57 | $221,59_{c}$ | 24,09 |
| Parque Santa Maria | 100 | 363,23 | $298,15_{e}$ | 46,79 |
| Pedras | 182 | 202,41 | $175,00_{d}$ | 38,2 |
| Planalto Ayrton Senna | 94 | 326,44 | $300,00_{b,e}$ | 44,01 |
| Sabiaguaba | 51 | 282,56 | $235,00_{a,b}$ | 59,03 |
| São Bento | 369 | 308,94 | $300,00_{b,e}$ | 29,58 |
| Siqueira | 433 | 212,98 | $200,00_{d}$ | 62,37 |

Fonte: elaboração do autor a partir dos dados da pesquisa.
Nota:
Apenas terrenos com a mediana do preço unitário inferior a R$ 300/m²
Letras iguais nas linhas indicam que as posições das medianas não diferem significativamente, $\alpha$=5%.

O Gráfico 3 demonstra o comportamento dos preços unitários (R$/m²) observados ao longo do período de corte de dados da amostra (5 anos), em escala logarítmica, para melhor visualização dada a dispersão dos dados dessa variável, e categorizado pela origem da informação (Oferta, ITBI ou Transação). Observa-se que os preços unitários referentes às ofertas permaneceram praticamente constantes no período e com mesma variabilidade (representada pelo intervalo interquartil). Dados provenientes das avaliações de ITBI tiveram uma queda mais acentuada em 2019, com mediana quase igual ao valor de 2015, o que pode sugerir uma retração no arbitramento da base de cálculo daquele imposto, mesmo que tardia, à crise do mercado imobiliário iniciada em 2015 e agravada em 2017. Os preços efetivamente transacionados também tiveram pouca flutuação no período, entretanto, apresentando maior dispersão em 2019.

**Gráfico 3 - *Boxplot* dos preços unitários observados na amostra (em escala lognormal) ao longo dos anos pela origem da informação.**

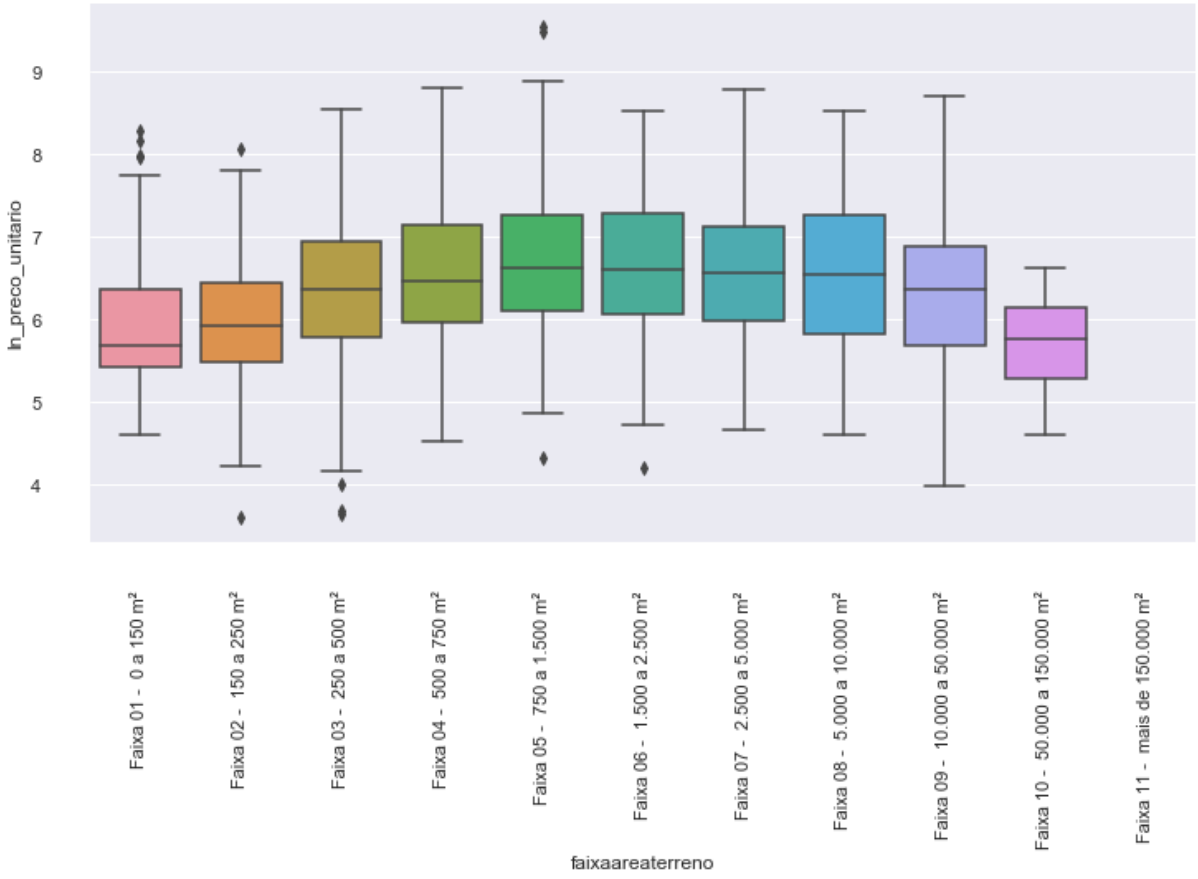



Fonte: elaboração do autor a partir dos dados da pesquisa.

Sabe-se que a área do terreno tem forte influência no seu preço. Entretanto, conforme Dantas (1998, p. 167) averte, não se pode partir aprioristicamente de um

comportamento de crescimento/decrescimento dos preços unitários com aumento/diminuição da área do terreno. A partir do Gráfico 4, observa-se um aumento dos preços unitários até faixa 05 com estabilização nas faixas 06 a 08 e decréscimo nas faixas 09 a 10. Essas últimas faixas correspondem a grandes terrenos, superiores a metragem padrão dos terrenos que ocupam toda uma quadra, com área de 10.000m² (100m x 100m). Nelas, se observa a aplicação do princípio da utilidade marginal decrescente, dada questões de liquidez e vocação destes para aproveitamento em conjuntos habitacionais, como bem ressalta aquele autor. Já nas faixas anteriores, isso não acontece, pois o mercado valoriza terrenos maiores, principalmente os de maior índice de aproveitamento (maior potencial construtivo), dada a possibilidade de maior incorporação de edificações (máximo aproveitamento possível).

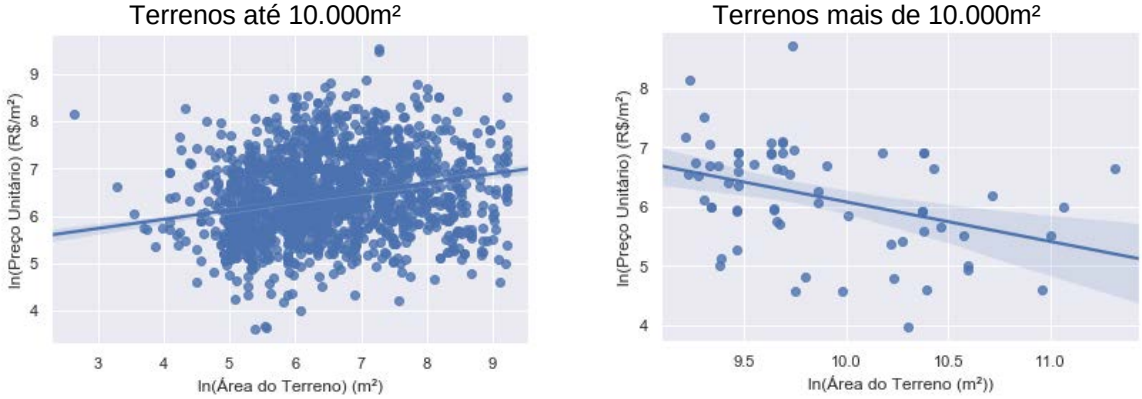
**Gráfico 4 - *Boxplot* dos preços unitários observados (R$/m²) no ano de 2019 e por faixa de área de terrenos.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 5, com 2 dispersões (e suas respectivas retas de regressão linear) para as categorias de terrenos com área até 10.000m² e mais de 10.000m², reforça o comportamento acima descrito[46].

**Gráfico 5 - Dispersão dos preços unitários (R$/m²) e área do terreno (m²) em escala logarítmica.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

### *5.1.2 Associação entre o preço unitário nas principais regionais*

Foram testadas várias faixas para o preço unitário dos terrenos, de modo a preservar o comportamento assimétrico da variável resposta e não violar um pressupostos do teste qui-quadrado, segundo o qual mais do que 20% das frequências esperadas sob a hipótese da independência não devem inferiores a cinco ou nenhuma delas deve ser igual a zero (SIEGEL; CASTELLAN JR., 2006). Desta feita, o preço unitário foi categorizado em sete faixas de valores: (1) R$ 30,95 a R$ 530,85; (2) R$ 530,95 a R$ 1.030,95, (3) R$ 1.030,95 a R$ 1.530,95, (4) R$ 1.530,95 a R$2.030,95, (5) R$ 2.030,95 a R$2.530,95, (6) R$ 2.530,95 a R$ 3.030,95 e (7) maiores que R$3.030,95.

A Tabela 4 exibe o resultado do teste qui-quadrado (p=0,000) de associação entre as faixas de preço nas principais regionais (1,3,4,5,7 e 8), com o resultado do teste de diferença das proporções nas categorias e as associações por meio dos resíduos qui-quadrado ajustados.

[46] Dantas (1998, p. 167), ao abordar essa questão na aplicação de modelos de regressão linear múltipla, sugere a utilização de um “Fator de Interação” dado pela a multiplicação entre a área do terreno e uma variável *dummy* de porte deste, segundo sua vocação. As variáveis “interação_incorporacao_vertical” e “interação_incorporacao_horizontal” representam essa abordagem, conforme se verá na seção 5.2.

As maiores associações significativas na segunda faixa de preço unitário (R$ 530,95 a R$ 1.030,95) foram com as regionais 5 e 7, as quais apresentaram diferenças nas proporções significativas.

Na terceira faixa prevalecem as regionais 4 e 5, as quais não apresentam diferenças significativas nas proporções e as maiores (iguais) associações significativas.

As regionais 1 e 3 nas maiores faixas de preço unitário dos terrenos. Os preços dos terrenos da regional 1 prevalecem nas faixas: R$ 1.530,95 a R$ 2.030,95, enquanto, os preços dos terrenos da regional 3 prevalecem nas duas últimas faixas.

Não primeira faixa, as diferenças das proporções de dados em todas essas regionais são significativas, com a regional 8 apresentando a maior associação nesta faixa.

