15º Congresso de Cadastro Multifinalitário e Gestão Territorial
3º Encontro de Professores de Cadastro Territorial
COBRAC
7 a 9 de NOVEMBRO | 2022 | EVENTO VIRTUAL
Florianópolis - SC
Realização: UFSC
Através do: PPGTG Programa de Pós-graduação em Engenharia de Transportes e Gestão Territorial
GOTT GRUPO DE OBSERVAÇÃO E TRANSFORMAÇÃO DO TERRITÓRIO

AVALIAÇÃO DE IMÓVEIS PARA FINS TRIBUTÁRIOS COM USO DE MODELOS DE MACHINE LEARNING BASEADOS EM ÁRVORES

Property appraisal for tax purposes using tree-based machine learning models

Carlos Augusto Zilli
Instituto Federal de Santa Catarina (IFSC)
Departamento de Ensino, Pesquisa e Extensão (DEPE)
carloszilli@gmail.com

Lia Caetano Bastos
Universidade Federal de Santa Catarina (UFSC)
Centro Tecnológico (CTC) - Departamento de Engenharia
lia.c.bastos@ufsc.br

Rogério Cid Bastos
Universidade Federal de Santa Catarina (UFSC)
Centro Tecnológico (CTC) - Departamento de Engenharia
rogerio.bastos@ufsc.br

Resumo:
O imposto sobre a propriedade imobiliária é um importante instrumento de política urbana e tem como base de cálculo o valor venal do bem imóvel, normalmente calculado por meio de processos de avaliações em massa. O objetivo deste estudo é analisar o desempenho preditivo dos algoritmos *random forest* e *gradient boosting* na avaliação em massa de imóveis urbanos, comparando seus resultados com os obtidos pelo modelo clássico de regressão. Foram utilizados dados de apartamentos dos bairros Centro, Trindade e Agronômica, na cidade de Florianópolis, Santa Catarina, Brasil, dos quais 80% foram utilizados como dados de treinamento e 20% como dados de teste. Os resultados mostraram que o modelo *gradient boosting* apresentou o melhor desempenho em todas as métricas analisadas (RMSE, MAPE, COD e R^2), com predições mais precisas, o que confirma a perspectiva desta técnica *de machine learning* na previsão do valor venal de apartamentos, demonstrando ser essa uma alternativa viável para a geração da base de cálculo de impostos de base imobiliária.

Palavras-Chave: Engenharia de Avaliações; Avaliação de Imóveis; *Random Forest*; *Gradient Boosting*.

Abstract
The property tax is an important instrument of urban policy and is based on the market value of the property, normally calculated through mass appraisal processes. The objective of this study is to analyze the predictive performance of random forest and gradient boosting algorithms in the mass evaluation of urban properties, comparing their results with those obtained by classical linear regression. Data from apartments in the Centro, Trindade and Agronomic neighborhoods, in the city of Florianópolis, Santa Catarina, Brazil, were used, of which 80% were used as training data and 20% as test data. The results showed that the gradient boosting model presented the best performance in all analyzed metrics (RMSE, MAPE, COD and R^2), with more accurate predictions, which confirms the perspective of this machine learning technique in predicting the market value of apartments, demonstrating that this is a viable alternative for generating the tax base for real estate taxes.

Keywords: Real Estate Engineering; Property Appraisal; Random Forest; Gradient Boosting.

1 INTRODUÇÃO

Conforme a Constituição Federal (1988), o IPTU é um imposto municipal de base imobiliária calculado a partir do valor venal da propriedade urbana, que pode ter alíquotas diferenciadas, de acordo com a localização e a utilização do imóvel.

O valor venal dos imóveis, que serve como base de cálculo para o IPTU, precisa ser corretamente determinado e periodicamente atualizado, por meio de sistemas de avaliação em massa de imóveis. Entretanto, se esses sistemas são falhos, a base de cálculo do IPTU torna-se ineficiente e os contribuintes acabam tendo diferentes níveis de tributação efetiva, causando o inequidade de tributação.

A avaliação em massa de imóveis tem se tornado cada vez mais importante devido à grande participação do mercado imobiliário nas medidas econômicas, que se tornou um dos indicadores de desenvolvimento dos países (Yilmazer *et al*., 2020).

Essas avaliações têm grande relevância na determinação da base de cálculo dos impostos de competência dos municípios, sendo também bastante utilizada no cálculo de indenizações e implantação de instrumentos de política urbana.

