VALOREM
Revista da Sociedade Brasileira de Engenharia de Avaliações

ESTIMATIVA DE DESEMPENHO DE MÉTODOS DE APRENDIZADO DE MÁQUINA BASEADOS EM ÁRVORES DE DECISÃO NA VALORAÇÃO DO SOLO NO MUNICÍPIO DE FORTALEZA, BRASIL

Estimating the performance of machine learning ensemble methods versus multiple regression for predicting land value in Fortaleza, Brazil

Antônio Augusto Ferreira de Oliveira
http://orcid.org/0000-0003-3890-4219
Secretaria das Finanças do Município de Fortaleza (SEFIN), Fortaleza, Brasil
augusto.oliveira@sefin.fortaleza.ce.gov.br

Sandro Ricardo Vasconcelos Bandeira
http://orcid.org/0000-0003-2730-025X
Secretaria das Finanças do Município de Fortaleza (SEFIN), Fortaleza, Brasil
sandro.bandeira@sefin.fortaleza.ce.gov.br

Carlyson Viana Alencar Silva
http://orcid.org/0000-0000-0000-0000
Universidade Federal do Ceará (UFC), Fortaleza, Brasil
carlysonviana@gmail.com

RESUMO

Este trabalho apresenta três abordagens de aprendizado de máquina (machine learning - ML) com modelos ensemble de árvore de decisão (DT) aplicados à avaliação em massa de imóveis urbanos, mais precisamente, no auxílio à modelagem de uma planta genérica de valores de terrenos para o Município de Fortaleza. Como benchmark para a verificação de acurácia e análise de desempenho, comparou-se métricas já consagradas na literatura internacional com a mesmas obtidas pela modelagem de preços hedônicos (regressão linear múltipla com ajuste de superfície de tendência). Em quase todas as métricas escolhidas, com exceção do "nível de avaliação", os modelos alternativos apresentados superaram o modelo clássico de regressão. Ressaltou-se a simplicidade de utilização de tais modelos, na sua "liberdade" quanto à presença de multicolinearidade entre as variáveis preditoras e com destaque para o ranking dos atributos mais importantes na formação do valor de mercado. Ao final, conclui-se premente a necessidade de se elaborar uma norma específica para avaliação em massa de imóveis que contemplem as novas abordagens de aprendizado de máquina com critérios claros de verificação de acurácia e performance.

Palavras-Chave: Avaliação em massa; Modelos hedônicos; Aprendizado de máquina; Árvores de decisão; Random forest; Gradient boosting.

ABSTRACT

This article presents three machine learning algorithms (ensemble models) applied in mass appraisal of urban land values in Fortaleza. These models were compared against hedonic regression combined with trend surface analysis. On almost all metrics, with the exception of the "sales ratio", machine learning models outperformed the classical regression model. Such modes are very simple to use and do not require the verification of the perfect multicollinearity hypothesis and still give us the most important features on model to a good predictive market value. Finally, we concluded by the need to develop a specific standard for mass appraisal of real estate that include new machine learning approaches with clear criteria for verification of accuracy and performance.

Keywords: Mass appraisal of real estate; Hedonic regression; Machine learning; Decision trees; Random forest; Gradient boosting regression.

Preenchimento dos Editores

INFORMAÇÕES SOBRE O ARTIGO
Submetido em 15/04/2022
Publicado em 15/06/2022

Comitê Científico Interinstitucional
Editor-Responsável: Carlos Augusto Zilli
(SEER/OJS – Versão 3)

1. INTRODUÇÃO

A avaliação em massa de imóveis tem grande relevância na determinação da base de cálculo dos impostos patrimoniais de competência dos municípios, tais como o imposto predial e territorial urbano (IPTU) e o Imposto de Transmissão Inter Vivos e Cessão de Direitos a ele relativos (ITBI). Ademais, é bastante utilizada no cálculo das indenizações das desapropriações e na implantação de instrumentos de política urbana onde o valor da terra/edificação seja de conhecimento necessário, tais como a outorga onerosa do direito de construir (OODC), as operações urbanas consorciadas (OUC), IPTU progressivo no tempo, dentre ouros.

A NBR 14.653 quedou-se quase que silente quanto às avaliações em massa[1]. A bem da verdade, os critérios ali expostos estão no âmbito das avaliações individuais de imóveis urbanos, exigindo um grande esforço dos profissionais da engenharia de avaliação quando são solicitados a fazer trabalho nessa seara. Tal esforço se reflete na necessidade de inovar no estabelecimento de critérios técnicos de desenvolvimento, aferição e de performance dos valores de mercado projetados, os quais se consubstanciam numa planta genérica de valores imobiliários (PGVI).

Não obstante a NBR 14653-2:2011 trazer o anexo E (informativo) com as “Recomendações para tratamento de dados por redes neurais artificiais”, novas abordagens baseadas em aprendizado de máquina (machine learning) (ML) já despontam no mercado como técnicas bastante promissoras na extração de conhecimento a partir dos dados e formulação de predições sobre os mesmos. Isso se deve basicamente aos seguintes fatores: a) melhoria nos algoritmos de predição e classificação de dados; b) evolução tecnológica de hardware com o consequente aumento capacidade de processamento (inclusive com o uso de processamento paralelo em vários núcleos de CPU); c) abundância de informações disponíveis na internet (big data), dentre outros.

A avaliação em massa de imóveis, pela necessidade intrínseca de se lidar com uma amostra grande de dados para inferência dos valores de mercado de uma região ampla, tem muito a ganhar com as novas técnicas e algoritmos de ML.

Este trabalho tem por objetivo a apresentação de uma avaliação em massa para a determinação dos valores unitários de mercado dos terrenos do Município de Fortaleza com aplicação de algoritmos baseados em árvores de decisão, quais sejam: decision tree regression (com bagging e boosting) e random forest (florestas aleatórias). Como benchmark, utilizar-se-á a regressão linear múltipla com ajuste de superfície de tendência, conforme recomendações de DANTAS (2014), para comparação das técnicas. Na verificação de acurácia e análise de desempenho para as novas abordagens aqui propostas, utilizar-se-á as métricas propostas por ANTIPOV et al. (2012) e CEH et al. (2018).

Na literatura brasileira especializada, até a presente data, não encontramos trabalho de avaliação em massa de terrenos de um município comparando técnicas tradicionais como o aqui proposto, o que revela, de certa forma, o ineditismo dessa abordagem.

2. REFERENCIAL TEÓRICO

Na literatura internacional, encontramos os seguintes trabalhos onde já se vislumbram a utilização de técnicas de ML aplicados à avaliação em massa: **a)** ANTIPOV e POKRYSHEVSKAYA (2012) para avaliação de apartamentos na cidade de São Petesburgo, Rússia; **b)** YOO, IM e WAGNER (2012) para a seleção de variáveis influenciantes a serem usadas em modelos hedônicos (dados de residências unifamiliares no condado de Onondada, Nova York, EEUU) e **c)** CEH et al. (2018) com artigo comparando a performance da abordagem RF frente ao modelo hedônico tradicional (MQO) na cidade de Liubliana, capital da República da Eslovênia, nas avaliações de apartamentos daquela cidade.

2.1. MÉTODOS DE MACHINE LEARNING

Conforme mencionado, esse trabalho se utilizará de um conjunto de algoritmos de ML conhecidos como árvores de decisão (decision trees) *e* seus melhoramentos, sendo estes últimos conhecidos como ensemble methods. Desta feita, além do algoritmo decision tree (DT), utilizar-se-á os algoritmos ensemble: bootstrap agregation (bagging), random forest (RF) (“florestas aleatórias”) e gradient boosting.

[1] A NBR 14653-2:2011 (ABNT) só se refere à avaliação em massa em 3 dispositivos: a) item 7.3.5.3, b) anexo C, item C.1.2 e anexo D, item D.1.2.

Para a implementação dos diversos algoritmos se utilizou das bibliotecas open-source escritas em python, principalmente scikit-learn. Nas modelagens por regressão linear múltipla, utilizou-se também o software estatístico R.

2.1.1. Machine Learning (ML)

Segundo GÉRON (2017), ML, ou aprendizado de máquina em português, "é a ciência (e arte) de programar computadores de tal forma que eles possam aprender com os dados". É uma "espécie" do "gênero" inteligência artificial, da mesma forma que as redes neurais artificiais também o são, sendo esta técnica já bastante utilizada no campo da engenharia das avaliações.