**Tabela 4 - Tabela de associação entre as faixas do preço unitário nas principais regionais.**

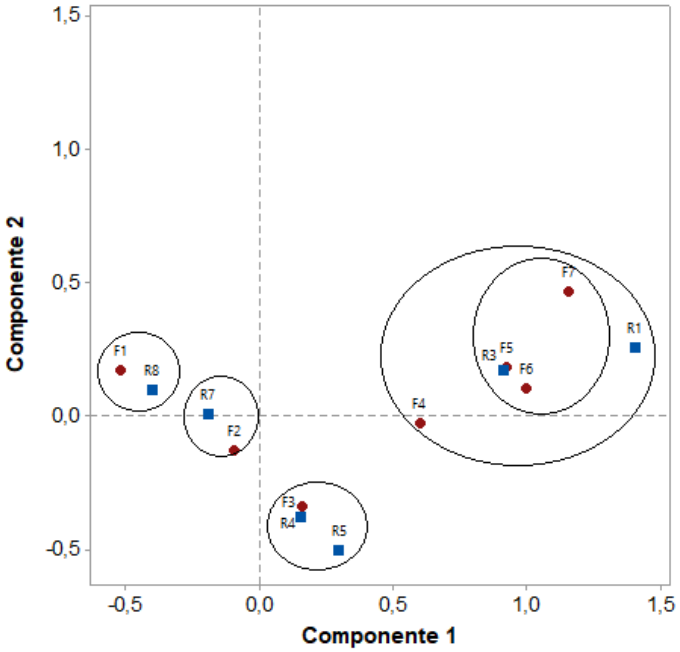
| Faixas preço unitário (R$) | Estatísticas | Regionais | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 3 | 4 | 5 | 7 | 8 | |
| 30,95 - 530,95 | Frequência | $0_a$ | $28_b$ | $60_c$ | $32_d$ | $480_e$ | $792_f$ | 1392 |
| | Resíduos | -6,70 | -11,70 | -4,30 | -8,20 | 3,50 | **11,70** | |
| 530,95-1.030,95 | Frequência | $4_a$ | $161_b$ | $105_{b,c}$ | $155_d$ | $443_c$ | $484_b$ | 1352 |
| | Resíduos | -6,00 | -1,60 | 0,40 | **3,30** | **2,30** | -1,10 | |
| 1.030,95-1.530,95 | Frequência | $25_{a,b}$ | $71_{b,c}$ | $82_a$ | $93_a$ | $132_c$ | $171_c$ | 574 |
| | Resíduos | 1,50 | -0,70 | **6,00** | **6,00** | -2,80 | -3,00 | |
| 1.530,95-2.030,95 | Frequência | $38_a$ | $96_b$ | $38_{b,c}$ | $44_c$ | $92_d$ | $69_e$ | 377 |
| | Resíduos | **7,40** | **6,30** | 1,90 | 1,90 | -1.8 | -6,10 | |
| 2.030,95-2.530,95 | Frequência | $28_a$ | $66_b$ | $9_c$ | $23_d$ | $34_c$ | $30_c$ | 190 |
| | Resíduos | **8,90** | **8,00** | -1,40 | 1,50 | -2,90 | -4,90 | |
| 2.530,95-3.030,95 | Frequência | $18_a$ | $47_a$ | $8_{b,c}$ | $17_c$ | $26_b$ | $12_d$ | 128 |
| | Resíduos | **6,80** | **7,10** | -0,50 | 1,70 | -1,90 | -5,20 | |
| Maiores 3.030,95 | Frequência | $22_a$ | $98_a$ | $11_b$ | $6_{b,c}$ | $29_b$ | $21_c$ | 187 |
| | Resíduos | **6,50** | **14,50** | -0,80 | -2,60 | -3,50 | -5,90 | |
| Total | | 135 | 567 | 313 | 370 | 1236 | 1579 | 4200 |

Fonte: elaboração do autor a partir dos dados da pesquisa.
Notas: letras iguais indicam que a posição das medianas não diferem significativamente, $\alpha$=5%.

Esse comportamento das faixas de preços com as regionais pode ser observado no **Erro! Autoreferência de indicador não válida.**, elaborado com a

técnica de análise de correspondência. Foi feita uma redução dimensional, obtendo-se os dois componentes que representam 94,90% dos dados.

**Gráfico 6 - Associação das principais regionais com as faixas de preço unitário dos terrenos através de análise de correspondência.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

## 5.2 Modelo de regressão linear múltipla

Foi estimado inicialmente um modelo por regressão linear múltipla pelo MQO, pelo procedimento *stepwise backward*, através da amostra de treinamento cujo resultados se encontram na Tabela 5:

**Tabela 5 - Estimação da equação pelo modelo de regressão linear múltipla na amostra de treinamento.**

| Dependente: ln(preco_imovel) | coeficiente | *p-value* | FIV |
|---|---|---|---|
| c | 4,9548 | 0,0000 | 239,3118 |
| regional_01 | 0,4649 | 0,0000 | 1,8576 |
| regional_02 | 0,1438 | 0,0020 | 1,2327 |
| regional_03 | 0,2774 | 0,0000 | 3,1290 |
| regional_04 | 0,3265 | 0,0000 | 1,7108 |
| regional_05 | 0,3270 | 0,0000 | 1,9058 |
| regional_06 | 0,2073 | 0,0000 | 2,5793 |
| regional_07 | 0,2304 | 0,0000 | 3,6123 |
| regional_08 | 0,3720 | 0,0000 | 3,3101 |

| **Dependente: ln(preco_imovel)** | **coeficiente** | ***p-value*** | **FIV** |
|---|---|---|---|
| regional_09 | 0,2278 | 0,0000 | 2,0515 |
| regional_10 | 0,2008 | 0,0000 | 2,2701 |
| regional_11 | 0,2498 | 0,0000 | 1,4406 |
| loteamento_condominio | 0,5974 | 0,0000 | 1,2669 |
| zona_incorporacao_vertical | 0,0417 | 0,0570 | 1,9205 |
| avenida | 0,1759 | 0,0000 | 1,5322 |
| via_01_expressa | 0,0997 | 0,0640 | 1,1158 |
| via_02_arterial_1 | 0,148 | 0,0000 | 1,3869 |
| via_05_paisagistica | 0,1812 | 0,0030 | 1,1039 |
| via_06_comercial | 0,3819 | 0,0000 | 1,3464 |
| numero_frentes | 0,0846 | 0,0000 | 1,4277 |
| ln_renda | 0,1845 | 0,0000 | 2,5792 |
| ln_testada | 0,0851 | 0,0000 | 8,1226 |
| ln_area_terreno | -0,0815 | 0,0000 | 8,1231 |
| percentual_area_preservacao | -8,5196 | 0,0000 | 1,0291 |
| esgoto | 0,0404 | 0,0260 | 1,8223 |
| galeria_pluvial | 0,0429 | 0,0070 | 1,9843 |
| pavimentacao_asfalto_concreto | 0,0684 | 0,0000 | 1,6110 |
| indice_aproveitamento_maximo_equivalente | 0,1412 | 0,0000 | 2,5194 |
| influencia_distancia_beiramar | 0,0082 | 0,0030 | 1,2149 |
| densidade_comercializacao_trecho | 0,09 | 0,0000 | 1,2885 |
| distancia_via_principal | -0,0002 | 0,0000 | 1,4801 |
| ln_valor_m2_terreno_face_quadra_iptu_2014 | 0,2923 | 0,0000 | 4,9667 |
| ln_assentamento_precario_area_percentual | -0,0096 | 0,0000 | 1,1968 |
| interacao_incorporacao_vertical | 7,09E-06 | 0,0000 | 1,5034 |
| ln_idh_educ | 1,3732 | 0,0000 | 3,7736 |
| ano_2016 | 0,0709 | 0,0000 | 1,5499 |
| ano_2017 | 0,0785 | 0,0000 | 1,4393 |
| ano_2018 | 0,1082 | 0,0000 | 1,5259 |
| ano_2019 | 0,035 | 0,0200 | 1,6639 |
| origem_oferta | 0,1106 | 0,0000 | 1,4783 |
| origem_transacao | 0,1128 | 0,0000 | 1,1560 |

| | | | |
|---|---|---|---|
| N | 6.567 | $R^2$ | 0,752 |
| Estatística F | 494,3 | $R^2$ ajustado | 0,750 |
| Probabilidade Jarque-Bera | 0,00 | AIC | 6824 |

Fonte: elaboração do autor a partir dos dados da pesquisa.
Notas:
O prefixo “ln_” antes da variável indica que foi aplicada uma transformação pelo logaritmo natural.
FIV: fator de inflação de variância.

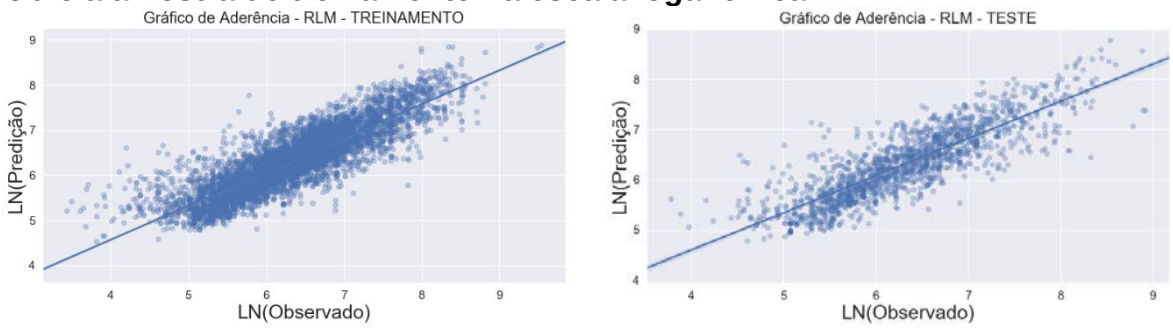
O modelo apresentou alto valor da estatística F rejeitando-se a hipótese nula, com todas as variáveis escolhidas sendo conjuntamente significantes. Com

exceção da variável “zona_incorporacao_vertical”, todas as variáveis foram significantes a 5%.

A multicolinearidade (não perfeita) nos modelos de regressão causa muita controvérsia na literatura. Dantas (1988, p. 133) ressalta que ela está presente nos dados de mercado imobiliário e cita o exemplo de glebas que geralmente estão situadas distantes do centro. Desta feita, um modelo que utilize as variáveis “porte” (relacionada à área do terreno) e “distância ao centro” apresentará alta colinearidade nestas. Sabe-se que a multicolinearidade pode alargar os intervalos de confiança dos coeficientes da regressão, tornando a estatística t menos confiável, entretanto, na maioria das vezes não afeta a qualidade da predição (GUJARATI; PORTER, 2011, p. 353). Dantas (ibidem), sobre o exemplo citado, afirma que o modelo pode ser utilizado, mas não se prestando para predições de glebas perto do centro. A multicolinearidade pode ser testada com a matriz de correlação das variáveis quando tomadas duas a duas (com atenção especial a resultados superiores a 0,80) (ABNT, 2011, p. 36) ou pelo cálculo, para cada regressor, do fator de inflação da variância (FIV). Muito embora haja críticas sobre a real utilidade do FIV, dada que as variâncias amostrais dos estimadores não dependem só do FIV (WOOLDRIDGE, 2016, p. 103), foram calculados todos os seus valores, conforme última coluna da Tabela 5, onde se observa que todos tiveram valores menores que 10. As variáveis “ln_testada” e “ln_area_terreno” foram as mais correlacionadas com as demais, apresentando os maiores valores de FIV.

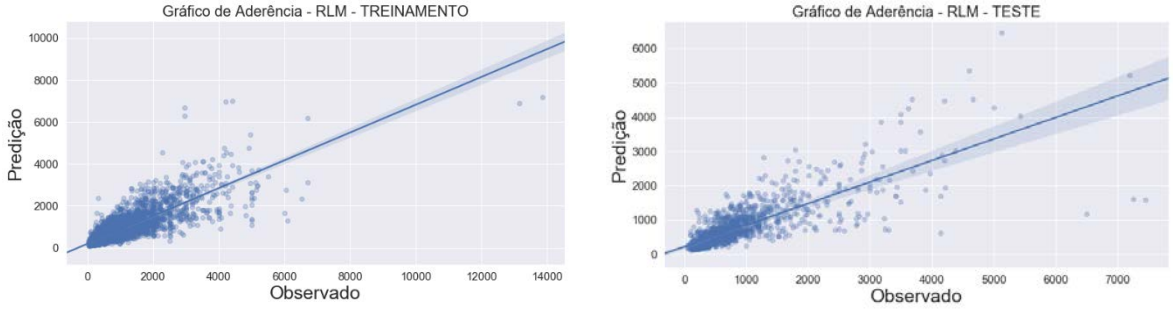
Todas as variáveis tiveram seu sinal de coeficiente de acordo com o esperado pela prática. Os gráficos de dispersão entre o observado e o predito, em valores de (RS/m²) estão representados para as amostras de treinamento e teste através do Gráfico 7 e Gráfico 8.

**Gráfico 7 - Dispersão do observado (R$/m²) x predito (R$/m²) do modelo MQO sobre a amostra de treinamento na escala logarítmica.**



Fonte: elaboração do autor a partir dos dados da pesquisa.