Segundo a associação internacional de avaliadores (IAAO, 2013), a avaliação em massa é o processo de avaliação de um grupo de propriedades a partir de determinada data usando dados comuns, métodos padronizados e testes estatísticos.

Entre as técnicas comumente usadas para a avaliação em massa de imóveis está a regressão linear múltipla (e.g. Uberti *et al*., 2018; Faria Filho *et al*., 2019; Benjamin *et al*., 2020). Há os autores que fizeram uso da geoestatística (Hornburg e Hochheim, 2017; Theodoro *et al*., 2019; Duarte, 2019) e, em alguns casos, regressões não paramétricas foram aplicadas com sucesso (e.g. Filho *et al*., 2005).

Os algoritmos de aprendizagem de máquina, ou em inglês *machine learning*, subcampo da inteligência artificial, são técnicas que chamam atenção pela sua superior capacidade de predição frente as abordagens clássicas. Entre os métodos de aprendizado de máquina, os mais comumente utilizados para a avaliação em massa de imóveis são as redes neurais artificiais (McCluskey et al., 1999; Verikas et al., 2002; Pelli Neto, 2006; Selim, 2009) e, mais recentemente, o aprendizado de máquina baseado em árvore (Antipov e Pokryshevskaya, 2012; Ceh *et al*., 2019; Hong *et al*., 2019; Oliveira, 2020; Yilmazer e Kocaman, 2020).

Objetiva-se, com este estudo, aplicar as técnicas de *Random Forest* (RF) e *Gradient Boosting* (GB), algoritmos de aprendizado de máquina baseados em árvores, e estimar o desempenho destas técnicas na predição do valor de imóveis para geração de uma planta de valores genéricos (PVG) de apartamentos localizados nos bairros Centro, Agronômica e Trindade, em Florianópolis, Brasil.

Foram utilizados 225 dados de mercado dos três bairros em estudo, coletados entre os meses de março e abril de 2020, disponíveis em Zilli (2020). As predições foram avaliadas por meio dos indicadores desempenho RMSE, COD e MAPE.

A motivação e a justificativa para o desenvolvimento deste estudo decorrem da necessidade premente de se ter um método mais preciso e, ao mesmo tempo justo, de avaliação em massa de imóveis para fins fiscais.

Com o desenvolvimento desse estudo, pretende-se gerar conhecimentos que contribuirão para a solução de problemas relacionados à avaliação em massa de imóveis, como, por exemplo, a geração de plantas de valores genéricos de prefeituras.

2 MACHINE LEARNING

O aprendizado de máquina é um subcampo da inteligência artificial que utiliza técnicas de estatística e matemática associadas à tecnologia para compreender padrões (características) de um conjunto de dados. Segundo Samuel (1959), precursor e criador do conceito de *machine learning*, o aprendizado de máquina é um campo de estudo que possibilita aos computadores a capacidade de aprender sem ser programado de forma explícita. Conforme Murphy (2012), definimos o aprendizado de máquina como um conjunto de métodos que podem detectar automaticamente padrões nos dados e, em seguida, usar os padrões descobertos para prever dados futuros ou para realizar outros tipos de tomada de decisão sob condição de incerteza.

Desta forma, é possível concluir que o objetivo principal deste subcampo da IA é construir modelos computacionais que podem adaptar-se e, conforme Mitchell (1997), aprender a partir da experiência contida no respectivo conjunto de dados.

2.1 Algoritmos Baseados em Árvore

Entre os algoritmos de aprendizado de máquina comumente utilizados estão aqueles baseados em árvores de decisão (*decision tree*). Segundo Grus (2015), uma árvore de decisão usa uma estrutura de árvore para representar uma série de caminhos de decisão possíveis e um resultado para cada um dos caminhos. Conforme Grus (2015), as árvores de decisão dividem-se em árvores de classificação (com saídas categóricas) e de regressão (com saídas numéricas).

2.2 Random Forest (RF)

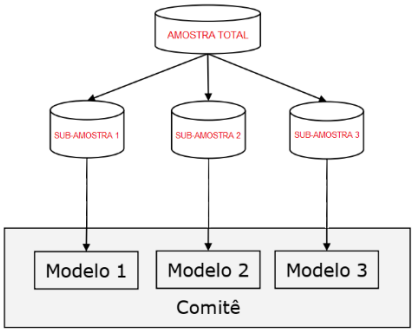
A técnica de *random forest* - *RF*, proposta por Breiman (2001), é a mais disseminada dentro do processo de modelos *ensemble* utilizando o método *bagging*.