2.1.2. Decision tree (DT) e Bagging

Decision tree (DT), ou árvore de decisão, é um método não paramétrico de ML usado tanto para problemas de classificação, como problemas de regressão. Essa técnica é também conhecida como CART (Classification and Regression Tree), tendo sido introduzida por Breiman et al. em 1984. Esse método se baseia na predição de valores alvos através de regras de decisão extraídas dos próprios atributos dos dados (PEDROGOSA et al., 2011). No caso de árvore de regressão, utilizada quando a variável alvo tem valor contínuo, tal qual o valor do imóvel a ser predito, a divisão dos dados nos diversos nós e ramos da árvore se dá de tal maneira a minimizar o erro quadrático médio (MSE - mean squared error) em cada ramo subdividido. O Anexo C traz um exemplo simples, mas explicativo do algoritmo DT, para a predição de valores unitários de terrenos em um loteamento na cidade de Fortaleza.

Métodos que melhoraram substancialmente a acurácia das predições de algoritmos baseados em árvores de decisão foram propostos inicialmente por HO (1995). Atualmente, são conhecidos na literatura como métodos ensemble. Partem do princípio da "sabedoria das multidões": é mais fácil acertar uma opinião com várias pessoas do que com um só "especialista", dado que este pode estar sujeito às suas próprias idiossincrasias (GÉRON, 2017). Ou seja, ao invés do algoritmo usar um único preditor, tem-se um conjunto de preditores não correlacionados, com suas predições servindo para extrair a predição final (a partir da média aritmética). Esses preditores podem ser provenientes de um único algoritmo, como também de uma combinação de algoritmos diferentes.

Bootstrap Aggregating, ou bagging, é um um algoritmo ensemble que usa a técnica de bootstrap. Esta consiste em dividir a amostra inicial (de treinamento) em subconjuntos aleatórios com reposição dos elementos escolhidos, executando o algoritmo de DT em cada subconjunto. Uma vez executado o algoritmo, a predição de um elemento não treinado se faz pela média de cada predição treinada. Com isso, temos uma sensível diminuição da variância, sendo esta bastante acentuada quando da utilização da técnica de árvore de decisão pura: uma simples variação nos dados de treinamento implica em uma alta variabilidade nos valores de predição.

2.1.3. Random Forest (RF)

A metodologia atualmente usada no algoritmo RF foi proposta por Breiman (2001) e consiste na divisão do nó de cada árvore da floresta com apenas um subconjunto de atributos escolhidos aleatoriamente. Esse último detalhe é o que difere RF de outros métodos ensemble, como o bagging, onde se utilizam todos os atributos para a melhor divisão possível. Segundo Breiman (2001) apud Liaw e Wiener (2002), essa estratégia torna o método RF mais performático que vários outros métodos consagrados, tais como análise discriminante, support vector machine (SVM) e redes neurais. Ademais, a correlação entre as árvores fica diminuída, acarretando na diminuição do overfiting, ou seja, quando se tem boas predições no conjunto de treinamento, mas não conseguindo o mesmo do conjunto de teste, indicando pobre poder de generalização.

Uma das causas para o sucesso de RF, além da sua alta taxa de acerto frente aos métodos tradicionais de análise de dados, inclusive inferência estatística, é a simplicidade na sua utilização. Basicamente, o analista só precisa especificar dois parâmetros: o número de árvores da "floresta" e o número de atributos aleatoriamente escolhidos na divisão do nó da árvore. LIAW e WIENER (2002) sugerem usar RF com 500 árvores e número de atributos aleatoriamente escolhidos em um terço (1/3) da quantidade total dos mesmos. Entretanto, no pacote scikit-learn, dentro da biblioteca para RF, existe uma biblioteca que otimiza automaticamente esses parâmetros (classe RandomizedSearchCV da biblioteca

sklearn.model_selection). Na amostra desse trabalho, a execução desse procedimento determinou 436 árvores (do total de 500 testadas) (hiperparâmetro n_estimators) e sete atributos escolhidos aleatoriamente na divisão de cada nó (hiperparâmetro max_features).

2.1.4. Gradient Boosting Regression

Gradient Boosting Regression (GBR) trabalha adicionando novos algoritmos preditores em cada etapa. O objetivo é tentar ajustar o novo preditor aos resíduos extraídos no passo anterior (na execução do preditor prévio) (GÉRON, 2013). Ao final, tem-se como predição a soma das predições do conjunto de preditores, de tal forma a minimizar os erros. A execução desse procedimento foi feita com 436 estágios (n_estimators) e sete atributos escolhidos aleatoriamente na divisão de cada nó (max_features).

2.1.5. Métricas de Desempenho e Acurácia

A fim de que se pudessem comparar os diversos modelos de predição de ML com a regressão linear múltipla pelo método dos mínimos quadrados ordinários (MQO), buscou-se utilizar no primeiro as mesmas variáveis[2], com pequenos ajustes, dentre eles: a) na eliminação das variáveis dummies, haja vista tais algoritmos não necessitarem dessa abordagem, o que diminui a quantidade de atributos necessários; b) na utilização direta (não transformada) de nenhum atributo e c) com a utilização direta das coordenadas planas UTM no datum SIRGAS 2000, x e y, com intuito de captar a influência espacial dos preços observados.

Cumpre por oportuno e relevante ressaltar, que o modelo de regressão linear com regressão linear múltipla e ajuste de superfície de tendência segue a metodologia proposta por DANTAS (2014), com pequenos ajustes, sendo o principal na utilização de um polinômio de 3º grau com as coordenadas dos centroides dos terrenos e suas interações.

Como medidas de desempenho e acurácia para comparação entre os diversos métodos adotaram-se as recomendações de Antipov et al. (2012) e Čeh et al. (2018). As seguintes métricas foram calculadas:

a) nível de avaliação (sales ratio) mediano (SR_m):

$$\mathrm{SR_m} = \text{mediana de } \frac{\text{preço predito}}{\text{preço real}} \tag{1}$$

b) coeficiente de dispersão (COD):

$$\mathrm{COD} = \frac{100}{\mathrm{SR_m}} \times \left(\frac{\sum_{i=1}^{n} |\mathrm{SR_i} - \mathrm{SR_m}|}{n} \right) \tag{2}$$

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

c) média percentual absoluta do erro (MAPE):

$$\mathrm{MAPE} = \frac{100}{n} \times \sum_{i=1}^{n} \left| \frac{\text{preço real}_i - \text{preço predito}_i}{\text{preço real}_i} \right| \tag{3}$$

Além das métricas acima, foram calculadas as seguintes por sua larga utilização na comparação entre modelos de ML:

d) raiz quadrada da média dos erros ao quadrado (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\text{preço real}_i - \text{preço predito}_i)^2}{n}} \tag{4}$$

[2] Utilizaremos a expressão atributos para se referir às variáveis independentes ou explicativas quando tratarmos de algoritmos de machine learning.

e) coeficiente de determinação (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\text{preço real}_i - \text{preço predito}_i)^2}{\sum_{i=1}^{n}(\text{preço real}_i - \text{preço real}_{\text{médio}})^2} \quad (5)$$

3. METODOLOGIA

3.1. ÁREA DE ESTUDO E DESCRIÇÃO DOS DADOS

O Município de Fortaleza é a capital do Ceará, com população estimada em 2,6 milhões de pessoas, segundo projeção do IBGE para o ano 2018, com área de 314,93 km², o que a torna extremamente adensada (7.786,44 hab/km² segundo o censo de 2010). De acordo com o cadastro imobiliário municipal da Secretaria das Finanças do Município de Fortaleza (SEFIN), possui aproximadamente 440 mil lotes (sendo cerca de 90 mil não edificados) e 777 mil inscrições municipais sujeitas à tributação do IPTU.