**Gráfico 8 - Dispersão do observado (R$/m²) x predito (R$/m²) do modelo MQO sobre a amostra de treinamento.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 8 mostra dados com uma boa aderência ao dados até o preço unitário observado de R$ 4.000/m², onde, a partir daí, começa uma grande dispersão do predito frente ao observado. Isso pode ser um erro aleatório ou falta de uma variável explicativa para terrenos de grande valor de mercado ou presença de heterocedasticidade. Sabe-se que esta última não causa viés ou inconsistência nos coeficientes betas da regressão MQO, mas invalida os erros padrão e as estatísticas de testes usuais (WOOLDRIDGE, 2016, p. 323) Para verificação da heterocedasticidade, aplicou-se o teste de Breusch-Pagan, sob a hipótese nula de homocedasticidade. Atestou-se a presença de heterocedasticidade[47] e, desta feita, calculou-se os novos erros padrão e estatísticas t robustas à heterocedasticidade pelo procedimento de White (HC1). Com esse procedimento, apenas a variável "via_01_expressa" não foi significante a 10% (p-value = 0,118), e as variáveis "zona_incorporacao_vertical" (p-value=0,056) e influencia_distancia_beiramar (p-value=0,053) não foram significantes a 5%.

O referido gráfico ainda apresenta grande erro em dois dados onde o preço unitário observado foi em torno de R$ 14 mil/m² (lado esquerdo, canto superior direito). Esses dados se referem a um mesmo terreno (oferta e declaração de ITBI) localizado na Av. Abolição, por detrás do Clube Náutico Atlético Cearense. Não se encontrou tecnicamente nenhuma justificativa para o preço observado, mas também nenhuma outra que justificasse a exclusão desses dados da modelagem.

Embora não atingida a normalidade dos resíduos, considerando-se o tamanho da amostra, os testes t e F serão válidos assintoticamente, sem prejuízos

[47] Estatística LM de 566,21 e p-value de 0.

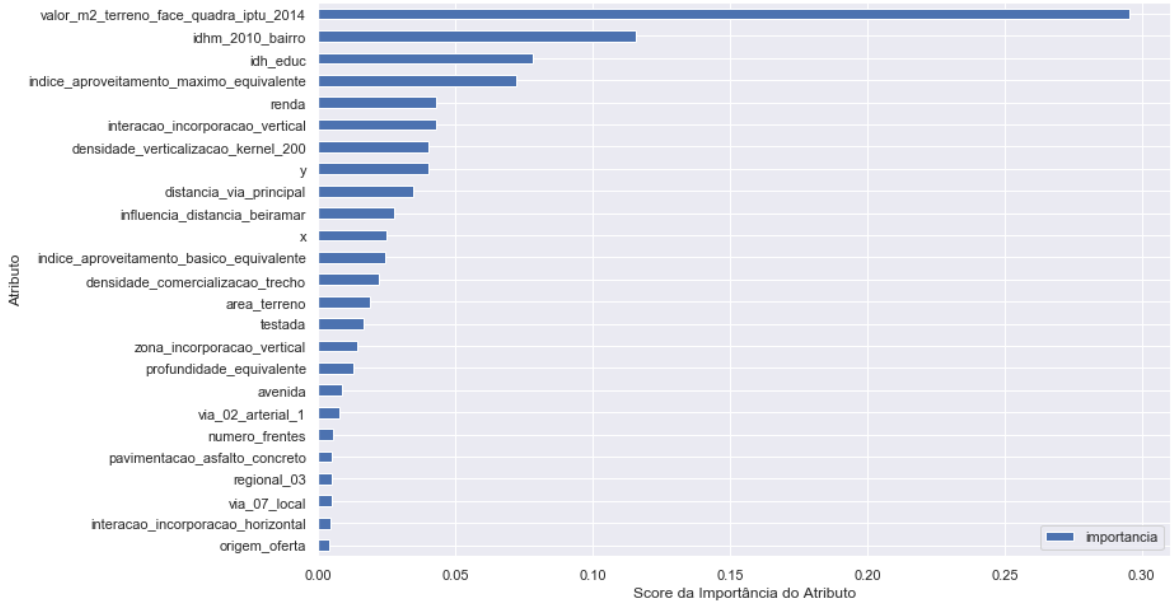
para o modelo e muito menos para as estimativas geradas (GUJARATI; PORTER, 2011, p. 327-328).

## 5.3 Modelo de florestas aleatórias

O modelo de florestas aleatórias foi estimado inicialmente sobre a mesma base de treinamento do modelo MQO. Foram utilizadas todas as variáveis independentes (atributos), conforme definidas na seção 3.4, acrescidas das coordenadas planas UTM no *datum* SIRGAS 2000 dos centroides dos terrenos, “x” e “y”, a fim de capturar uma certa dependência espacial. Através de um processo de pesquisa dos melhores conjuntos de hiperparâmetros para otimizar o erro quadrático médio numa validação cruzada de cinco partições (*folders*), chegou-se aos seguintes valores daqueles: 700 árvores, com profundidade máxima de 12 níveis, 17 atributos escolhidos aleatoriamente em cada nó e com, no mínimo, três dados em cada folha.

O Gráfico 9 mostra os 25 atributos com maior escore segundo o cálculo do próprio algoritmo. As 13 variáveis mais importantes se referem a variáveis de localização. Em seguida, as variáveis área e comprimento da testada principal, como representantes de variáveis do tipo estrutural. Apesar da defasagem da atual PGV, o seu valor base de face de quadra se mostrou como a variável mais importante na otimização da divisão dos nós das árvores e minimização do erro. As variáveis de IDH, seja na sua dimensão completa, seja na sua dimensão educação se mostram bastante importantes. O índice de aproveitamento máximo equivalente se mostrou mais importante que o índice de aproveitamento básico equivalente, o que demonstra uma maior explicação dos preços pelo potencial construtivo adicional (diferença entre os dois índices). A variável “renda” como *proxy* espacial ocupa a quinta posição, comprovando a hipótese de correlação forte entre renda e preço de imóvel.

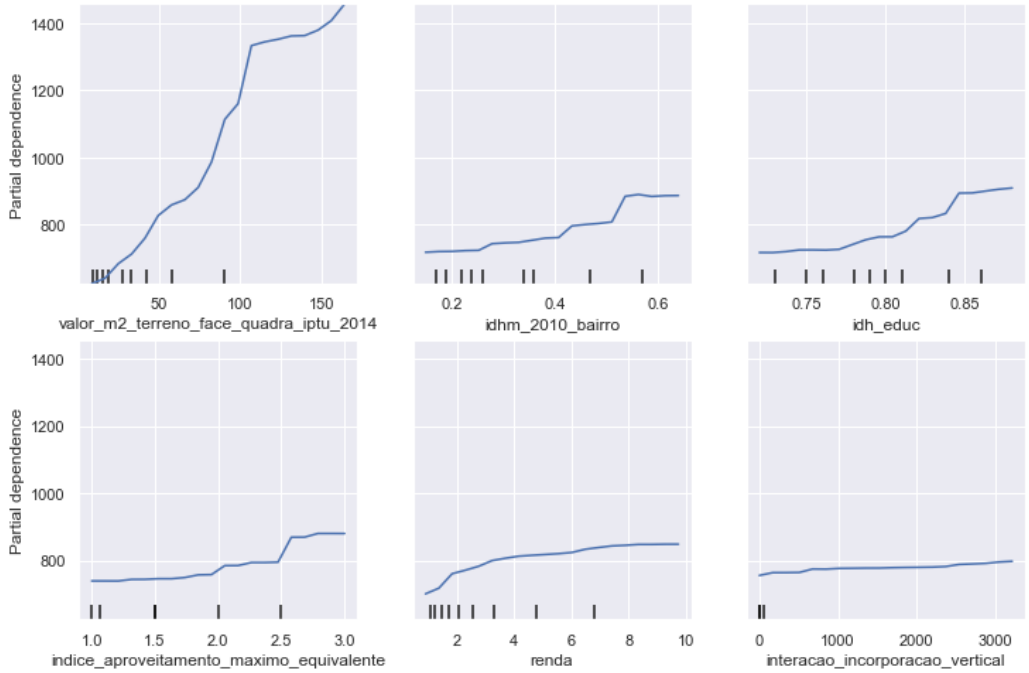
**Gráfico 9 - Escore de importância dos 25 principais atributos do modelo RF.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

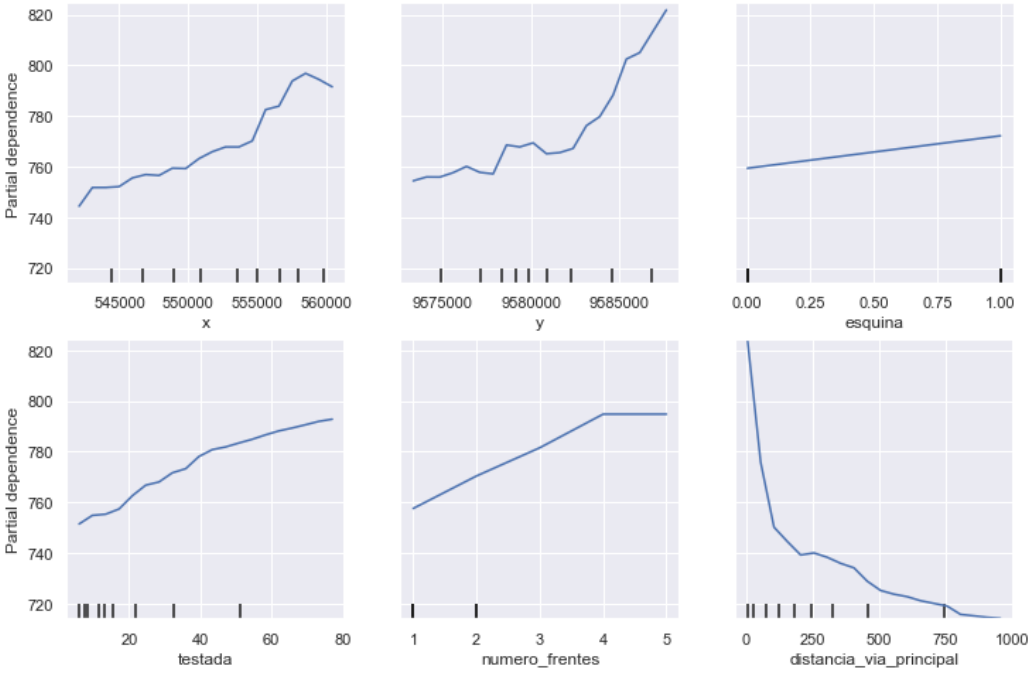
Como comentado na seção 4.2.6, os gráficos de dependência parcial se mostram úteis para visualizar a variação da variável dependente com determinado atributo, testando-se desta feita, as hipóteses iniciais esperadas de comportamento daquela variável. É o que se apresenta nos Gráfico 10 e Gráfico 11, com destaque para o decaimento do preço unitário à medida que aumenta a distância do terreno à via principal mais próxima (canto inferior direito do Gráfico 11):

**Gráfico 10 - Dependência parcial do preço unitário com algumas variáveis mais importantes do modelo RF.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

**Gráfico 11 - Dependência parcial do preço unitário com algumas variáveis mais importantes do modelo RF.**

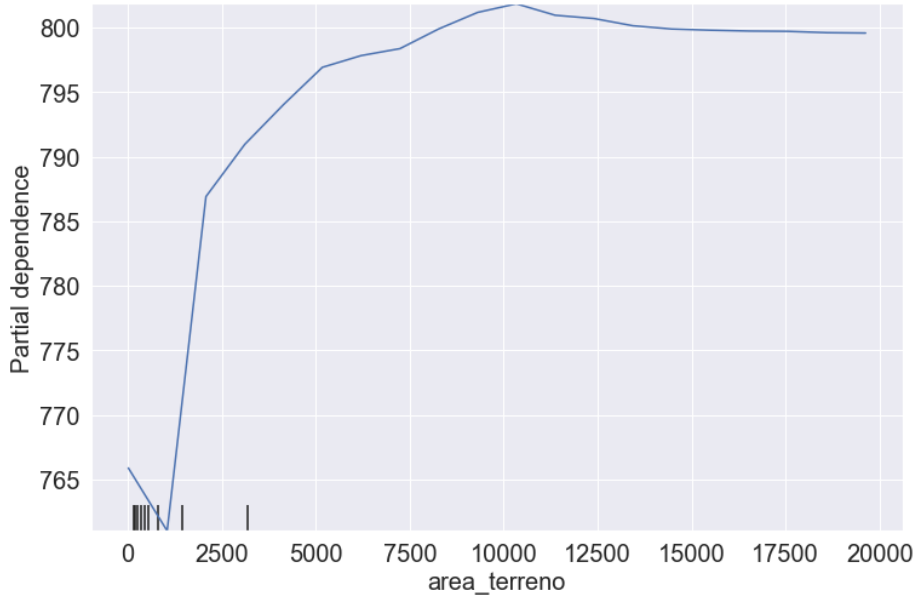



Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 12 apresenta importante relação do preço unitário com a área do terreno. Observa-se que o preço unitário decai com o aumento da área, passa a

aumentar até a área aproximada de 10.000m² e depois retorna a cair. Portanto, três faixas de comportamento de acordo com a área. Para a primeira faixa, temos os terrenos de pequenas dimensões (lotes urbanos padrão de até 750m²). Para a terceira faixa, temos os grandes terrenos de área superior à quadra urbana padrão (10.000m²). Nessas duas faixas, se aplica o princípio da utilidade marginal decrescente, o preço unitário (R$/m²) decai à medida que a área aumenta. Nos terrenos da faixa intermediária, por serem adequadas à incorporação imobiliária, isso não acontece: quanto maior a área, maior é o preço observado, dada a maior possibilidade de empreender unidades com alto padrão de acabamento.