Conforme James *et al*. (2013), na técnica *random forest* cria-se uma série de árvores de decisão a partir da amostra de treinamento e, ao construí-las, cada vez que uma divisão em uma árvore é considerada, uma amostra aleatória de p preditores é escolhida como candidatos à divisão do conjunto completo de m preditores. Deve-se tomar uma nova amostra de p preditores em cada divisão e, normalmente, escolhemos $p \approx m^{0,5}$ - o número de preditores considerados em cada divisão é aproximadamente igual à raiz quadrada do número total de preditores.

O efeito da aleatoriedade é reduzir a variância sem afetar o viés. Outro benefício disso é que não há necessidade de podar as árvores (MARSLAND, 2015). Entretanto, há outro parâmetro de difícil escolha, que é o número de árvores a colocar na floresta. No entanto, podemos resolver esse problema continuando a construir árvores até que o erro pare de diminuir. Liaw *et al*. (2002) sugerem usar 500 árvores e número de atributos aleatoriamente escolhidos em um terço (1/3) da quantidade total dos mesmos. Na biblioteca *random forest* do *software* R há pacotes que otimizam esses parâmetros.

Segundo Friedman *et al*. (2001), a ideia principal do *random forest* é reduzir a correlação de árvores de decisão do método *bagging*, sem aumentar muito a variância, através da seleção aleatória das variáveis de entrada. Desta forma, para cada árvore de decisão gerada, a média esperada de m árvores é a mesma esperada para qualquer uma delas e, portanto, somente ocorre a redução da variância. A Figura 1 mostra a técnica de *random forest*.

Figura 1: Método de Regressão *Random Forest*.



Fonte: adaptado de Mayrink (2016).

Uma vez que o conjunto de árvores é treinado, o resultado da floresta aleatória é a média aritmética de todas as predições individuais dadas por $y_1, ..., y_k$.

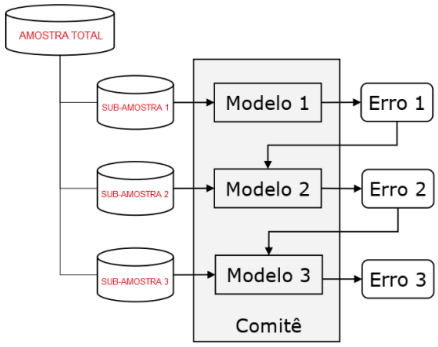
2.3 Gradient Boosting (GB)

O método *gradient boosting – GB* é utilizado para a resolução de problemas de classificação e regressão e consiste em uma série de combinações de modelos aditivos (modelos fracos), estimados iterativamente, resultando em um modelo forte.

Conforme Geron (2017), a técnica de *gradient boosting* 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).

O algoritmo *gradient boosting* consiste, portanto, em um processo iterativo aditivo em que o método inicia com uma previsão constante, cujo valor corresponde à média da variável de resposta na amostra de treinamento. Segundo Mayrink (2016), a cada iteração, um novo termo é adicionado ao modelo corrente, com o objetivo de reduzir gradualmente o erro de previsão. Assim, as atualizações são calculadas seguindo o sentido inverso do gradiente da função objetivo, em relação às aproximações correntes. O processo repete-se até que uma determinada condição de parada seja satisfeita, por exemplo, um número máximo de iterações. O esquema da Figura 2 mostra a técnica de *gradiente boosting*.

Figura 2: Método de Regressão *Gradient Boosting*.



Fonte: adaptado de Mayrink (2016).

Uma vez que o conjunto de árvores é treinado, o resultado da floresta pela técnica de *gradient boosting* é o próprio y_k. Tem-se, nesse caso, uma arvore de decisão final construída com ponderações dos resultados de cada uma das arvores anteriores.

2.4 Medidas de desempenho das modelagens

Como medidas de desempenho das diversas modelagens realizadas, utilizou-se as recomendações de Ceh *et al*. (2018) e os padrões estabelecidos pela norma IAAO (2013). Sendo assim, as seguintes métricas foram consideradas no presente estudo:

RMSE ⇒ A raiz do erro quadrático médio calcula a raiz da média dos erros do modelo ao quadrado, com isso, os valores maiores terão mais importância do que os menores, onde y_i corresponde ao valor unitário observado, $\hat{y}_i$ corresponde ao valor unitário ajustado e n é o número de observações. O cálculo é descrito pela Eq. 01:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad \text{Eq. 01}$$