Os 18.584 dados para composição de amostra para esse trabalho foram disponibilizados pela SEFIN e representam: a) os preços de transações e ofertas obtidos do "observatório urbano de valores" (OUV) e b) as declarações de valores do imóvel nas guias de ITBI pelo contribuinte. Essa amostra forma uma estrutura de dados em painel correspondendo o período dos anos de 2009 a julho de 2018, cuja distribuição espacial se observa na figura abaixo:

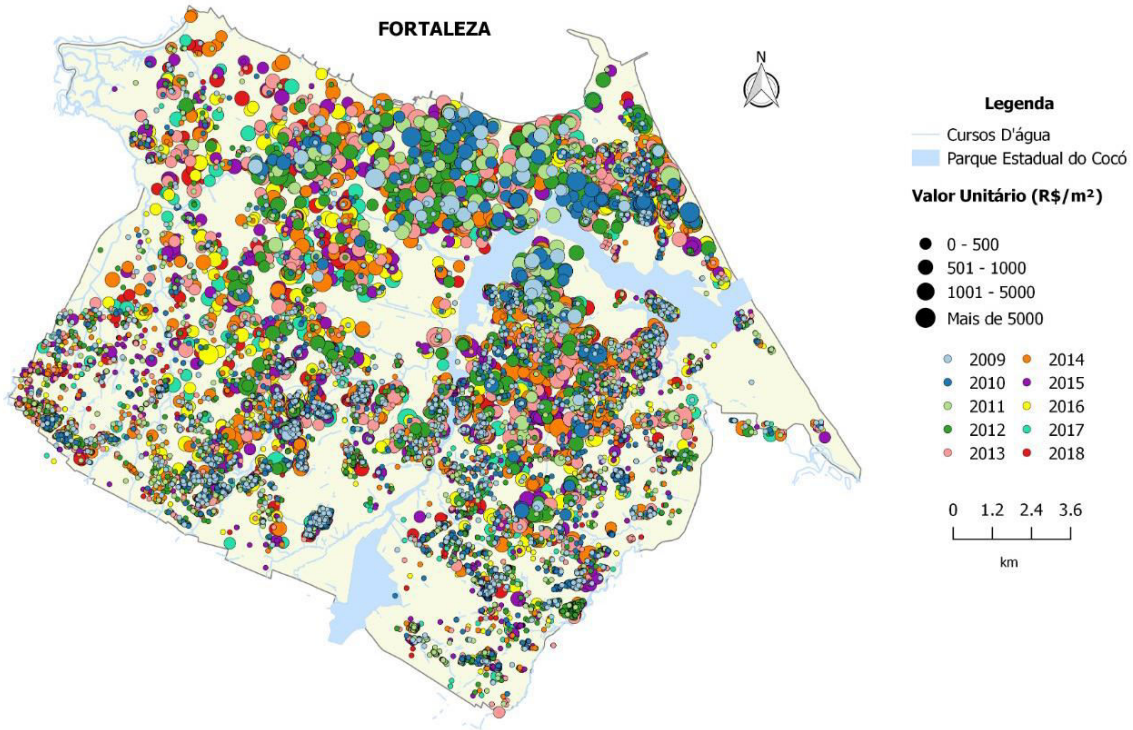

Figura 1 - Distribuição da amostra de terrenos (transações, ofertas e declarações de ITBI) do ano de 2009 a jul/2018 – 18.584 dados.

A descrição das variáveis explicativas utilizadas na regressão linear múltipla com ajuste de superfície de tendência encontra-se na tabela 4 do Anexo A. A listagem dos atributos para os algoritmos de ML encontra-se na tabela 5 do Anexo B.

3.2. AMOSTRAS DE TREINAMENTO/TESTE, ERRO DE GENERALIZAÇÃO E OVERFITTING

Um aspecto importante na avaliação de modelos utilizando-se de ML deve ser ressaltado: a divisão da amostra em treinamento e teste. Esse procedimento nos dá uma ideia de como se comportará o modelo ajustado nos novos casos a ele submetidos. Em termos de avaliação de imóveis, o quão preciso será nossa predição para o(s) imóvel(is) avaliando(s)? Em procedimentos de avaliação em massa, a resposta a essa pergunta é de suma importância, haja vista que a predição (projeção do modelo) se dará em uma enorme quantidade de dados.

A divisão da amostra em treino/teste executa o algoritmo apenas sobre os dados de treino. Uma vez ajustado o modelo nestes, procede-se a predição sobre os dados de teste. O erro medido no teste servirá de medida para o "erro de generalização" (generalization error ou out-of-sample error) (GÉRON, 2013). Ademais, a divisão treino/teste evita que o algoritmo "fotografe" os dados, com excelente poder de acerto sobre os dados a ele apresentados, mas com baixíssimo poder de explicação nos novos dados (overfitting).

Nesse trabalho, utilizamos a divisão 80% para dados de treinamento e 20% para dados de teste com escolha aleatória entre os mesmos, seguindo recomendação de GÉRON (2013). Não fizemos o mesmo para a aplicação do modelo de regressão múltipla (MQO), pois esse não é um procedimento comum no campo da engenharia de avaliações. Entretanto, recomendamos fortemente a adoção de tal prática pela vantagem de vislumbrarmos o futuro "erro de generalização" e podermos comparar com um valor normativo aceitável (quem sabe numa futura norma de avaliação em massa de imóveis urbanos...). Ressalte-se que apesar da divisão aleatória da amostra entre treino e teste, uma vez executada, esta passa a ser a mesma para todos os algoritmos utilizados, a fim de que a comparação não tenha nenhum viés.

4. RESULTADOS E DISCUSSÃO

4.1. MULTICOLINEARIDADE NA REGRESSÃO LINEAR E FATOR DE INFLAÇÃO DE VARIÂNCIA (VIF)

A avaliação em massa caracteriza-se pela escolha de uma grande quantidade de variáveis independentes, a fim de que se possa ajustar um modelo razoável apto a explicar à variabilidade dos valores de mercado de vários imóveis de uma região. Ocorre que a muitas dessas variáveis podem estar correlacionadas com as demais induzindo a uma "inflação na variância", ou seja, alargam "os intervalos de confiança para os verdadeiros coeficientes β_1, β_2, ..., β_k e tornando a estatística t seja menos confiável (DOANE e SEWARD, 2014). Não obstante a matriz de correlação abaixo mostra correlações baixas, quando tomadas as variáveis duas a duas, a tabela 1 mostra o fator de inflação de variância que mede a relação da variável com todas as demais.

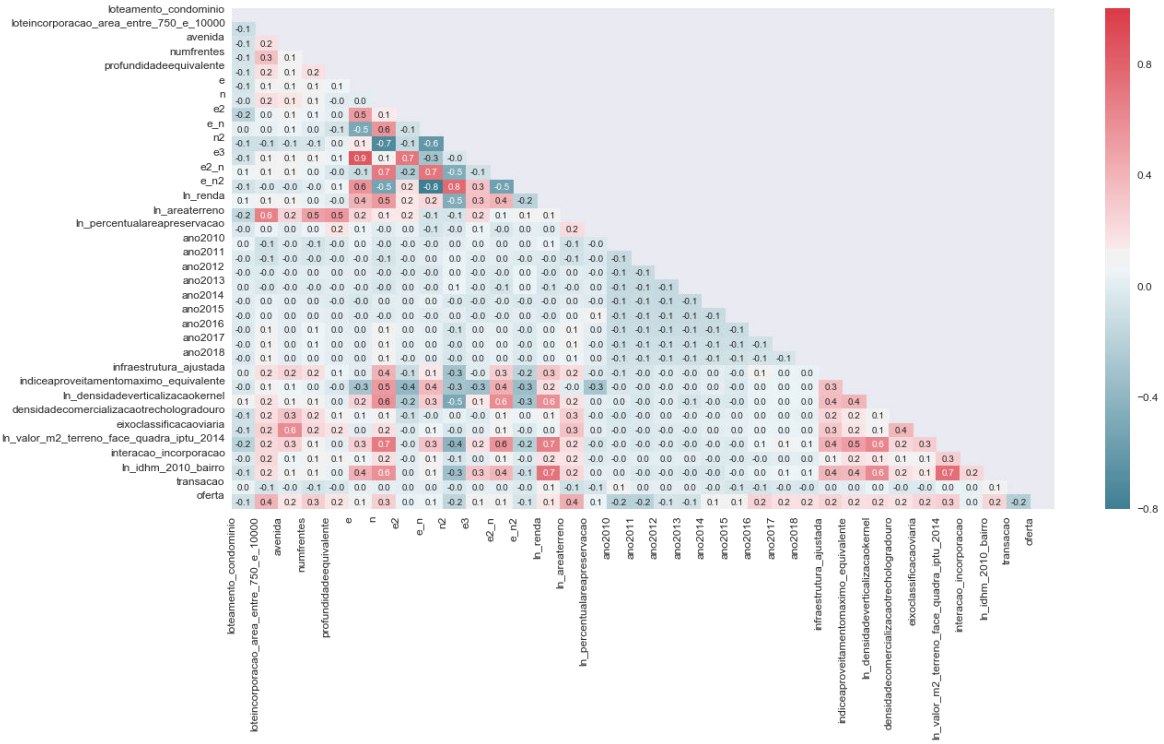

Figura 2 - Correlação entre os regressores no modelo de regressão linear múltipla com ajuste de superfície de tendência.