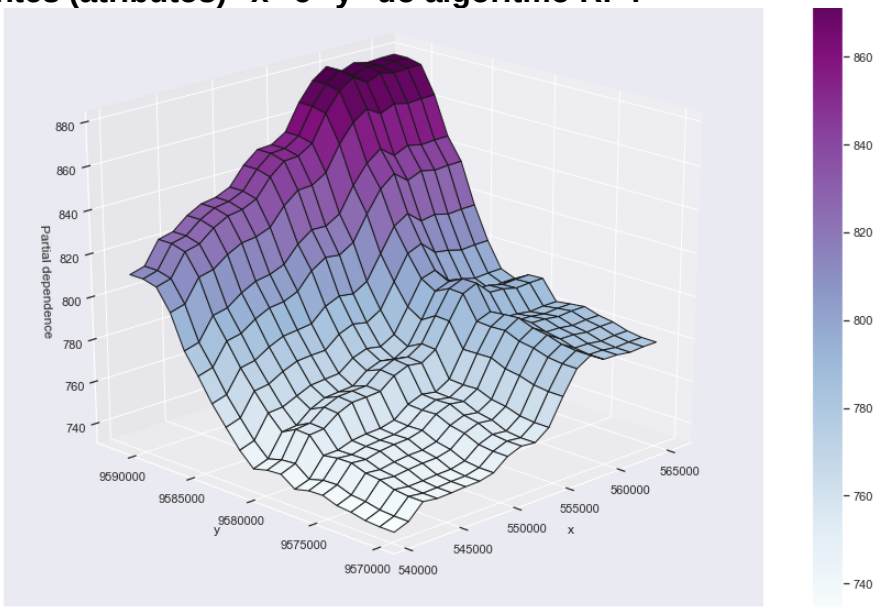
**Gráfico 12 - Dependência parcial do preço unitário com a área do terreno.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

Outro ponto que merece destaque é a importância relativa da direção norte-sul, representada pela variável “y” frente à direção leste-oeste, representada pela variável “x”. Isso indica claramente a valorização do litoral norte como importante região de incorporação imobiliária, atividade comercial e turística, onde há escassez e demanda elevada de terrenos com grande potencial construtivo. O Gráfico 13 de dependência parcial dos preços unitário observados com as duas variáveis de coordenadas UTM, “x” e “y”, dá uma visualização espacial do comportamento ora explicado. Através desse mesmo gráfico, observa-se que a região sudoeste apresenta os menores preços unitários, exatamente onde se situam as áreas mais pobres do município, regionais 10 e 12.

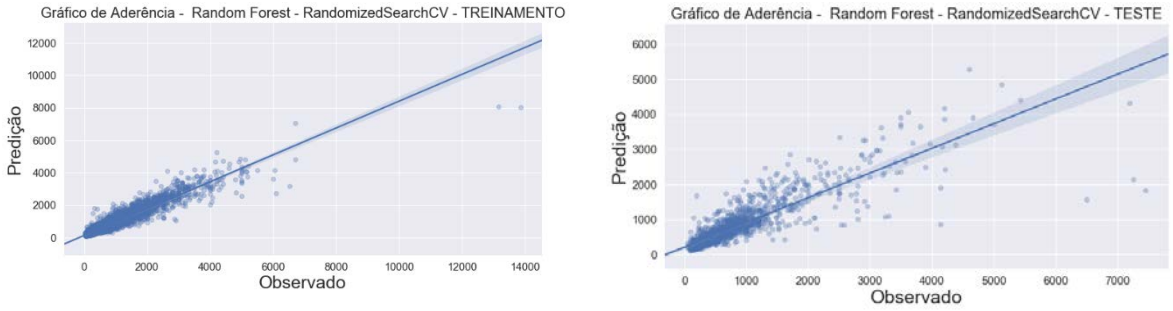
**Gráfico 13 - Dependência parcial 3D dos preços unitários com as variáveis independentes (atributos) "x" e "y" do algoritmo RF .**




Fonte: elaboração do autor a partir dos dados da pesquisa.

Através do Gráfico 14 (à esquerda), também se observa a falta de ajuste dos dois terrenos com preços em torno de R$ 14 mil/m², entretanto, a aderência dos dados se mostra muito mais eficaz. O coeficiente de correlação de Pearson para as duas variáveis (observado e predito) foi de 0,951. O coeficiente de determinação ($R^2$) para os dados *out-of-bag*[48] foi de 0,78.

**Gráfico 14 - Dispersão do observado (R$/m²) x predito (R$/m²) do modelo RF sobre a amostra de treinamento e teste.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

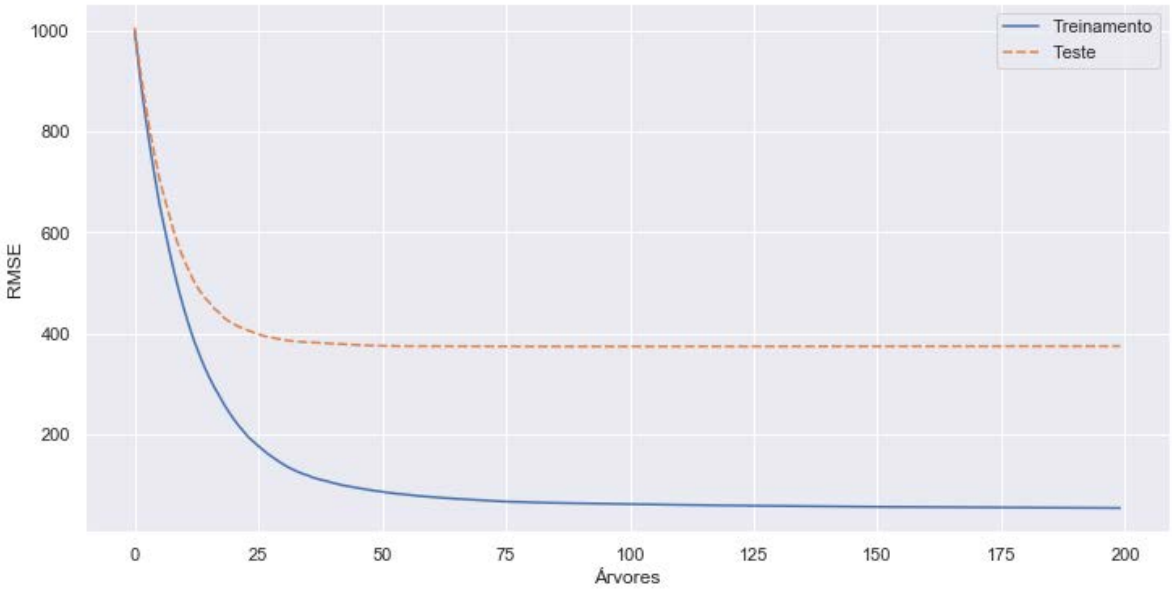
[48] Como o algoritmo de florestas aleatórias seleciona aleatoriamente apenas um subconjunto dos dados para cada árvore a ser treinada, por *bootstrap*, haverá dados que não foram selecionados, conhecidos como *out-of-bag* (*OOB*). Para cada observação, pode-se selecionar as árvores onde o dado não foi selecionado e inferir o potencial de predição do modelo, sem a necessidade de se utilizar os dados de teste. Pode-se demonstrar matematicamente que a proporção aproximada de dados OOB é de 36,88% para cada amostra *bootstrap*.

O Gráfico 14 (à direita) mostra a dispersão entre os valores observados e estimados sobre a base de teste. O coeficiente de correlação de Pearson para as duas variáveis foi de 0,866.

## 5.4 Modelo XGBoost

O modelo XGBoost foi estimado sobre a mesma base de treinamento dos modelos anteriores. Para a otimização dos hiperparâmetros, dado que o algoritmo XGBoost é guloso, rapidamente provocando sobreajustamento, definiu-se inicialmente a profundidade máxima da árvore igual a 12 (mesmo valor do algoritmo de florestas aleatórias) e escolha aleatória de 1/3 dos atributos para a divisão do nó. A partir daí, aumentou-se incrementalmente o número de preditores (árvores) e verificou-se a redução no MAE e RMSE, tanto nos dados de treinamento, como nos de teste.

**Gráfico 15 - Variação da raiz quadrado do erro médio (RMSE) no treinamento e teste com a quantidade de árvores no algoritmo XGBoost.**

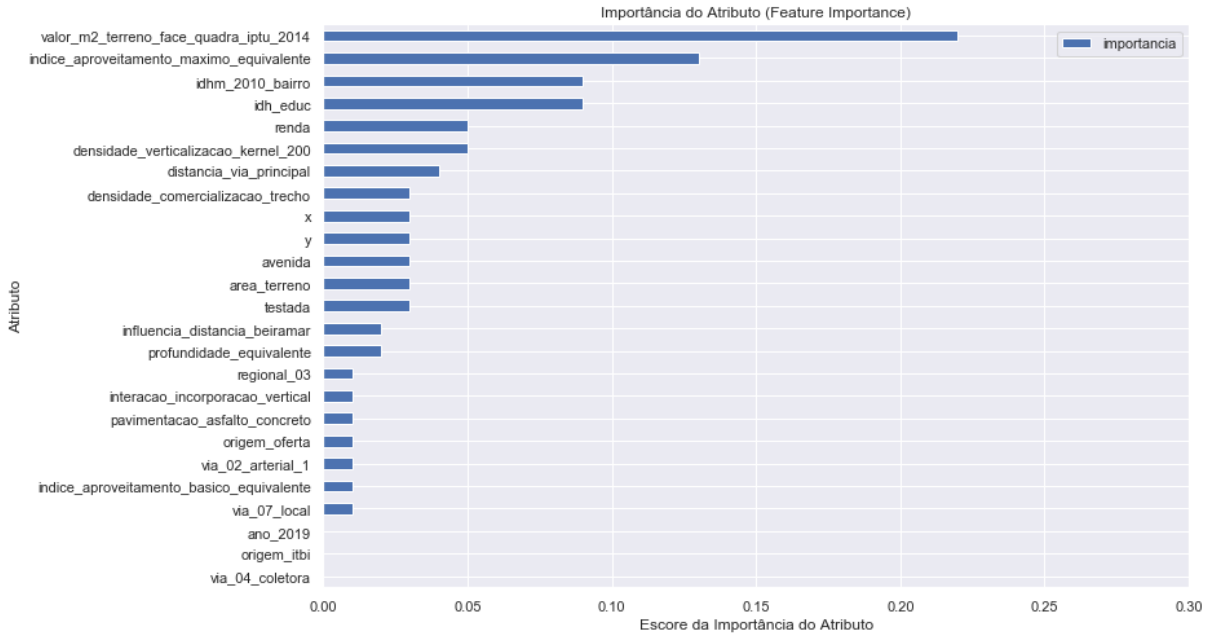



Fonte: elaboração do autor a partir dos dados da pesquisa.