MAPE ⇒ O erro médio absoluto calcula a porcentagem da média dos erros do modelo em valor absoluto, portanto é aplicada uma modulação à subtração. O peso de importância aos erros é dado de maneira linear. O cálculo é descrito pela Eq. 02:

$$MAPE = \frac{100}{n} \times \sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad \text{Eq. 02}$$

COD ⇒ O coeficiente de dispersão da mediana é o desvio médio expresso em termos percentuais do nível em que cada propriedade foi avaliada em relação à mediana do valor avaliado dividido pelo valor de mercado. O coeficiente mede a variabilidade (uniformidade) das avaliações e pode ser obtido pela Eq. 03 em que $\hat{R}$ é o nível de avaliação, do inglês *sales ratio*, de cada um dos imóveis individualmente e n é o número total de dados da amostra.

$$COD = \frac{100}{\hat{R}} \times \frac{\sum_{i=1}^{n}|R_i - \hat{R}|}{n} \qquad \text{Eq. 03}$$

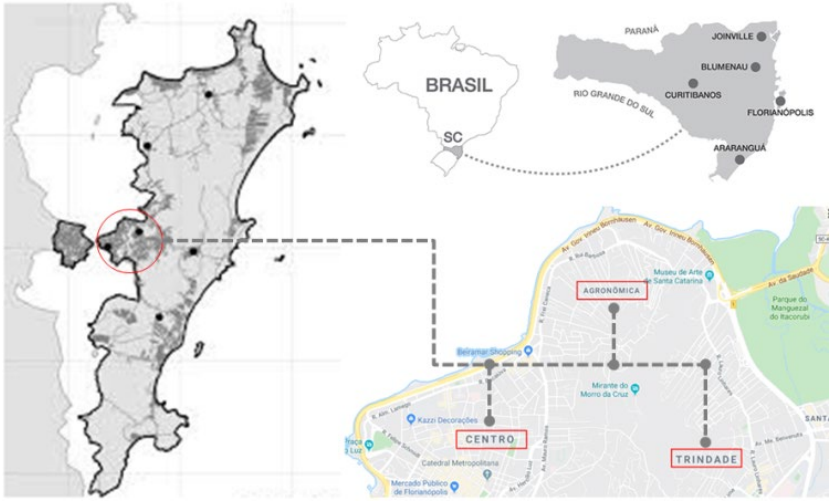
R2 ⇒ O coeficiente de determinação é uma métrica de avaliação que mede quão perto o valor esperado está do valor observado e pode ser descrito como a porcentagem de variação da variável resposta do modelo (Eq. 04), em que $\bar{y}$ corresponde ao valor unitário médio observado.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{Eq. 04}$$

3 ÁREA DE ESTUDO, MATERIAIS E MÉTODOS

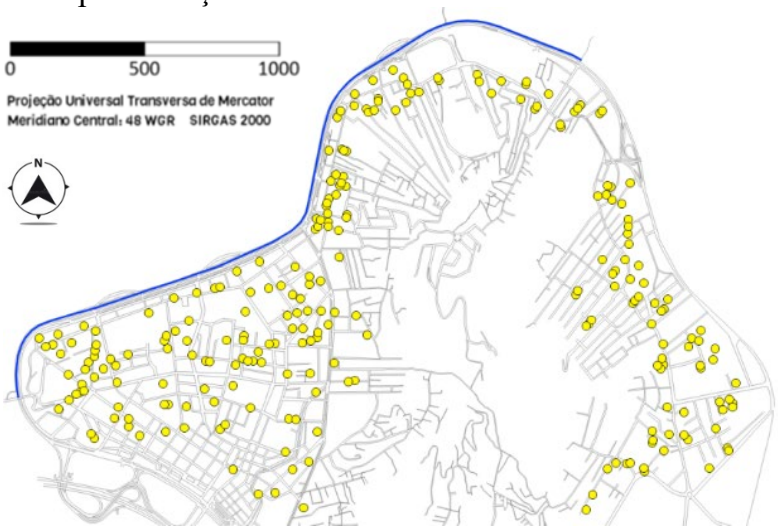
Esta pesquisa tem como área de estudo os bairros Agronômica, Centro e Trindade, no município de Florianópolis, estado de Santa Catarina, ao sul do Brasil. A área de estudo do qual foram extraídos os dados pode ser vista na Figura 3.

Figura 3: Localização dos bairros analisados em Florianópolis/Brasil.