O VIF é calculado para cada regressor e pode ser definido como (DOANNE e SEAWARD, 2014): $\mathrm{VIF_j} = \frac{1}{1-\mathrm{R_j^2}}$, onde $\mathrm{R^2_j}$ é o coeficiente de determinação quando se faz a regressão do regressor $\mathrm{X_j}$ contra todos os demais regressores. Caso o preditor Xj não esteja correlacionado com os demais, seu valor de $\mathrm{R^2_j}$ será 0 (zero) e seu VIF será 1 (op. cit.). Segundo aqueles autores, regressores com VIF superior a 10 (dez),

podem indicar a remoção dos mesmos. Neste caso, verifica-se que seis variáveis tem inflação de variância forte: “areaterreno”, “percentualareapreservacao”, “valor_m2_terreno_face_quadra_iptu_2014”, “idhm_2010 _bairro”, “indiceaproveitamentomaximo_equivalente” e “infraestrutura_ajustada”. Como são extremamente importantes para explicar o comportamento do mercado de preços do mercado imobiliário, sugere-se a aplicação da técnica de redução de dimensionalidade, podendo ser a análise de componentes principais.

Tabela 1 - Fator de inflação de variância (VIF) para os regressores do modelo de regressão linear múltipla.
OBS: não consideradas as variáveis de polinômio de tendência do 3º grau.

Posição	**Variável**	**Fator de Inflação da Variância (VIF)**	**Interpretação (DOANNE e SEAWARD, 2014)**
1	ln_areaterreno	61.8	Inflação da variância forte
2	ln_percentualareapreservacao	45.3	Inflação da variância forte
3	ln_valor_m2_terreno_face_quadra_iptu_2014	41.6	Inflação da variância forte
4	ln_idhm_2010_bairro	19.8	Inflação da variância forte
5	indiceaproveitamentomaximo_equivalente	13.1	Inflação da variância forte
6	infraestrutura_ajustada	11.2	Inflação da variância forte
7	eixoclassificacaoviaria	9.7	Inflação da variância moderada
8	numfrentes	7.7	Inflação da variância moderada
9	ln_renda	6.0	Inflação da variância moderada
10	oferta	2.4	Inflação da variância moderada
11	ln_densidadeverticalizacaokernel	2.4	Inflação da variância moderada
12	ano2014	2.2	Inflação da variância moderada
13	ano2015	2.2	Inflação da variância moderada
14	ano2013	2.2	Inflação da variância moderada
15	ano2012	2.2	Inflação da variância moderada
16	ano2010	2.2	Inflação da variância moderada
17	profundidadeequivalente	2.2	Inflação da variância moderada
18	loteincorporacao_area_entre_750_e_10000	2.1	Inflação da variância moderada
19	ano2016	2.1	Inflação da variância moderada
20	ano2011	2.1	Inflação da variância moderada
21	densidadecomercializacaotrechologradouro	1.9	-
22	ano2017	1.8	-
23	avenida	1.8	-
24	ano2018	1.6	-
25	loteamento_condominio	1.3	-
26	interacao_incorporacao	1.2	-
27	transacao	1.2	-

Ressalta-se que nas abordagens alternativas apresentadas baseadas em árvores de decisão, não há que se preocupar na multicolinearidade dos atributos, nem muito menos qualquer dos demais pressupostos do modelo de regressão (GRÖMPING, 2009 apud YOO et al. 2012).

4.2. DESEMPENHO E ACURÁCIA

A tabela 2 abaixo apresenta as métricas definidas no item 2.1.5. Como supramencionado, os valores apresentados na coluna “Teste” indicam o “erro de generalização”, nos dando uma ideia como se comportará o modelo na predição final de novos dados a ele apresentados.

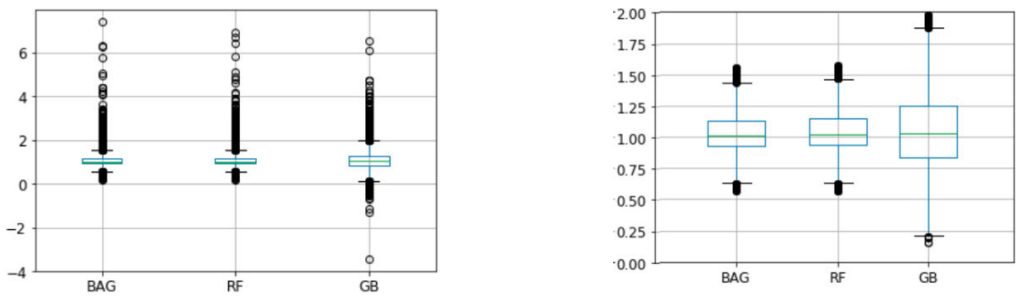

Figura 3 - Boxplot do nível de avaliação (SR) comparativo entre os modelos DT com ensemble. A figura da esquerda apresenta dados com outliers. A da direita com corte dos outliers (pelo multiplicador 1,5 do intervalo interquartil).

Tabela 2 - Comparativo de desempenho e acurácia entre os modelos testados nas amostras de treino e teste.

	Regressão Múltipla	Bagging		Random Forest (RF)		Gradient Boosting Regression (GBR)	
	Amostra	Treino	Teste	Treino	Teste	Treino	Teste
Nível de Avaliação (SR_m)	0,99	1,01	1,02	1,01	1,03	1,04	1,04
Média Perc. Abs. do Erro (MAPE) (%)	32,03	7,30	18,41	7,37	18,12	27,41	29,22
Coef. de Dispersão (COD) (%)	32,19	8,51	22,01	8,57	21,79	30,58	31,73
RMSE (R$/m²)	324,44	92,75	226,93	91,43	221,25	208,34	250,89
R^2	0,84 (lin.) 0,72 (ñ lin.)	0,98	0,85	0,98	0,86	0,89	0,82

Verifica-se que o algoritmo RF obteve o melhor resultado para as métricas MAPE, COD e RMSE, sendo estas intrínsecas ao “erro de generalização”. Os resultados do algoritmo bagging são bastante semelhantes ao RF. Bagging tornou-se vencedor apenas quanto ao nível de avaliação, pois obteve um valor ligeiramente menor de 1,02 frente a 1,03 de RF.

A Portaria 511 de 2009 do Ministério das Cidades, que estabelece as diretrizes para o cadastro multifinalitário, dispõe, no art. 30, §5º, que COD superior a 30% (trinta por cento) “indica falta de homogeneidade nos valores e a necessidade de atualização”. Apenas os algoritmos bagging e RF passaram pelo critério.

Todos os métodos baseados em árvores de decisão foram superiores no quesito poder de explicação do modelo, medido pelo R^2, se considerarmos essa medida no modelo de regressão múltipla na forma não linear (sem a transformação logarítmica da variável dependente).