Através do Gráfico 15, observa-se redução do RMSE no treinamento até a utilização de 200 árvores. Já quanto ao teste, a redução se inicia com 25 e 50 árvores respectivamente. A escolha do número de árvores, no total de 200, se deu por conta da redução dos erros no treinamento.

O Gráfico 16 mostra a importância relativa entre os 25 principais atributos, entendida esta como o ganho total na otimização da função de perda ao se utilizar o referido atributo na divisão do nó. Novamente, o atributo “valor_m2_terreno_face_quadra_iptu_2014” se apresenta como o mais importante. Comparando-se com o Gráfico 9, de importância relativa do modelo das florestas aleatórias, vê-se que o “indice_aproveitamento_maximo_equivalente” desbanca os índices “idhm_2010_bairro” e “idh_educ”. O atributo “renda” ocupa a quinta posição em ambos algoritmos. Observa-se ainda que o atributo “numero_frentes” estava na 20ª posição no Gráfico 9 e não aparece entre os 25 do Gráfico 16. Os atributos “origem_itbi” e “ano_2019” passam a compor no algoritmo XGBoost relevância maior.

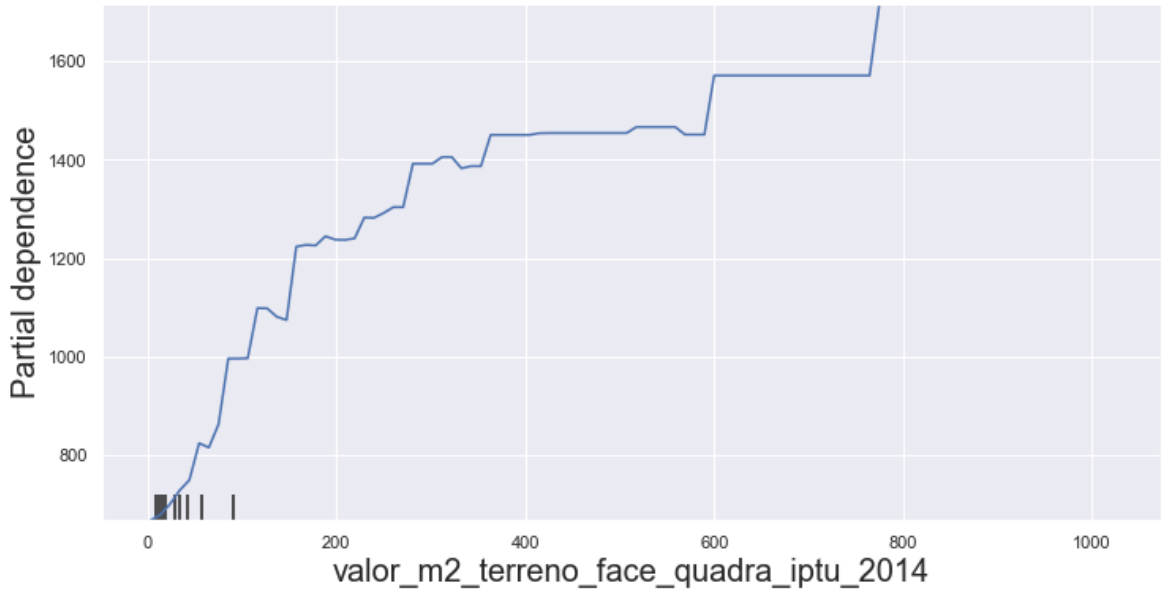
**Gráfico 16 - Escore de importância dos 25 principais atributos do modelo XGBoost.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

Dada a importância do atributo "valor_m2_terreno_face_quadra_iptu_2014", representa-se a variação do preço unitário com à medida que se modifica o valor desse atributo. Está claro, através do Gráfico 17, que a relação é positiva, mas não linear.

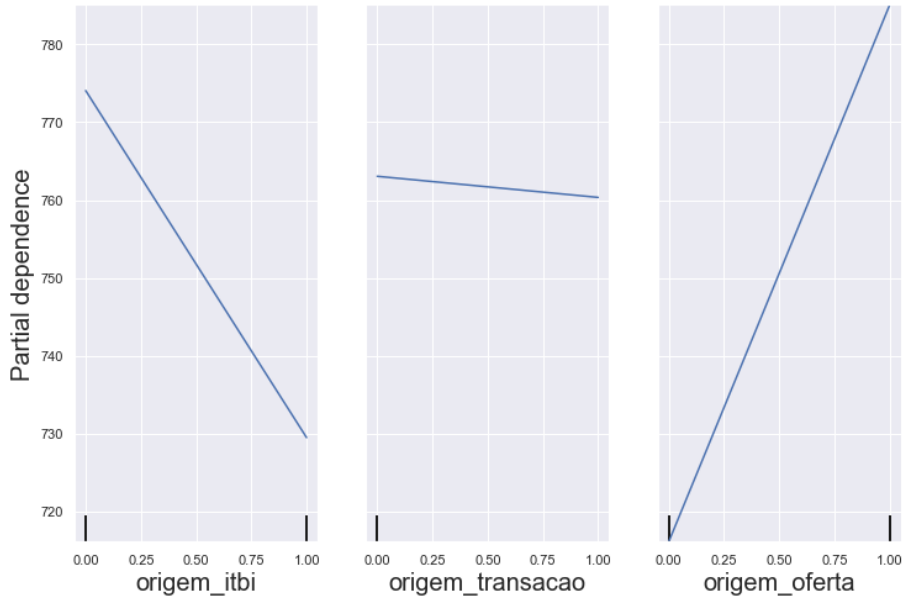
**Gráfico 17 - Dependência parcial do preço unitário com "valor_m2_terreno_face_quadra_iptu_2014" no modelo XGBoost.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 18 mostra como o preço unitário varia com a origem da informação, seja ela proveniente de uma avaliação de ITBI, de um valor considerado como transação ou como uma oferta de mercado imobiliário. Conclui-se que as avaliações de ITBI são inferiores aos valores declarados, que por sua vez são inferiores aos valores de oferta.

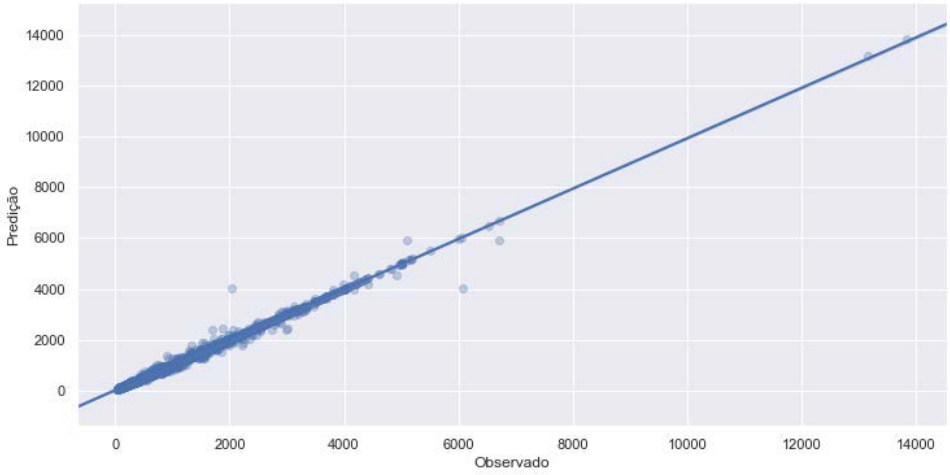
**Gráfico 18 - Dependência parcial do preço unitário com a origem da informação no modelo XGBoost.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 19 demonstra como o algoritmo XGBoost é propício ao sobreajustamento. Observa-se uma dispersão mínima dos dados na reta de 45º, inclusive com resíduo quase zero sobre os dois dados atípicos com preços unitários em torno de R$14mil/m². O coeficiente de correlação de Pearson para as duas variáveis foi de 0,998.

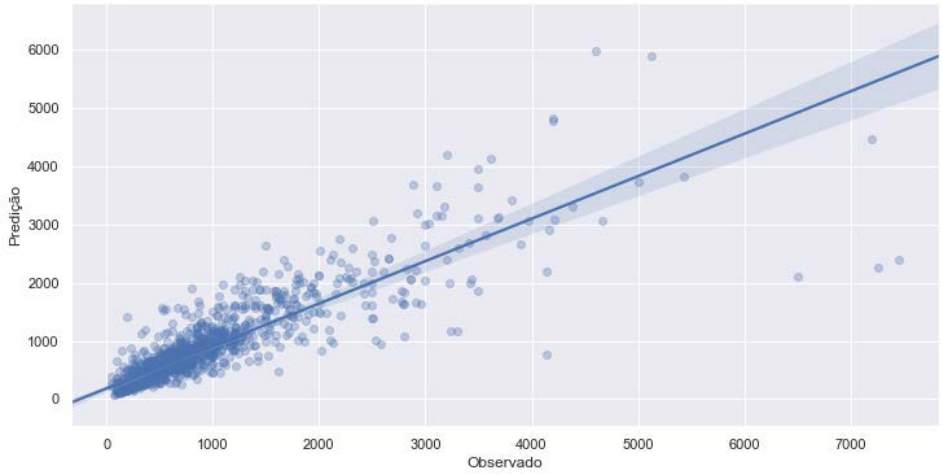
**Gráfico 19 - Dispersão do observado (R$/m²) x predito (R$/m²) do modelo XGBoost sobre a amostra de treinamento.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 20 mostra a dispersão entre o observado e o predito sobre os dados de teste. Também se observam erros crescentes para preços unitários a partir de R$ 2.000/m². O coeficiente de correlação de Pearson para as duas variáveis foi de 0,878, sendo superior ao correspondente do modelo das florestas aleatórias (0,866).

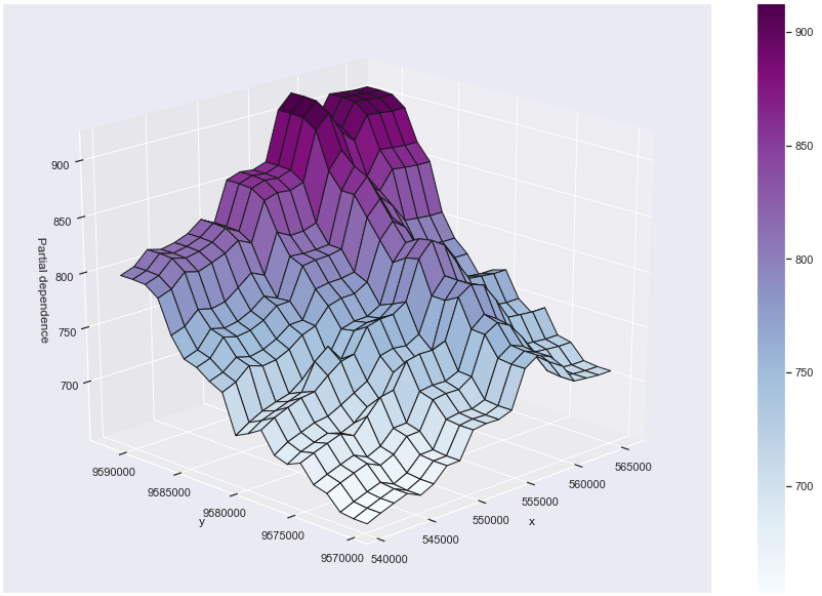
**Gráfico 20 - Dispersão do observado (R$/m²) x predito (R$/m²) do modelo XGBoost sobre a amostra de teste.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 21 demonstra a dependência parcial dos preços unitários com os atributos “x” e “y”. Comparando-se com o Gráfico 13, observam-se mudanças nos planos mais acentuadas, o que indica um maior sobreajustamento aos dados realizado pelo algoritmo XGBoost.