Como material para essa pesquisa, utilizou-se dados de mercado coletados entre março e abril de 2020. Os dados foram tratados no *software* R® 3.5.3, onde se realizou a análise exploratória, geração de gráficos e realização de testes estatísticos. O *software* R® foi utilizado, ainda, para a modelagem da regressão clássica e dos algoritmos de *random forest* e *gradiente boosting*. Os 225 dados, separados por bairros, ficaram distribuídos em 107 no bairro Centro, 51 no bairro Agronômica e 67 no bairro Trindade. Todos os 225 dados, com as respectivas descrições, estão presentes em Zilli (2020) e podem ser espacializados, conforme a Figura 4.

Figura 4: Espacialização dos dados de mercado utilizados neste estudo.



Na Figura 4, os pontos amarelos representam a posição espacial de cada um dos 225 dados dos apartamentos coletados em Florianópolis. A linha em azul da figura (Avenida Beira-Mar Norte) representa um suposto polo de valorização. Para este estudo, foram consideradas as variáveis valor unitário (VU), em R$/m²; área privativa (AP), em m²; distância à Avenida Beira Mar (DM), em metros; número de dormitórios (ND), banheiros (NB) e vagas garagens (NG), em unidades; existência de piscina (PS), dicotômica de códigos 0 e 1; e padrão construtivo (PC), de códigos alocados do padrão baixo (1), médio (2) e alto (3). Com o objetivo de captar a influência espacial dos preços observados, utilizou-se a variável bairro, sendo Trindade (TR) e Agronômica (AG) os bairros adicionados ao modelo de regressão

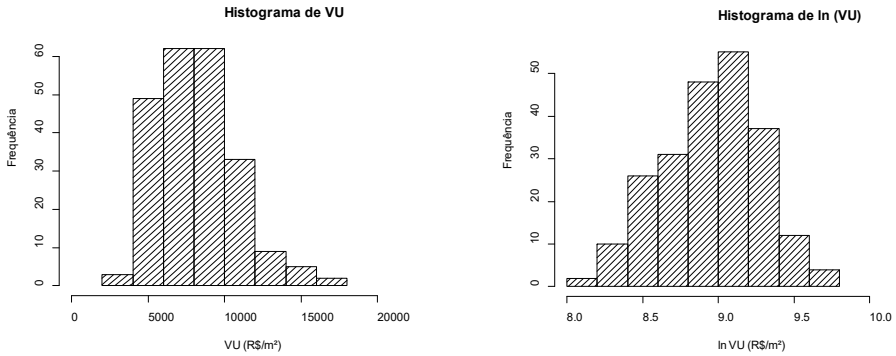
Visando comparar efetivamente os algoritmos de aprendizado de máquina com a regressão clássica, buscou-se utilizar em todas as modelagens as mesmas variáveis, na mesma escala e com as respectivas transformações. Para a regressão linear clássica utilizou-se os 225 dados de mercado disponíveis e, no processo de modelagem, foram descartados os pontos considerados *outliers*. Obteve-se, então, o melhor modelo de regressão linear clássica, que não violava nenhum dos pressupostos básicos da inferência estatística.

Com a amostra restante, excluindo-se os dados descartados na etapa anterior, realizou-se a divisão de 80% dos dados para treinamento e 20% dos dados para teste, com escolha aleatória entre os mesmos, seguindo recomendação de Geron (2017). 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.

4 RESULTADOS E DISCUSSÃO

A análise exploratória dos dados foi realizada tanto na variável explicada quanto nas variáveis explicativas, utilizando como ferramentas estatísticas os gráficos de transformação *Box-Cox*, dispersão, momentos, correlações e histogramas. Na Figura 5, pode-se ver os histogramas de frequência para a variável dependente VU em escala original e transformada.

Figura 5: Histogramas de frequências da variável valor unitário (VU e ln VU).



Pode-se verificar que na escala original (VU) os dados apresentavam leve padrão de assimetria positiva com curva platicúrtica, com os dados assimétricos para direita. Já, quando a variável valor unitário passa pela transformação logarítmica (ln VU), há uma correção, resultando em uma curva de assimetria levemente negativa e platicúrtica.

Fez-se a verificação de pontos influenciantes pelo método da Distância de Cook. Com relação aos *outliers*, considerou-se a variação de ± 2,5 desvios em torno da média por ser uma amostra consideravelmente grande, para dados imobiliários. Desta forma, quatro dados foram eliminados da amostra de trabalho, sendo eles: o ponto influenciante AP_026 (8.295,00) e os três *outliers* AP_115 (6.627,00), AP_131 (16.290,00) e AP_206 (14.070,00). Os modelos de regressão construídos neste estudo foram determinados com os 221 pontos restantes, sem esses quatro pontos e toda a amostra foi reorganizada para ser utilizada nas etapas seguintes.