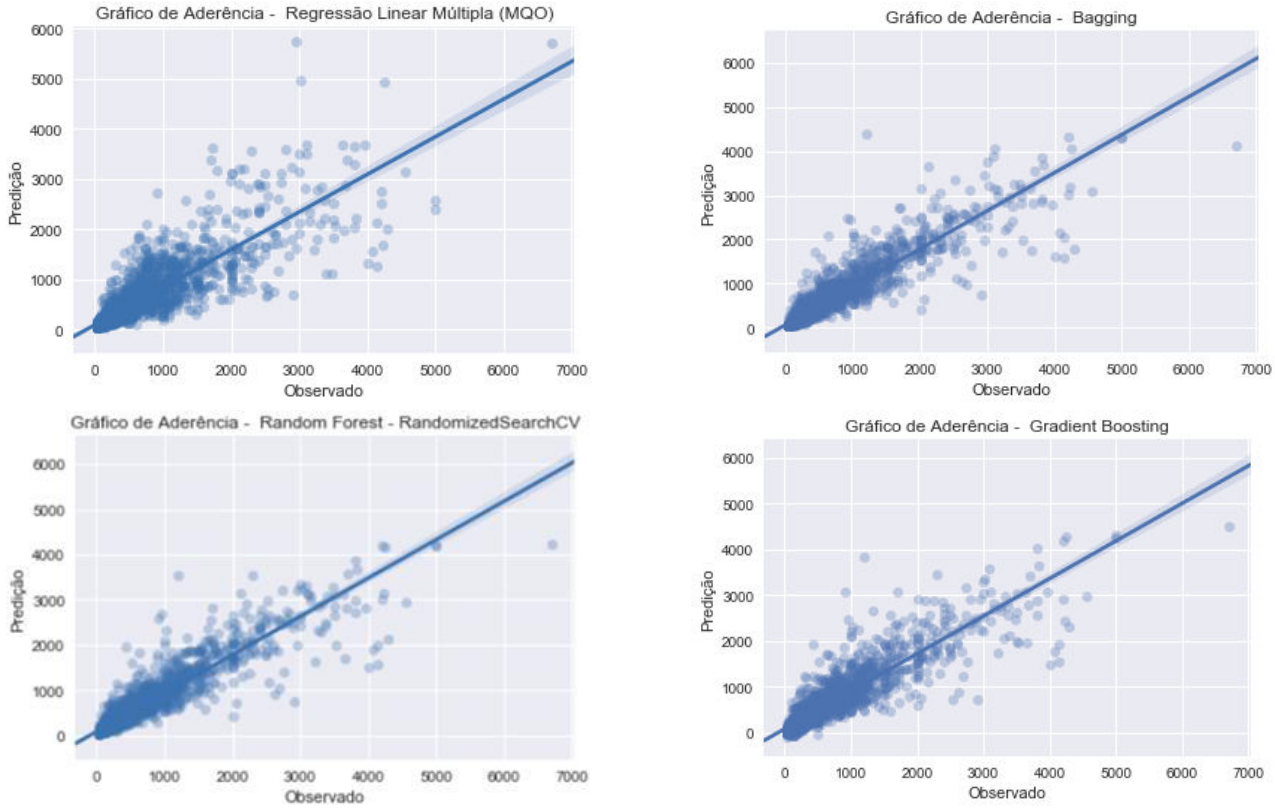

Figura 4 - Gráficos de aderência dos diversos modelos.

Pela análise visual da figura 4 acima, observa-se que os modelos com os algoritmos RF e bagging apresentam uma maior aderência dos valores preditos aos valores observados. Convém ressaltar, que o gráfico superior à esquerda representa a projeção dos valores do modelo de regressão linear múltipla sobre os mesmos dados de teste dos demais modelos de ML. Mesmo tendo tal modelo trabalhado com esses dados de testes no seu ajuste, haja vista que nessa abordagem não se separou a amostra de treinamento daquela, não conseguiu o mesmo uma melhor aderência.

4.3. IMPORTÂNCIA DAS VARIÁVEIS (FEATURE IMPORTANCE)

Um produto interessante da aplicação dos métodos baseados em árvore é a ordenação dos atributos por ordem de importância.

YOO, IM e WAGNER (2012) foram os pioneiros em aplicar a abordagem de ML para selecionar as variáveis mais importantes na modelagem hedônica.

A ideia subjacente à ordenação por importância é a de que atributos relevantes aparecem mais vezes nos nós iniciais da árvore, enquanto atributos menos importantes aparecem mais próximos às folhas (nós finais) (GÉRON, 2013). Isto decorre do próprio algoritmo que escolhe a divisão do nó baseado no atributo que minimiza a soma dos quadrados dos resíduos em cada ramo dividido.

A importância dos atributos pelo algoritmo RF se encontra na tabela 3 abaixo.

Tabela 3 - Importância dos atributos pelo algoritmo Random Forest.

	Atributo	Importância
1	valor_m2_terreno_face_quadra_iptu_2014	26%
2	densidadeverticalizacaokernel	15%
3	anodado	11%
4	idhm_2010_bairro	10%
5	origem_informacao	6%
6	y	6%
7	renda	4%
8	x	4%
9	indiceaproveitamentomaximo_equivalente	3%
10	areaterreno	3%
11	profundidadeequivalente	2%
12	densidadecomercializacaotrechologradouro	2%
13	eixoclassificacaoviaria	2%
14	interacao_incorporacao	2%
15	infraestrutura_ajustada	1%
16	numfrentes	1%
17	avenida	1%
18	gleba_lote_incorporacao	0%
19	loteamento_condominio	0%
20	percentualareapreservacao	0%
21	assentamentoprecarioareapercentual	0%

O atributo “valor_m2_terreno_face_quadra_iptu_2014” foi o mais importante. Isso indica que os valores base da planta genérica de valor anteriormente elaborada (em 2003, com valores corrigidos monetariamente até 2014) guardam forte pertinência com os valores atuais de mercado. Entretanto, sabe-se que a cidade se expandiu bastante em várias direções, mas principalmente ao longo da direção sudeste (eixo da Av. Washington Soares). Houve também o desenvolvimento de loteamentos em condomínios fechados, principalmente nos bairros Maraponga e Cajazeiras. Os valores dessa variável nesses locais são extremamente baixos, o que indica uma necessidade de análise mais acurada nessas regiões. Apesar da utilização de variável específica para captar essa última particularidade (variável “loteamento_condominio”), talvez por conta da pouca quantidade de dados de terrenos nessa condição, o algoritmo RF colocou-a na antepenúltima posição. Recomenda-se um modelo próprio para terrenos em condomínios.

O atributo “densidadeverticalizacaokernel” representa a medida de densidade de lotes com condomínios verticais com elevador, sejam residenciais ou comerciais. Seu cálculo foi feito com técnicas de sistema informações geográficas (SIG), qual seja, com a geração mapa de calor para a concentração de lotes naquela situação. Funciona como uma proxy espacial e indica regiões com alta densidade de incorporação imobiliária, o que eleva os preços de terrenos no entorno.

O aspecto temporal está contemplado no atributo “anodado” e assume valores inteiros de 2009 até 2018. Devido à variação dos preços observados ao longo do tempo, assumiu a 3ª posição em importância.

O atributo “origem_informacao” assume 3 (três) valores a depender da origem de coleta dos dados: 1) transação, 2) ITBI e 3) oferta. Obviamente, os preços observados variam a depender da origem, o que levou o atributo a ocupar a 5ª (quinta) posição.

As coordenadas UTM, “x” e “y”, se mostraram importantes, com destaque para a coordenada “y” ocupar uma posição de maior relevância em relação à coordenada “x”. Isso era esperado, pois na posição

norte da cidade está situada a região mais nobre da cidade, com decréscimo de índice de desenvolvimento humano à medida que se desloca para o sul. Não por acaso, o IDH bairro se mostrou bastante importante.

Terrenos com área de preservação ambiental e situados em assentamentos precários não são numerosos na amostra, o que também pode explicar a penúltima e última importância, respectivamente. Sobre a última posição ocupada pelo atributo “assentamentoprecarioareapercentual” cabe uma observação. Esse corresponde à área do terreno ocupada por assentamentos precários, segundo metodologia própria do IBGE para a sua determinação. A fim de que cada dado de terreno tivesse um valor para essa variável diferente de 0 (zero), já que a grande maioria dos terrenos não estão ocupados pelos mesmos, seria mais interessante a utilização da distância do terreno ao assentamento precário mais próximo. Desta feita, espera-se maior influência dessa variável.

YOO, IM e WAGNER (2012) citam a estratégia de Díaz-Uriarte e Alvarez de Andrés (2006) que sugerem eliminar 20% das variáveis menos significativas. Nesse trabalho, entretanto, como se observou acima, as mesmas representam situações específicas de mercado próprios, o que exige modelos próprios. Portanto, não basta apenas eliminar, mas modelar cada situação independentemente. Por exemplo, ao invés de termos um modelo genérico para terrenos, poderíamos dividir em nas seguintes modelagens: a) lotes padrões, b) lotes situado em zonas de incorporação, c) terrenos sujeitos a restrições ambientais, d) glebas, e) terrenos invadidos ou localizados em assentamentos precários dentre outros.

5. CONCLUSÕES

Esse trabalho procurou apresentar a abordagem de aprendizado de máquina (ML) com modelos ensemble de árvore de decisão (DT) aplicados à avaliação em massa de imóveis urbanos, mais precisamente, no auxílio à modelagem de uma planta genérica de valores de terrenos para o município de Fortaleza, Ceará.