**Gráfico 21 - Dependência parcial 3D dos preços unitários com as variáveis independentes (atributos) "x" e "y" do algoritmo XGBoost.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

## 5.5 Estimativas de desempenho

A Tabela 6 sintetiza as métricas dos três modelos, inclusive com os resultados de treinamento e teste. A análise dos resultados deve ser realizada para os dados de teste, pois representa a capacidade de generalização dos modelos a novos dados que não se submeteram a treinamento.

**Tabela 6 - Comparativo das estimativas de desempenho entre os modelos analisados.**

| Modelo | MQO | | RF | | XGBoost | |
|---|---|---|---|---|---|---|
| | Treino | Teste | Treino | Teste | Treino | Teste |
| Nível de avaliação | 0,98 | 0,99 | 1,04 | 1,04 | 1 | 1,01 |
| COD (%) | 33,48 | 34,45 | 21,51 | 27,13 | 2,88 | 24,05 |
| RMSE (R$/m²) | 430,83 | 466,65 | 249,52 | 391,05 | 54,59 | 375,05 |
| MAE (R$/m²) | 226,82 | 236,33 | 132,1 | 185,14 | 17,13 | 171,29 |
| MAPE (%) | 32,93 | 34,09 | 22,76 | 28,64 | 2,88 | 24,40 |

Fonte: elaboração do autor a partir dos dados da pesquisa.

Como se observa, o modelo XGBoost teve melhor performance em todos as métricas avaliadas. O modelo das florestas aleatórias ficou numa situação intermediária, sendo melhor que o modelo MQO em todas as métricas. Apenas os modelos de aprendizado de máquina tiveram a métrica de uniformidade de avaliação aceitável, dentro do estabelecido pela Portaria 511/09, cujo limite (MAPE) é 30%. Apesar do modelo XGBoost apresentar métricas de treinamento que possam indicar um sobreajustamento, isso não se confirmou com os resultados nos dados de teste, tendo, nestes, performance inclusive superior ao modelo das florestas aleatórias. Não é à toa que esse algoritmo é o campeão nas competições de aprendizado de máquina realizadas pela comunidade Kaggle.

A Tabela 7 mostra uma estatística descritiva comparativa dos resíduos dos três modelos, onde se observa que o modelo XGBoost tem menor mediana, menor desvio padrão e menor valor máximo. Apenas sua média é superior ao modelo de florestas aleatórias.

**Tabela 7 - Comparativo da descritiva dos resíduos dos modelos.**

| | Resíduos ($y - \hat{y}$) | | |
|---|---|---|---|
| | XGBoost | RF | MQO |
| N | 8209 | 8209 | 8209 |
| média | 4,62 | 2,04 | 59,62 |
| desv. pad. | 174,65 | 283,55 | 434,18 |
| min | -1991,27 | -1497,84 | -3747,70 |
| 25% | -9,51 | -90,25 | -98,94 |
| 50% | -0,45 | -18,12 | 6,45 |
| 75% | 7,84 | 50,02 | 121,54 |
| max | 5060,94 | 5818,25 | 6657,97 |

Fonte: elaboração do autor a partir dos dados da pesquisa.

Para se testar se existe realmente diferenças estatísticas entre o modelo MQO e os de aprendizado de máquina, aplicou-se o teste de postos sinalizados de Wilcoxon que tem como hipótese nula (teste bilateral) a igualdade de medianas, e hipótese alternativa, a diferença delas. Este teste foi escolhido, dado a não normalidade observada na distribuição dos resíduos, segundo o teste de Shapiro-Wilk. Em todos os três testes, realizados dois a dois, rejeitou-se a hipótese nula de que as medianas dos erros são iguais com significância de 1%[49]. De certa forma, isso nos leva a concluir que existem diferenças significativas entre os modelos, e, como as medidas de performance apresentadas dependem de alguma maneira destes resíduos, há também diferenças estatisticamente significantes em tais medidas.

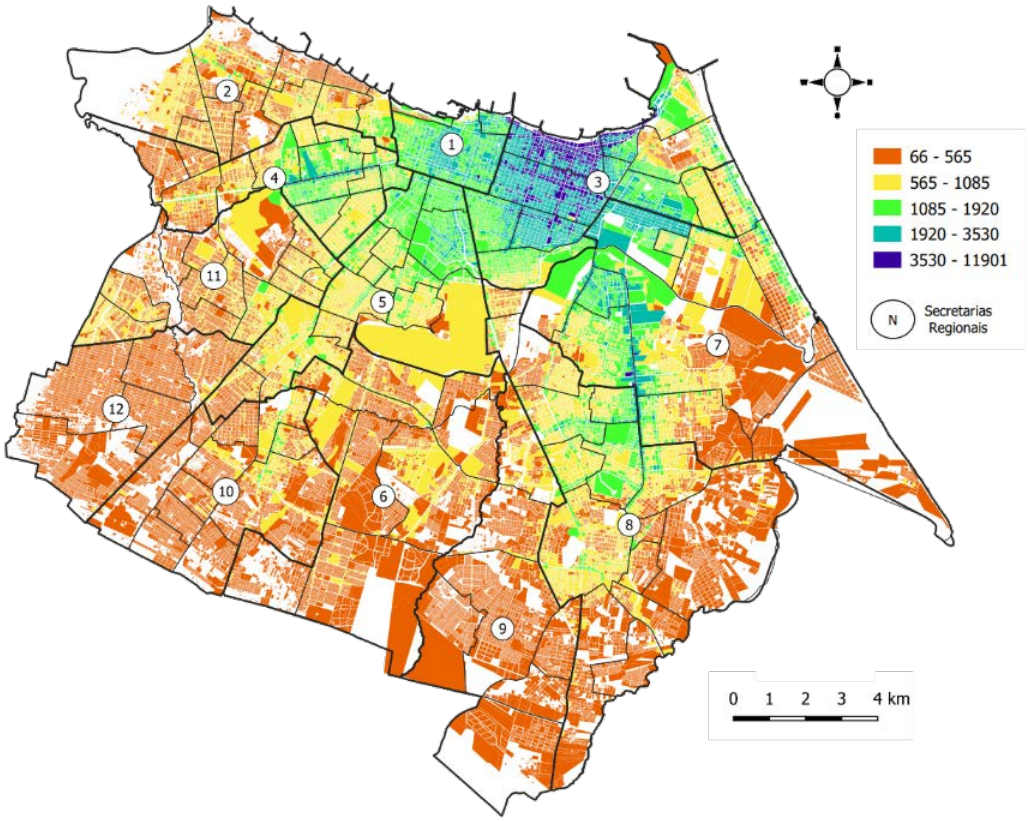
## 5.6 Proposição de uma PGV para terrenos urbanos de Fortaleza

Nesta seção, a partir do algoritmo vencedor, XGBoost, propõe-se uma PGV para os terrenos do Município de Fortaleza. De acordo com o cadastro imobiliário da SEFIN, a quantidade de lotes georreferenciados participantes do lançamento do IPTU 2020 é de 374.095. Portanto, far-se-á a predição da parcela territorial de todos esses

[49] As estatísticas para XGBoost, florestas aleatórias e MQO foram 14040932, 15257686 e 11447853.

lotes, cujo resultado pode ser visto através do Mapa 6. Vale ressaltar que os valores foram preditos com a origem da informação sendo ITBI (mais conservador) e o ano da informação para 2019.

**Mapa 6 - Predição dos valores de mercado da parcela territorial dos imóveis de Fortaleza pelo algoritmo XGBoost.**

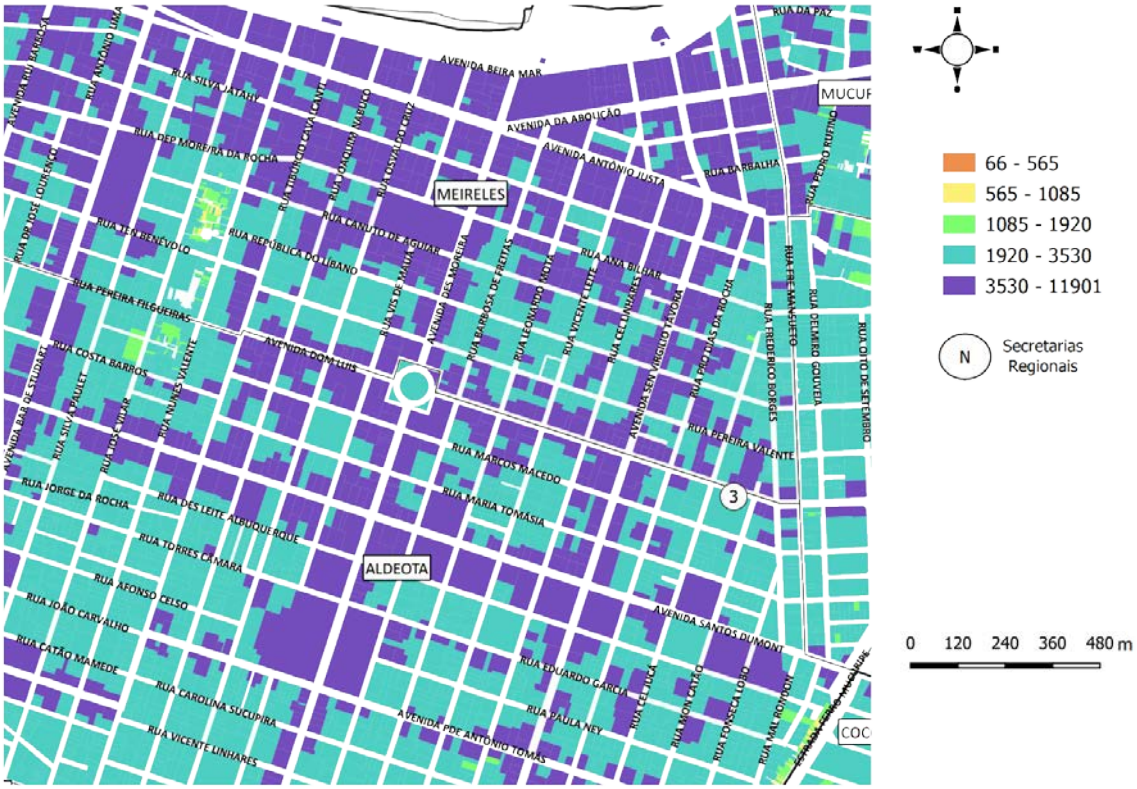



Fonte: elaboração do autor a partir dos dados da pesquisa.
Notas:
Considerado o valor de origem da informação como ITBI (e não transação).
Os valores se referem ao ano de 2019.

Pelo Mapa 6 também se observam os seguintes destaques: **i)** as avenidas Santos Dumont, Washington Soares e Bezerra de Menezes são indutoras de valorização da terra; **ii)** a regional 3 possui os terrenos mais valorizados, e quanto mais perto da Av. Beira Mar, maior o valor; **iii)** toda a região oeste da cidade, região sul e sudeste concentram os valores de terra mais baixos e **iv)** os bairros da Praia do Futuro apresentam entre R$ 1.000/m² e R$ 2.000/m².

O Mapa 7 mostra um detalhe do Mapa 6 nas regiões dos bairros mais valorizados. Meireles e Aldeota:

**Mapa 7 – Detalhe da predição dos valores de mercado da parcela territorial dos imóveis de Fortaleza pelo algoritmo XGBoost nos bairros Meireles e Aldeota.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

Através da Tabela 8, se observa que as regionais 01 e 03 têm as maiores média de valores, as regionais 09, 10 e 12, as menores médias. A regional 03 tem o maior valor unitário (no bairro Meireles, na quadra da Av. Abolição por conta do dado atípico na amostra) e a maior dispersão deles. As regionais 07 e 08 apresentam valores máximo elevados por conta dos lotes com frente para a Av. Washington Soares.