4.1 Modelo Clássico de Regressão (MCR)

Para se conseguir o melhor MCR, que pudesse explicar o mercado imobiliário da forma mais fidedigna possível, realizou-se diversas simulações, com e sem transformações na variável dependente e, após excluir quatro dados da amostra, obteve-se um modelo que não violava nenhum dos pressupostos básicos da regressão. O modelo é apresentado na Eq. (05):

$$\ln(VU) = \beta_0 + \beta_1 \times (AP) + \beta_2 \times (DM) + \beta_3 \times (ND) + \beta_4 \times (NG) + \beta_5 \times (NG) + \beta_6 \times (PS) + \beta_7 \times (PC) + \beta_8 \times (AG) + \beta_9 \times (TR) \quad \text{Eq. 05}$$

Na Tabela 1, pode-se observar as estatísticas de cada um dos regressores do modelo MCR. Todos se mostraram significativos ao nível de significância de 10,0%, atendendo a situação mais desejável da norma de avaliações NBR 14.653-2 (2011).

Tabela 1: Estatísticas relativas aos parâmetros do modelo de MCR.

Variável	Coeficientes	Erro Padrão	Estatística t	Significância
Intercepto	8,43127	0,05575	151,222	0,00000
AP	– 0,00284	0,00046	– 6,21926	0,00000
DM	– 0,00011	0,00003	– 3,78418	0,00020
ND	0,05689	0,02100	2,70851	0,00731
NB	0,02039	0,01597	2,27665	0,08313
NG	0,16208	0,02109	7,68651	0,00000
PS	0,08523	0,02655	3,21025	0,00153
PC	0,21299	0,01662	12,8125	0,00000
AG	– 0,08141	0,02954	– 2,75558	0,00637
TR	– 0,11917	0,03175	– 3,75334	0,00023

Os sinais dos regressores confirmam a expectativa do mercado imobiliário local, tornando-os coerentes. O teste F de Snedecor mostra que o modelo foi significante ao nível de 1,0%. O teste de Jarque-Bera indica normalidade dos resíduos ao nível $\alpha = 5{,}0\%$ onde se obteve *p*-valor = 0,165. O teste de Breusch-Pagan indica homecedasticidade ao nível $\alpha = 5{,}0\%$, onde se obteve *p*-valor = 0,1036. A análise gráfica da variável valor unitário (ln VU) versus cada uma das variáveis explicativas, ambas em escala transformada, indica tendência linear nos dados. O teste de Inflação da Variância (VIF) teve seu valor máximo na variável área privativa ($VIF_{AP} = 5{,}738$), o que indica ausência de multicolinearidade. Gujarati *et al*. (2018, p. 348)

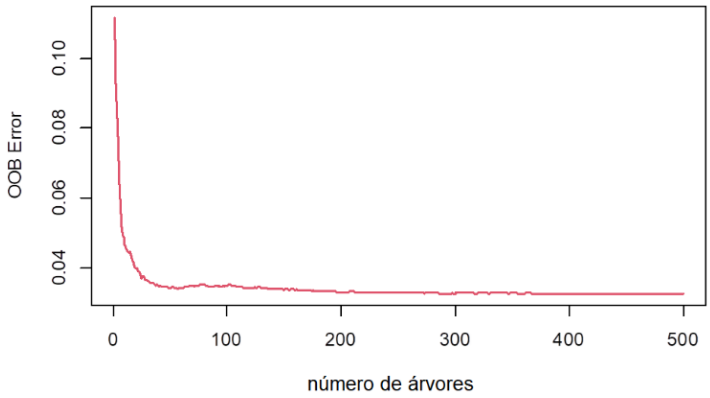
dizem que se VIF for superior a 10,0, a variável é tida como altamente colinear. Nesse sentido, todas as variáveis explicativas deste estudo tiveram sua colinearidade aceita.

Ressalta-se que nas abordagens alternativas apresentadas baseadas em árvores de decisão, não há que se preocupar na multicolinearidade dos atributos, muito menos com qualquer um dos demais pressupostos do modelo clássico de regressão (Gromping, 2009 *apud* Yoo *et al*. 2012). Verifica-se, portanto, que o modelo adotado não violou os pressupostos, sendo aprovado em todos os testes realizados, mostrando-se um modelo estatisticamente correto para explicar o mercado imobiliário da área em estudo.