As avaliações em massa exigem uma quantidade enorme de dados para uma boa projeção de valores, e, por isso, representam um campo fértil na utilização de modelos de aprendizado de máquina.

Muito embora a NBR 14653-2:2011 traga metodologias alternativas para a avaliação em massa, como a regressão espacial, e até mesmo a aplicação de redes neurais artificiais (RNA), entendemos que as inovações dos novos algoritmos de ML aqui apresentados se mostram mais fáceis de serem utilizados do que a RNA, pois a necessidade de tunning dos hiperparâmetros se mostra bastante reduzida em relação àquela.

Observou-se que o algoritmo random forest (RF) teve performance e acurácia superior ao modelo tradicional de preços hedônicos representado pela regressão linear múltipla com ajuste de superfície de tendência (com polinômio de 3º grau), bem como todos os demais apresentados.

Verificou-se também que não há necessidade de atendimento dos pressupostos do método dos mínimos quadrados ordinários (MQO) para a utilização dos modelos de árvore de decisão. Ficou patente, que se adotarmos modelos hedônicos, há a necessidade de utilização de técnica auxiliar para reduzir a multicolinearidade das variáveis explicativas com as demais, o que pode ser feito com a técnica de análise de componentes principais.

Mesmo que no primeiro momento, por lacuna normativa, o engenheiro avaliador não adote os valores de predição final através do aprendizado de máquina aqui proposto, pode este profissional se utilizar do ranking de importância das variáveis para a seleção das mesmas no seu modelo hedônico.

Não é praxe nas avaliações em massa a repartição da amostra em treinamento e em teste. Como foi apresentado, isso é comum nos algoritmos de aprendizado de máquina, pois permite inferirmos o “erro de generalização”, ou seja, como o modelo se comportará na predição de novos valores. Essa ideia deveria “ser importada” para a “engenharia de avaliações em massa”, com critérios claros de acurácia e precisão esperados para as métricas apuradas na amostra de teste, tais como, o nível de avaliação (SR), o coeficiente de dispersão (COD), a raiz quadrada da média dos resíduos ao quadrado (RSME) e escore R^2.

A bem da verdade, como toda inovação científica, mais estudos são necessários para formarmos uma base sólida de aplicação desses algoritmos para a engenharia de avaliações. Entretanto, não podemos nos furtar de quebrar os paradigmas que permeiam nosso campo de estudo. Urge que as futuras normas de avaliação de imóveis urbanos levam em consideração a tendência e o avanço da ciência da computação com incorporação dos métodos aqui apresentados, dentre outros que já fazem parte da comunidade acadêmica envolvida com data science e machine learning (ML).

REFERÊNCIAS

ANTIPOV, E.A.; POKRYSHEVSKAYA, E.B. **Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics.** Expert Syst. Appl. 2012, 39, 1772–1778.

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.

ČEH, M.; KILIBARDA, M.; LISEC, A.; BAJAT, B. **Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments.** ISPRS Int. J. Geo-Inf. 2018, 7, 168.

DANTAS, Rubens Alves. **Prestação de serviços de assessoria na atualização da planta genérica de valores de Fortaleza**. 05 de abril de 2014.
Relatório de Atividades. Secretaria das Finanças do Município de Fortaleza.

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

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

HO, Tin Kam. **Random Decision Forests**. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

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

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

YOO, S.; IM, J.; WAGNER, J.E. **Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY**. Landsc. Urban Plan. 2012, 107, 293–306.

ANEXO A

Tabela 4 - Lista de variáveis explicativas usadas no modelo de regressão múltipla. Fonte: elaboração própria.

Nome da Variável	Descrição
loteamento_condominio	Variável que indica se o terreno está localizado em condomínio fechado (1) ou não (0)
loteincorporacao_area_entre_750_e_10000	Variável que indica se o terreno tem área entre 750 e 10.000m² (1 em caso de ter, 0, caso contrário)
avenida	Variável que assume valor 1(um) se o dado está situado em avenida ou rodovia e zero em caso contrário.
numfrentes	Variável que indica o número de frentes do terreno
profundidadeequivalente	É definida como resultante matemática da divisão da área total do imóvel pela sua frente efetiva e conjugado com esta, nos fornecerá a característica geométrica de um terreno.
e, n, e2,n2, e_n, e2_n, e_n2, e3, n3	Variáveis sempre mantidas na escala direta (sem transformação), representadas por um polinômio de tendência do terceiro grau, com objetivo de filtrar as variações dos preços a grande escala espacial, composto pelas coordenadas UTM, na projeção SIRGAS 2000, dos centroides dos imóveis (E e N, representando a longitude e latitude, respectivamente), padronizadas em termos da média geral dos lotes existentes na cidade (Xm, Ym) e convertidas para km, onde E = (X-550.173,35)/1000 e N = (Y-9.582.962,63)/1000.
renda	Variável de macrolocalização, representada pela renda média do chefe da família, em salários mínimos, ajustada a uma superfície de tendência construída pelo processo de *Krigeagem,* tomando-se como base os dados de renda média do responsável no setor censitário, divulgada pelo censo do IBGE (2010). Krigeagem elaborada por DANTAS, 2014.
areaterreno	Variável que indica sua área territorial (em m²).
percentualareapreservacao	Variável representando a área de preservação (ZPA1) que atinge o imóvel, segundo o plano diretor do Município de Fortaleza (PDPFor) (em m²). Em casos de zero absoluto, para não inviabilizar sua transformação logarítmica, devemos considerar 0,01.
2010 a 2018	Variáveis que assumem valor 1(um) se o dado corresponde a uma amostra oriunda do ano em questão ou zero em caso contrário.
infraestrutura_ajustada	Variável representando a soma dos elementos de infraestrutura, presentes em qualquer um dos trechos de logradouro para qual o imóvel tem frente, a saber: água, esgoto, galeria pluvial, sarjeta, iluminação pública e pavimentação. Essa variável foi ajustada da seguinte forma: a) valor 1 (um) se esta soma é menor igual a 3; b) 2 (dois) se esta soma é 4; c) 3 (três) se esta soma é 5 e d) 4 (quatro) se esta soma é 6.
indiceaproveitamentomaximo_equivalente	Variável que representa o índice de aproveitamento máximo equivalente (no caso de mais de uma zona cortando o lote, determina-se a ponderação das áreas) onde está situado o imóvel segundo plano diretor. Em casos de zero absoluto, para não inviabilizar sua transformação logarítmica, devemos considerar 0,01.
densidadeverticalizacaokernel	Variável que indica a concentração de lotes com condomínios verticais com elevador. Interpolação da densidade por kernel.
densidadecomercializacaotrechologradouro	Variável 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 no trecho de logradouro. Em casos de zero absoluto, para não inviabilizar sua transformação logarítmica, devemos considerar 0,01.
eixoclassificacaoviaria	Variável que assumiu valores: 1(um) para vias locais e paisagísticas; 2(dois) para vias coletoras; 3(três) para vias expressas e arteriais dos tipos I e II, conforme classificação viária do cadastro imobiliário (LUOS).
valor_m2_terreno_face_quadra_iptu_2014	Variável indicando o valor unitário (R$/m²) base do terreno para o lançamento do IPTU, referente ao ano 2014.
interacao_incorporacao	Interação entre as seguintes variáveis: areaterreno * indiceaproveitamento_equivalente * densidadeverticalizacaokernel * loteincorporacao_area_entre_750_e_10000
idhm_2010_bairro	Variável do índice de desenvolvimento humano do ano de 2010 referente ao bairro do imóvel.
transacao	Variável representando as TRANSAÇÕES do mercado imobiliário ou valor declarado pelo contribuinte nas declarações de ITBI quando essa declaração esteja dentro do limite de mais ou menos 5% do avaliado pelo setor de avaliação de imóveis da célula de gestão de ITBI da Secretaria Municipal das Finanças do Município de Fortaleza. Agrupada com OFERTA. Quando OFERTA e TRANSAÇÃO são iguais a 0 (ZERO) simultaneamente, indica que o dado é uma avaliação de ITBI.
oferta	Variável que indica se o dado foi colhido no mercado imobiliário como oferta (caso em que assume 1).