**Tabela 8 - Estatística descritiva da predição dos valores unitários das parcelas territoriais do Município de Fortaleza pelo XGBoost.**

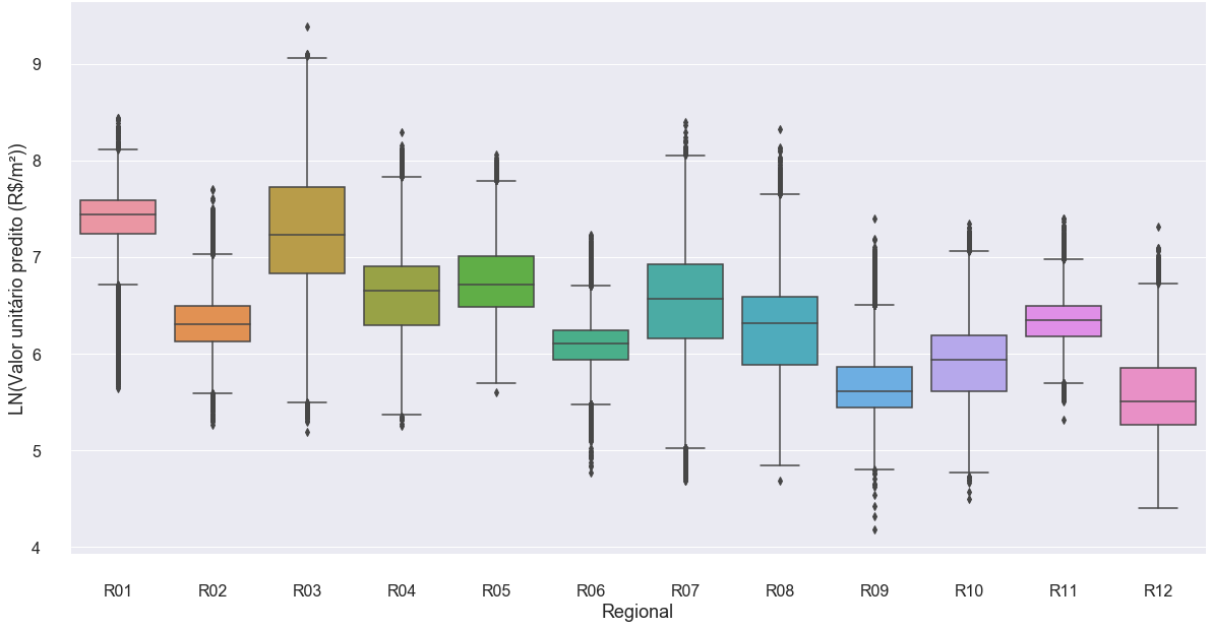
| Reg | N | Média | Desv. Pad | Mín. | 25% | 50% | 75% | Máx. |
|---|---|---|---|---|---|---|---|---|
| 01 | 11.349 | 1.680,49 | 560,33 | 283,26 | 1.397,91 | 1.700,86 | 1.983,80 | 4.646,07 |
| 02 | 41.465 | 574,98 | 168,91 | 192,21 | 459,42 | 544,72 | 659,69 | 2.227,36 |
| 03 | 29.080 | 1.673,21 | 1.071,21 | 179,84 | 930,27 | 1.380,39 | 2.276,98 | 11.901,20 |
| 04 | 36.028 | 818,14 | 359,26 | 190,53 | 541,33 | 778,73 | 1.002,63 | 4.019,89 |
| 05 | 35.166 | 933,54 | 397,20 | 271,17 | 652,05 | 828,14 | 1.102,56 | 3.165,48 |
| 06 | 32.150 | 455,40 | 129,84 | 118,32 | 377,72 | 445,96 | 514,32 | 1.388,14 |
| 07 | 24.131 | 831,59 | 527,07 | 107,91 | 474,83 | 711,29 | 1.013,40 | 4.459,75 |
| 08 | 43.745 | 592,78 | 301,97 | 108,34 | 357,73 | 553,77 | 727,14 | 4.148,13 |
| 09 | 24.322 | 311,01 | 124,79 | 65,63 | 230,19 | 273,03 | 352,66 | 1.643,43 |

| Reg | N | Média | Desv. Pad | Mín. | 25% | 50% | 75% | Máx. |
|---|---|---|---|---|---|---|---|---|
| 10 | 31.139 | 395,01 | 149,85 | 89,91 | 273,67 | 378,35 | 489,68 | 1.551,79 |
| 11 | 35.330 | 582,01 | 147,30 | 203,02 | 481,99 | 568,08 | 665,08 | 1.633,76 |
| 12 | 30.190 | 281,99 | 119,62 | 81,16 | 192,60 | 246,79 | 346,47 | 1.504,24 |
| Tot. | 374.095 | 704,24 | 562,88 | 65,63 | 382,9 | 545,48 | 810,53 | 11.901,20 |

Fonte: elaboração do autor a partir dos dados da pesquisa.

O Gráfico 22 demonstra mais apropriadamente a dispersão dos dados entre as regionais e indica a necessidade de remoção do dado *outlier* na regional 03.

**Gráfico 22 - Boxplot da predição dos valores unitários das parcelas territoriais do Município de Fortaleza pelo XGBoost por regional.**




Fonte: elaboração do autor a partir dos dados da pesquisa.

Com exceção das regionais 01, 03 e 07, a dispersão dos valores é satisfatória. Essa alta dispersão pode ser explicada pela heterogeneidade socioeconômica dessas regionais. Por exemplo, na regional 01 temos o bairro Moura Brasil, de população de baixa renda, assim como o bairro Cais do Porto e Vicente Pinzon na regional 03 e Sabiaguaba a regional 07.

# 6 CONCLUSÕES

Neste trabalho, foi feita a avaliação em massa com a utilização de modelos de aprendizado de máquina aplicados aos terrenos urbanos do Município de Fortaleza. A amostra foi constituída de dados de 8.209 de terrenos no período de 01/01/2015 a 31/12/2019, colhidos através de OUV. Muitas das variáveis explicativas utilizadas foram obtidas através de técnicas de geoprocessamento com o cruzamento espacial da posição geográfica do terreno com outras camadas existentes no cadastro multifinalitário a fim de se incorporar o máximo de informação espacial aos dados.

Foi elaborada uma análise exploratória de dados para estudo do comportamento da variável resposta, preço unitário dos terrenos, com algumas variáveis qualitativas e quantitativas, por meio de testes não paramétricos de igualdades de medianas e análise de correspondência. A exploratória foi de suma importância para a validação do modelo final, testando as premissas empíricas de comportamento dos preços observados. Os dados originais da amostra foram divididos aleatoriamente em treinamento e teste na proporção de 80%-20% para todos os modelos comparados.

Em seguida, foi estimado um modelo clássico de regressão linear múltipla com MQO, tendo como variável dependente o preço unitário transformado na escala logaritmo natural, para servir de *benchmark* na comparação de desempenho com os modelos de aprendizado de máquina de florestas aleatórias e *XGBoost*. O modelo apresentou significância conjunta para todas as variáveis escolhidas. Todas estas foram significantes a 5%, excetuando a variável “zona_incorporacao_vertical”. Em relação à verificação de multicolinearidade, todas apresentaram fator de inflação da variância menor que 10, tendo o sinal de seu coeficiente de acordo com o esperado pela prática e análise exploratória inicial dos dados.

Para os modelos de aprendizado de máquina, foi escolhida a melhor combinação de hiperparâmetros por validação cruzada. Também foram equacionados a relação entre viés-variância, poder de generalização preditiva e o sobreajustamento.

O modelo de florestas aleatórias teve como conjunto de variáveis explicativas mais importantes para explicação do comportamento dos preços unitários dos terrenos obtido: o valor unitário (R$/m²) base do terreno para o lançamento do IPTU, referente ao ano 2014, IDHM do Bairro, índice de aproveitamento máximo

equivalente, renda e interação incorporação vertical. O índice de aproveitamento máximo equivalente se mostrou mais importante que o índice de aproveitamento básico equivalente, podendo ser um indicativo da maior explicação dos preços pelo potencial construtivo adicional através dos instrumentos urbanísticos do plano diretor. A importância da variável “renda”, ocupando a quinta posição, evidencia a forte correlação entre a renda e preço unitário do terreno, evidenciando resultados anteriores de outros trabalhos de avaliação imobiliária. Através dos gráficos de dependência parcial apresentados, foi possível observar diversas relações entre os atributos e a variável resposta, validando o comportamento inicial esperado para os preços de imóveis. Por exemplo, através destes gráficos observou-se que o preço unitário decai à medida que aumenta a distância do terreno à via principal mais próxima, bem como a dependência espacial deste preço com a localização dos terrenos, representada pelas coordenadas planas UTM do centroide deles.

Com o modelo XGBoost, o conjunto de variáveis explicativas mais importantes para explicação do comportamento dos preços unitários dos terrenos foram: “valor_m2_terreno_face_quadra_iptu_2014”, índice de aproveitamento máximo equivalente, “idhm_2010_bairro” e “idh_educ” e “renda” (este último ocupou a quinta posição em ambos modelos de aprendizado de máquina). Também foi possível observar a relação positiva e não linear entre o "valor_m2_terreno_face_quadra_iptu_2014" e o preço unitário do terreno. O preço unitário variou com a origem da informação, de modo que pode ser um indício que as avaliações de ITBI são inferiores aos valores declarados, que por sua vez são inferiores aos valores de oferta. E análogo ao modelo de floresta aleatória, foi observado a dependência espacial do preço com a localização dos terrenos.

Para testar se a diferença de performance entre os modelos é estatisticamente significante, utilizou-se o teste de postos sinalizados de Wilcoxon, que rejeitou a hipótese nula de medianas dos resíduos iguais entre todos os pares de modelos. O modelo XGBoost apresentou o melhor desempenho em todos as métricas avaliadas. O modelo de florestas aleatórias ficou numa situação intermediária, melhor do que o modelo MQO, também em todas as métricas. Verificou-se ainda que os modelos de aprendizado de máquina são os únicos com predições em conformidade com a Portaria 511/09 do Ministério das Cidades e a norma do IAAO.

A partir modelo do XGBoost, foi proposta uma PGV para todas as parcelas territoriais do Município de Fortaleza. De todo o exposto, concluiu-se ser possível a

aplicação de aprendizado de máquina nas avaliações em massa para fins fiscais de terrenos urbanos, bem como a combinação desta com as técnicas tradicionais de preços hedônicos, mormente quanto à seleção das variáveis explicativas mais importantes para este modelo. Conforme demonstrado nessa pesquisa, os métodos de aprendizado de máquina não estão sujeitos a pressupostos rígidos de aplicação, conseguem explicar relações não lineares entre as variáveis explicativas e explicada e apresentam resultados finais compatíveis com o observado na dinâmica de preços do mercado imobiliário. Neste passo, espera-se o reconhecimento dessas técnicas emergentes nas normas de avaliações de imóveis urbanos da ABNT, principalmente, numa possível norma para avaliação em massa, desde que esta foque nos desempenhos das predições finais dos modelos e não nos detalhes dos métodos a serem utilizados, dada a sua dinamicidade de atualizações.

Como trabalhos futuros relacionados ao tema, sugerem-se **i)** aperfeiçoamento da modelagem MQO com a redução da quantidade de variáveis explanatórias com técnicas de análise de componentes principais ou análise fatorial; **ii)** utilização das florestas aleatórias com interpolação de seus resíduos através de geoestatística; **iii)** utilização da econometria espacial para considerar a correlação espacial dos terrenos em substituição ao modelo MQO; **iv)** aplicação de modelos modelos lineares generalizados e **v)** a aplicação de modelos de aprendizagem profunda (*deep learning*), dentre outros algoritmos emergentes do campo da ciência de dados.

# REFERÊNCIAS


ALMEIDA, Eduardo. **Econometria espacial aplicada**. Campinas, SP: Alínea, 2012.