4.2 Modelo de Random Forest (RF)

O modelo de *random forest* foi estimado considerando as mesmas variáveis adotadas no modelo de regressão clássica. Os 221 dados utilizados no modelo clássico foram divididos em 80% para treinamento e 20% para validação. Calculou-se o *out-of-bag error* (OOB) para o modelo *random forest*. Segundo Breiman (1996), esse erro é um indicador importante para medir o erro de estimativa de *random forest* usando *bagging* para reamostrar dados usados para treinamento. Esse indicador é utilizado para determinar o número ideal de árvores a serem usadas no modelo de regressão. O diagrama OOB é visualizado na Figura 6.

Figura 6: Diagrama *out-of-bag error* para modelagem *random forest*.



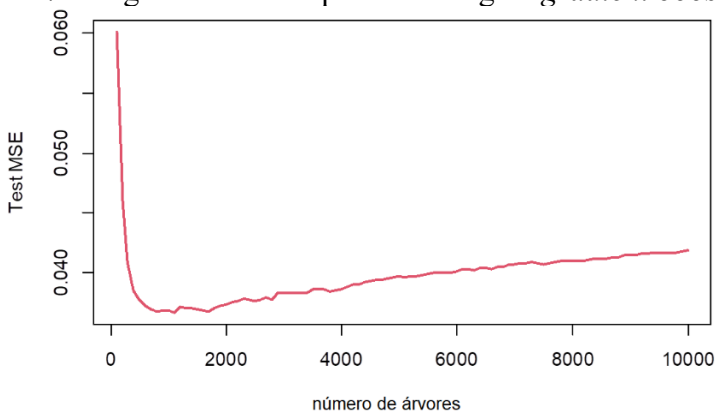
A Figura 6 mostra o gráfico do erro OOB obtido após a realização do RF. O valor da árvore onde o erro fica aproximadamente constante é $x = 50$. Por esta razão, foi adotado 50 árvores nesse algoritmo. Verifica-se que a partir desse número, o erro OOB mínimo é obtido. Se o número de árvores escolhido for maior que 50, o OOB não muda significativamente, mas o custo operacional aumentará. Desta forma, utilizando a biblioteca "randomForest" do *software* R®, elaborou-se a modelagem *random forest* considerando 50 árvores e $p \approx m^{0,5} \Rightarrow p \approx 9^{0,5} = 3$ variáveis em cada iteração.

4.3 Modelo de Gradient Boosting (GB)

O modelo de *gradient boosting* foi estimado considerando os mesmos dados de treinamento das modelagens anteriores. O número de atributos utilizados para a modelagem *gradient boosting* foi o mesmo que na modelagem *random forest*.

Para definição do número de árvores, algumas simulações na biblioteca "gbm" do R foram realizadas, para diversos níveis de profundidade, verificando a variação do erro para diferentes números de árvores. Pode-se observar que o erro é mínimo quando o número de árvores está entre 500 e 1000, conforme a Figura 7.

Figura 7: Diagrama de erros para modelagem *gradient boosting*.



Observou-se, então, que 800 árvores poderiam apresentar um resultado adequado para o algoritmo *gradient boosting*. Para modelagem dos dados utilizou-se, ainda, um fator de encolhimento de 0,01 e um nível de profundidade de 8,0.

4.4 Desempenho das Modelagens

O resumo com as métricas de desempenho descritas e analisadas neste estudo pode ser visualizada na Tabela 2. Por meio destes resultados, pode-se observar que o poder de explicação (R^2) do modelo *gradient boosting* (GB) se mostrou superior ao modelo clássico de regressão linear e ao modelo *random forest* (RF).

Tabela 2: Resumo com as métricas de desempenho das modelagens.

	MCR	***Random Forest***		***Gradient Boosting***	
	AMOSTRA	TREINO	TESTE	TREINO	TESTE
RMSE (R$/m²)	1294,00	946,24	1218,45	882,73	1128,20
MAPE (%)	12,41	9,56	11,64	7,84	10,14
Ratio R	1,01	0,99	1,02	1,00	1,01
COD (%)	12,26	9,84	12,02	7,82	10,12
Coeficiente R^2	0,751	0,812	0,774	0,871	0,784

Com relação ao nível de avaliação, a norma IAAO (2013) recomenda que esteja próximo da unidade. Verifica-se que o modelo de *gradient boosting*, para os dados de treinamento, se mostrou mais adequado, com nível de avaliação esperado.