ANEXO B

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 3.237e+00  5.054e-02  64.045  < 2e-16 ***
loteamento_condominio                       5.234e-01  1.604e-02  32.638  < 2e-16 ***
loteincorporacao_area_entre_750_e_10000     6.822e-02  1.057e-02   6.453 1.13e-10 ***
avenida                                     1.131e-01  1.181e-02   9.577  < 2e-16 ***
numfrentes                                  2.919e-02  5.965e-03   4.893 1.00e-06 ***
profundidadeequivalente                    -2.741e-04  6.482e-05  -4.228 2.37e-05 ***
e                                           3.311e-02  2.080e-03  15.921  < 2e-16 ***
n                                           3.967e-02  2.007e-03  19.763  < 2e-16 ***
e2                                         -2.447e-03  2.243e-04 -10.910  < 2e-16 ***
e_n                                         3.211e-03  2.929e-04  10.962  < 2e-16 ***
n2                                         -5.985e-04  2.526e-04  -2.370   0.0178 *
e3                                         -1.787e-04  1.778e-05 -10.051  < 2e-16 ***
e2_n                                       -8.811e-04  4.043e-05 -21.792  < 2e-16 ***
e_n2                                        2.044e-04  4.109e-05   4.974 6.61e-07 ***
ln_renda                                    1.954e-01  6.890e-03  28.362  < 2e-16 ***
ln_areaterreno                             -3.678e-02  4.978e-03  -7.388 1.55e-13 ***
ln_percentualareapreservacao               -3.906e-02  5.071e-03  -7.703 1.39e-14 ***
ano2010                                     2.886e-01  1.264e-02  22.826  < 2e-16 ***
ano2011                                     6.030e-01  1.300e-02  46.387  < 2e-16 ***
ano2012                                     9.550e-01  1.276e-02  74.847  < 2e-16 ***
ano2013                                     1.191e+00  1.277e-02  93.306  < 2e-16 ***
ano2014                                     1.415e+00  1.311e-02 107.872  < 2e-16 ***
ano2015                                     1.494e+00  1.343e-02 111.280  < 2e-16 ***
ano2016                                     1.544e+00  1.432e-02 107.825  < 2e-16 ***
ano2017                                     1.562e+00  1.545e-02 101.104  < 2e-16 ***
ano2018                                     1.564e+00  1.717e-02  91.119  < 2e-16 ***
infraestrutura_ajustada                     3.653e-02  3.230e-03  11.311  < 2e-16 ***
indiceaproveitamentomaximo_equivalente      7.281e-02  8.900e-03   8.181 3.00e-16 ***
ln_densidadeverticalizacaokernel            1.459e-02  9.433e-04  15.462  < 2e-16 ***
densidadecomercializacaotrechologradouro    7.882e-02  1.303e-02   6.051 1.47e-09 ***
eixoclassificacaoviaria                     7.919e-02  6.415e-03  12.344  < 2e-16 ***
ln_valor_m2_terreno_face_quadra_iptu_2014   3.155e-01  6.848e-03  46.075  < 2e-16 ***
interacao_incorporacao                      6.312e-08  1.225e-08   5.151 2.62e-07 ***
ln_idhm_2010_bairro                         7.255e-02  1.333e-02   5.442 5.33e-08 ***
transacao                                   1.964e-01  9.431e-03  20.825  < 2e-16 ***
oferta                                      2.368e-01  8.975e-03  26.388  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4017 on 18548 degrees of freedom
Multiple R-squared:  0.8399,    Adjusted R-squared:  0.8396
F-statistic:  2781 on 35 and 18548 DF,  p-value: < 2.2e-16
```

Figura 5 - Resultado da regressão linear múltipla (RLM) com ajuste de superfície de tendência com polinômio de 3º grau. Saído do software R. Fonte: elaboração própria.

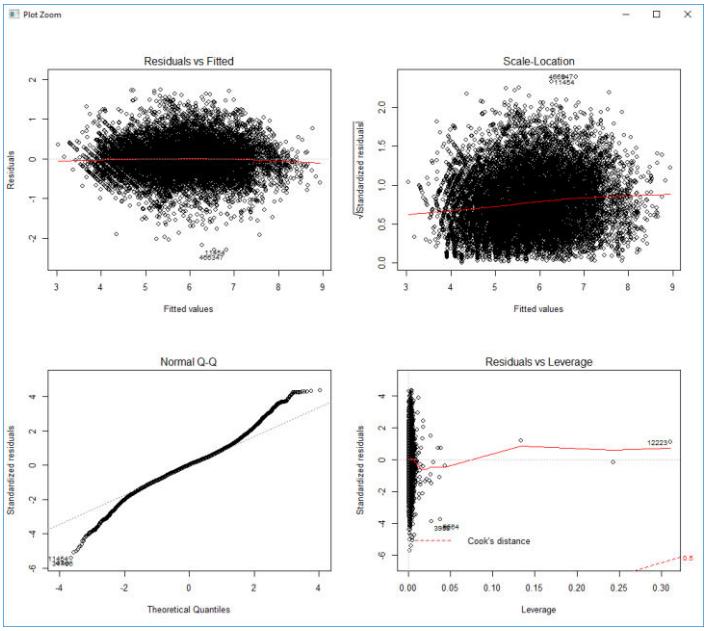

Figura 6 - Resultado gráfico da RLM com ajuste de superfície de tendência com polinômio de 3º grau. Saído do software R. Fonte: elaboração própria.

Tabela 5 - Lista de atributos usados nos modelos baseados em árvores de decisão

Variável	Desv.Pad.	Mínimo	Mediana	Média	Máximo
x	5.576	540.609	553.892	552.704	565.371
y	4.156	9.570.227	9.579.296	9.579.769	9.591.599
loteamento_condominio	0,21	0	0	0,05	1,00
gleba_lote_incorporacao[3]	0,43	0	1,00	1,17	2,00
avenida	0,33	0	0	0,12	1,00
numfrentes	0,58	1	1	1,26	10
renda	3,18	0,32	1,98	3,21	29,57
areaterreno	8.894,52	24,00	390,00	1.441,20	672.810,00
profundidadeequivalente	56,73	2,03	33,00	42,14	3.737,83
percentualareapreservacao	0,11	0,01	0,01	0,03	1,00
assentamentoprecarioareapercentual	24,22	0,01	0,01	6,33	100,00
anodado	2,62	2.009	2.013	2.013	2.018
infraestrutura_ajustada	1,11	1,00	3,00	2,97	4,00
indiceaproveitamentomaximo_equivalente	0,56	0,01	1,50	1,41	3,00
densidadeverticalizacaokernel	69,01	0,00	1,28	21,72	676,87