AMORIM, William Nilson de. **Ciência de dados, poluição do ar e saúde.** 2019. 153 f. Tese (Doutorado) - Curso de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

ANSELIN, Luc. **Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity**. Geographical Analysis, [s.l.], v. 20, n. 1, p.1-17, 3 set. 1988. Wiley. http://dx.doi.org/10.1111/j.1538-4632.1988.tb00159.x. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1538-4632.1988.tb00159.x. Acesso em: 20 fev. 2020.

ANSELIN, Luc; REY, Sergio J. **Modern Spatial Econometrics in Practice**: a guide to GeoDa, GeoDaSpace and PySAL. Geoda Press LLC. Chicago, IL, 2014. 368 p.

ANTIPOV, Evgeny A.; POKRYSHEVSKAYA, Elena B.. **Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics.** Expert Systems With Applications, [s.l.], v. 39, n. 2, p.1772-1778, fev. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2011.08.077.

ARBIA, Giusepe. 2014. **A Primer for Spatial Econometrics With Applications in R**. London : Palgrave Macmillan, 2014. 230 p.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-1: avaliação de bens.** Rio de Janeiro, 2019.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-2: imóveis urbanos**. Rio de Janeiro, 2011.

BARROS, Luana. Fortaleza deve aumentar número de Secretarias Regionais para 12. Diário do Nordeste, Fortaleza. 2019. Disponível em: <https://diariodonordeste.verdesmares.com.br/editorias/politica/fortaleza-deve-aumentar-numero-de-secretarias-regionais-para-12-1.2137143>. Acesso em: 17 ago. 2019.

BONET, Jaime; MUÑOZ, Andrés; MANNHEIM, Carlos Pineda. **El potencial oculto:** factores determinantes y oportunidades del impuesto a la propiedad inmobiliaria en América Latina. Washington, D.C: Bid, 2014. 145 p.

BRASIL. Ministério das Cidades. Gabinete do Ministro. **Portaria nº 511, de 7 de dezembro de 2009.** Diretrizes para a criação, instituição e atualização do Cadastro Territorial Multifinalitário (CTM) nos municípios brasileiros. Diário Oficial da União, Brasília, DF, 08 dez. 2009.

BRASIL. Lei Complementar nº 10257, de 10 de julho de 2001. Regulamenta os arts. 182 e 183 da Constituição Federal, estabelece diretrizes gerais da política urbana e dá outras providências. **Estatuto da Cidade**. Brasília, DF, 10 jul. 2001. Disponível

em: <http://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10257.htm>. Acesso em: 02 jan. 2020.

BREIMAN, Leo. **Random Forests**. Machine Learning, [s.l.], v. 45, n. 1, p.5-32, 2001. Springer Science and Business Media LLC. http://dx.doi.org/10.1023/a:1010933404324.

BROWNLEE, Jason. XGBoost with python: gradient boosted trees with XGBoost and scikit-learn. [S.l.: s.n.], 2016. 115 p.

BRUNO, Artur; FARIAS, Airton de. **Fortaleza:** uma breve história. 2. ed. Fortaleza: Edições Demócrito Rocha, 2015. 264 p.

CARRANZA, Juan Pablo et al. **Random Forest como Técnica de Valuación Masiva del Valor del Suelo Urbano: uma Aplicación para la Ciudad de Río Cuarto, Córdoba, Argentina**. In: Congresso Brasileiro de Cadastro Técnico Multifinalitário e Gestão Territorial, 13., 2018, Florianópolis. Anais... Florianópolis: UFSC, 2018.

ČEH, Marjan et al. Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. Isprs International Journal Of Geo-information, [s.l.], v. 7, n. 5, p.168-170, 2 maio 2018. MDPI AG. http://dx.doi.org/10.3390/ijgi7050168.

CHEN, Tianqi; GUESTRIN, Carlos. **XGBoost: a scalable tree boosting system**. Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining - Kdd '16, [s.l.], p.1-13, 2016. ACM Press. http://dx.doi.org/10.1145/2939672.2939785.

CODES, B. N. Avaliação dos preços de imóveis na cidade de fortaleza, com a utilização de redes neurais artificiais, para a composição do ITBI. 2018. 78 f. Dissertação (Mestrado em Engenharia Civil)-Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Civil: Estruturas e Construção Civil, Universidade Federal do Ceará, Fortaleza, 2018.

DANTAS, Rubens A. **Engenharia de Avaliações, Uma Introdução à Metodologia Científica**. São Paulo: Pini, 1998.

DANTAS, Rubens A. **Modelos Espaciais aplicados ao mercado habitacional – um estudo de caso para a cidade de Recife**. Recife, 2003 (Tese – Universidade Federal de Pernambuco – UFPE).

DANTAS, Rubens A. **Prestação de serviços de assessoria na atualização da planta genérica de valores de Fortaleza.** Fortaleza: Dantas Engenharia, 2014. 51 p.

DE CESARE, Cláudia M.; CUNHA, Eglaísa Micheline Pontes. **Avaliação em massa de imóveis para fins fiscais:** discussão, análise e identificação de soluções para problemas e casos práticos**.** Brasília: Ministério das Cidades, 2012. 116 p.

DOANE, David P., SEWARD, Lori E. **Estatística Aplicada à Administração e Economia**, 4. ed. Porto Alegre: AMGH, 2014.

ERBA, Diego Alfonso. **El catastro territorial em América Latina y el Caribe**. Cambrige, Ma: Lincoln Institute Of Land Policy, 2008. 428 p. Disponível em: <https://www.lincolninst.edu/sites/default/files/pubfiles/el-catastro-territorial-america-latina-full.pdf>. Acesso em: 01 nov. 2019.

FLORÊNCIO, Lutemberg de Araújo. **Engenharia de avaliações com base em modelos GAMLSS.** 2010. 125 f. Dissertação (Mestrado) - Curso de Estatística, Universidade Federal de Pernambuco, Recife, 2010.

FORTALEZA. Fundação de Desenvolvimento Habitacional de Fortaleza (HABITAFOR). Prefeitura Municipal de Fortaleza. **PLHIS-FOR:** Plano Local de Habitação de Interesse Social de Fortaleza. 2010. Disponível em: <http://salasituacional.fortaleza.ce.gov.br:8081/acervo/documentById?id=fcd18692-a091-4677-ac71-346c5cff1010>. Acesso em: 11 fev. 2020.

FORTALEZA (Município). **Lei Complementar nº 62, de 02 de fevereiro de 2009**. Institui o Plano Diretor Participativo do Município de Fortaleza e dá outras providências. Fortaleza, CE, Disponível em: <https://portal.seuma.fortaleza.ce.gov.br/fortalezaonline/portal/legislacao/>. Acesso em: 27 jan. 2019.

FORTALEZA (Município). **Lei Complementar nº 236, de 11 de agosto de 2017**. dispõe sobre o parcelamento, o uso e a ocupação do solo no Município de Fortaleza, e adota outras providências. Fortaleza, CE. Disponível em: <https://portal.seuma.fortaleza.ce.gov.br/fortalezaonline/portal/legislacao/>. Acesso em: 27 jan. 2019.

FORTALEZA (Município). **Lei Complementar nº 278, de 23 de dezembro de 2019**. Altera dispositivos da Lei Complementar nº 0176, de 19 de dezembro de 2014, que dispõe sobre a organização e a estrutura administrativa do Poder Executivo Municipal e dá outras providências. Fortaleza, CE. Disponível em: <https://sapl.fortaleza.ce.leg.br/media/sapl/public/normajuridica/2019/12861/lc_278-2019.pdf>. Acesso em: 10 fev. 2020.

FORTALEZA. PREFEITURA MUNICIPAL DE FORTALEZA. **A Cidade.** 2020. Disponível em: <https://www.fortaleza.ce.gov.br/a-cidade>. Acesso em: 11 jan. 2020.

FORTALEZA. Secretaria Municipal de Desenvolvimento Econômico (SDE) (Org.). **Desenvolvimento humano, por bairro, em Fortaleza.** 2014. Disponível em: <https://en.calameo.com/read/0032553521353dc27b3d9>. Acesso em: 11 fev. 2020.

FRENTE NACIONAL DE PREFEITOS. **Anuário multicidades:** finanças dos Municípios do Brasil. Vitória: Aequus, 2020.

GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.** O'Reilly Media, Inc., 2017.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. MIT Press, 2016, disponível em http://www.deeplearningbook.org. Acesso em: 10 fev. 2020.

GUJARATI, Damodar N.; PORTER, Dawn C. **Econometria básica.** 5. ed. Porto Alegre: Amgh Editora Ltda., 2011. 924 p.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning:** data mining, inference, and prediction. 2. ed. Standford: Springer, 2008. 764 p.

HOFFMAN, Donna L.; FRANKE, George R.. **Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research**. Journal Of Marketing Research, [s.l.], v. 23, n. 3, p.213-227, ago. 1986. JSTOR. http://dx.doi.org/10.2307/3151480.

IBGE – Instituto Brasileiro de Geografia e Estatística. Cidades@. 2019. Disponível em: <https://cidades.ibge.gov.br/brasil/ce/fortaleza/panorama>. Acesso em: 10 fev. 2020.

INTERNATIONAL ASSOCIATION OF ASSESSING OFFICERS (IAAO). 2010. **Standards on Ratio Studies**. International Association of Assessing Officers: Kansas City, Missouri, US.

KAGGLE INC (Org.). **Your home for data science.** 2020. Disponível em: <https://www.kaggle.com/>. Acesso em: 15 fev. 2020.

LIAW, A.; WIENER, M. **Classification and regression by random Forest**. R News, 2002, 2, 18–22.

LONGLEY, Paul A. et al. **Sistemas e Ciência da Informação Geográfica**. 3. ed. Porto Alegre: Bookman, 2013. 540 p.

LUCENA, José Mario Pereira de. **O mercado habitacional no Brasil.** 1981. 376 f. Tese (Doutorado) - Curso de Economia, Instituto Brasileiro de Economia, Fundação Getúlio Vargas, Rio de Janeiro, 1981.

OLIVEIRA, Antônio A. F. de.; BANDEIRA, Sandro R. V.; SILVA, Carlyson V. A. **Estimativa de desempenho de métodos de aprendizado de máquina baseados em árvores de decisão frente à regressão múltipla na valoração do solo no Município de Fortaleza, Ceará.** In: Simpósio Brasileiro da Sociedade Brasileira de Engenharia de Avaliações, 8., 2018, João Pessoa. João Pessoa: Sobrea, 2018.

PAULINO, C. D. e SINGER, J.M. Análise de Dados Categorizados. São Paulo: Editora Edgar Blucher, 2006, 629p.

PAZOLINI, Tiago Umberto. **Observatório de valores imobiliários:** modelagem conceitual. 2019. 90 f. Dissertação (Mestrado) - Curso de Engenharia Civil, Engenharia de Transportes e Gestão Territorial, Universidade Federal de Santa Catarina, Florianópolis, 2019.

PEDREGOSA *et al.* **SCIKIT-LEARN:** machine learning in python, JMLR 12, pp. 2825-2830, 2011.

PROVOST, Foster; FAWCETT, Tom. **Data science para negócios.** Rio de Janeiro: Alta Books, 2016. 408 p. Tradução de : Data science for business.

RONAGHAN, Stacey. **The mathematics of decision trees, random forest and feature importance in scikit-learn and spark.** 2018. Disponível em: <https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>. Acesso em: 15 fev. 2019.

SEFIN. **Secretaria das Finanças do Município de Fortaleza.**

SIEGEL, S.; CASTELLAN Jr, N. J. **Estatística Não-Paramétrica para Ciências do Comportamento**, 2.ed. Porto Alegre, 2006, 448p.

SUROWIECKI, James. **The wisdom of crowds:** why the many are smarter than the few and how collective wisdom shapes business. New York: Anchor Books, 2004. 336 p.

WOOLRIDGE, Jeffrey M. **Introdução à econometria:** uma abordagem moderna. 6. ed. São Paulo: Cengage Learning, 2016. 848 p.

YOO, Sanglim; IM., Jungho; WAGNER, John E. Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. **Landscape And Urban Planning**, [s.l.], v. 107, n. 3, p.293-306, set. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.landurbplan.2012.06.009.