densidadecomercializacaotrechologradouro	0,26	0,01	0,07	0,18	1,00
eixoclassificacaoviaria	0,63	1,00	1,00	1,30	3,00
valor_m2_terreno_face_quadra_iptu_2014	72,85	2,05	22,02	42,29	1.146,86
interacao_incorporacao	260.770,12	0,01	0,01	30.797,62	16.236.007,50
idhm_2010_bairro	0,16	0,12	0,25	0,31	0,95
origem_informacao[4]	0,87	0	0	0,64	2,00
valorunitario	612,00	11,70	250,00	470,71	13.857,32

[3] Assume valores de 0 a 2, caso seja um lote, um lote com vocação de incorporação ou uma gleba.
[4] Assume valores de 0 a 2 caso seja dados de transação, ITBI ou oferta, respectivamente.

ANEXO C

Apresentamos a seguir um exemplo do algoritmo de árvore de decisão para a predição do valor de mercado dos lotes do loteamento “Cidade Ecológica”, em Fortaleza-CE, baseado em amostra de 20 (vinte) ofertas colhidas no interior do loteamento. Para fins de simplificarmos o entendimento, considerou-se como atributos informadores do valor observado apenas a testada e a área do terreno. Os dados colhidos em 2017 estão dispostos na tabela 6 abaixo:

Tabela 6 - Ofertas de terrenos colhidas em 2017 no Loteamento Cidade Ecológica. Identificação da divisão do algoritmo *decision tree* (DT) e suas respectivas predições (pela média de todos os dados no nó).

id	Testada (m)	Área (m²)	Valor Unit. (R$/m²)	Média da divisão
MI1912567	6,00	150,00	500,00	**516,67**
MI1912623	**6,00**	150,00	533,33	
MI1912143	**11,20**	360,00	272,22	**357,64**
MI1912564	13,40	360,00	361,11	
MI1912278	12,50	361,57	373,37	
MI1912140	12,00	364,17	356,98	
MI1912243	12,00	370,76	296,69	
MI1911050	12,00	370,76	404,57	
MI1912622	15,00	376,80	318,47	
MI1909982	15,00	**376,80**	477,71	
MI1912131	16,00	**400,00**	237,50	**277,27**
MI1910780	12,37	408,21	316,01	
MI1912501	17,48	531,86	259,47	
MI1898210	17,48	531,86	310,23	
MI1909821	42,98	550,00	272,73	
MI1911054	46,83	553,11	271,19	
MI1910540	15,12	568,81	290,08	
MI1909981	12,00	599,43	325,31	
MI1912084	14,24	970,38	288,55	
MI1912566	21,64	1.388,51	201,66	

O algoritmo DT tenta escolher no nó raiz, qual o atributo e qual o valor limite para o mesmo, onde a soma da média dos resíduos ao quadrado (MSE) nos dois ramos seja mínima. Esse procedimento continua indefinidamente até que se tenha o número mínimo de dados na folha (nó final), ou se tenha alcançado o número máximo de divisões estipulado, dentre outros critérios preestabelecidos (esses ajustes são os “hiperparâmetros” do algoritmo). Quando um novo valor é submetido à árvore de decisão, diversas comparações são feitas em cada nó, até chegarmos na folha correspondente. Nessa folha encontramos a média de todos os dados ali presentes.

A Figura 7 abaixo tenta explicar o funcionamento do algoritmo de maneira gráfica. Foi determinado para o mesmo, como hiperparâmetro, o nível máximo de “profundidade” igual a 2 (dois) (max_depth = 2).

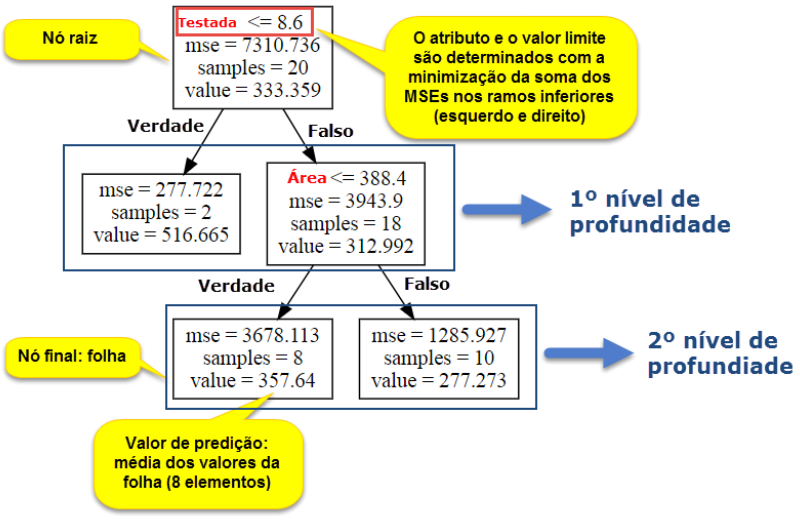

Figura 7 - Representação gráfica de uma árvore de decisão com profundidade máxima de 2 níveis. “*Samples*” representam o número de dados que satisfazem a condição de divisão do nó.

O primeiro atributo escolhido que minimiza MSE foi “testada” e o seu respectivo valor limite foi de 8,6m[5]. Nessa situação temos apenas 2 dados na amostra satisfazendo essa condição, o que leva à média de R$ 516,67/m² (média de R$ 500/m² e R$ 533,33/m²). Qualquer terreno cuja testada seja inferior a 8,6m terá como predição aquele respectivo valor. Caso o terreno tenha testada superior a 8,6m, adentramos no segundo nó, que determinou pela minimização do MSE, a área e o valor de 388,40m² como critério de divisão[6]. Observe na imagem abaixo que 18 dados correspondem a essa situação (total da amostra menos os 2 dados já classificados no critério anterior). Caso o terreno satisfaça essa condição (o que ocorre para 8 dados), o seu valor de predição será R$ 357,64/m², caso contrário, R$ 277,27/m² (com 10 dados).

A figura 8 apresenta outra árvore de decisão sobre os mesmos dados, mas com um nível de profundidade igual a 5 (cinco).

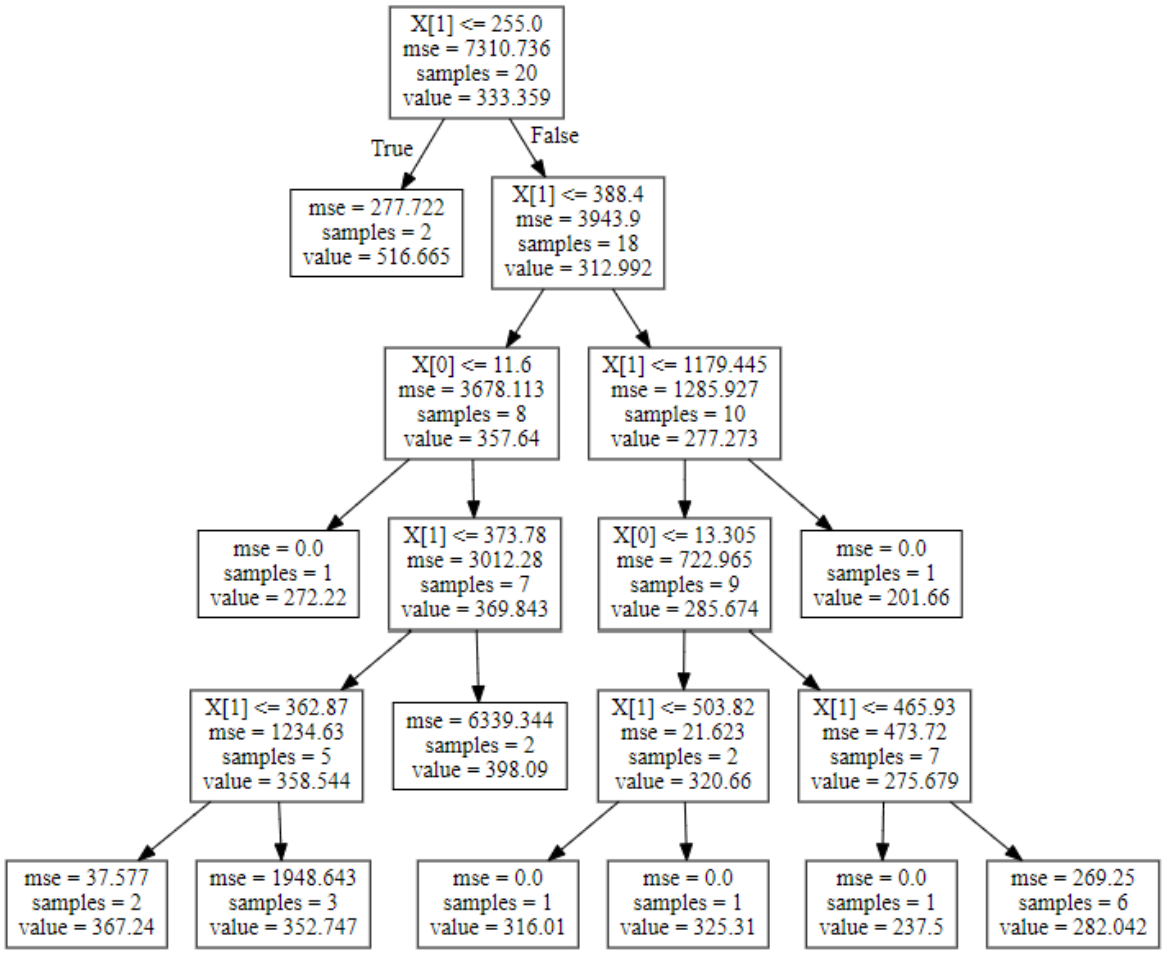

Figura 8 - Representação gráfica de uma árvore de decisão com profundidade máxima de 5 níveis.

[5] Esse valor é exatamente a média de testada entre os valores de 6,00 e 11,20 que correspondem aos último e primeiro valores observados de testada para a primeira divisão.
[6] Seguindo o mesmo raciocínio, observe que o valor de 388,40 é a média entre 376,80 e 400,00, valores extremos entre a segunda e terceira classe resultantes da segunda divisão.