Considerando o coeficiente de dispersão da mediana (COD), verifica-se que a modelagem *gradient boosting* mostrou-se superior, apresentando menor valor e indicando maior uniformidade nas avaliações. Para imóveis em regiões heterogêneas, a norma IAAO

(2013) sugere coeficientes de dispersão inferiores à 15%. Observa-se que o erro médio absoluto (MAPE) das modelagens foi inferior ao limite estabelecido pela Portaria 511/09 do Ministério das Cidades, que é de 30%. O modelo *gradient boosting* teve melhor desempenho nesta métrica.

Com relação à raiz quadrada do erro quadrático médio, observa-se que a modelagem *gradient boosting* também teve melhor desempenho, com valor aproximadamente 7% menor que a modelagem *random forest*. Verifica-se que, apesar de a modelagem *gradient boosting* apresentar métricas de treinamento que permitem indicar um sobreajustamento, isso não se confirmou nos valores preditos pela amostra de teste, mostrando, neste caso, um ajuste tão bom quanto as outras duas modelagens realizadas.

Por fim, é importante destacar que as avaliações em massa exigem uma quantidade grande de dados para uma boa projeção de valores (Oliveira *et al*., 2018). Para fins de aprendizado de máquina, a quantidade utilizada nesta pesquisa é considerada baixa, o que demonstra uma limitação deste estudo. Entretanto, novas pesquisas estão sendo realizadas com dados abertos e em grande quantidade, que serão publicadas em momento oportuno.

5 CONCLUSÕES

Procurou-se, neste estudo, investigar o desempenho dos algoritmos *random forest* e *gradient boosting* em comparação ao modelo clássico de regressão na predição do valor de mercado de apartamentos da região de Florianópolis.

Os resultados obtidos com este estudo são uma demonstração de que a precisão de avaliações em massa baseadas em aprendizado de máquina pode ser elevada. Essa constatação é importante, pois é necessário que se obtenha uma estimativa precisa do valor de mercado dos imóveis, a fim de se conduzir a um processo de avaliação em massa confiável e bem sucedido.

Verificou-se que os algoritmos de aprendizado de máquina foram capazes de explicar a variável dependente com maior precisão, produzindo resultados ligeiramente melhores do que a regressão lineares clássica, muito usada em processos de avaliação em massa.

Tradicionalmente, os modelos de regressão linear clássica enfrentam perda significativa na precisão dada a simplificação da realidade à qual o modelo está sujeito. Observa-se que parte desta precisão pode ser recuperada pelos modelos de aprendizado de máquina, sendo estes capazes de rastrear com maior sucesso a complexidade do processo de avaliação dos imóveis, resgatando informações que o modelo de regressão clássica não foi capaz de capturar.

Destaca-se, também, que outro ponto relevante nas avaliações em massa por algoritmos de aprendizado de máquina baseados em árvore é que para estes não há a necessidade de atendimento dos pressupostos básicos da inferência estatística (GROMPING, 2009).

Observou-se, ainda, que os modelos de aprendizado de máquina apresentam um *ranking* de importância das variáveis explicativas utilizadas nas modelagens. O profissional de engenharia pode fazer uso desta informação para seleção das mesmas no seu modelo.

Finalmente, este estudo demonstrou que os algoritmos de aprendizado de máquina baseados em árvore aqui abordados são capazes de fazer melhores previsões de valores para os imóveis, incorporando informações que o modelo de MCR ignora, diminuindo, assim, as distorções de valores presentes nas plantas de valores genéricos, possibilitando, desta forma, uma tributação mais justa e equânime, pilares estes, para se atingir justiça fiscal.

Agradecimentos

Este trabalho foi apoiado pelo Fundo de Apoio à Manutenção e ao Desenvolvimento da Educação Superior (FUMDES) e pelo Instituto Federal de Santa Catarina (IFSC), Brasil.

Referências

ABNT. ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-2: Avaliação de Bens.** Parte 2: Imóveis Urbanos. Rio de Janeiro, 2011. 53 p.

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.

BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil**. Brasília: DF, 1988.

BREIMAN, L. **Random forests. Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. **Bagging predictors**. Mach. Learn. 24, 1996, 123–140.

BUCKLAND, Michael K. Information as thing. **Journal of the American Society for Information Science**, v. 42, n. 5, p. 351-360, June 1991.

CEH, 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.

DUARTE, D.C.O. **Análise multicritério e geoestatística aplicadas na avaliação em massa de imóveis urbanos.** 2019. 150 f. Tese (Doutorado em Engenharia Civil) - Universidade Federal de Viçosa, Viçosa. 2019

FARIA FILHO, R.F., Gonçalves, R.M.L, Luiz, H.T.G.**Statistical models for generating the plants of generic values: an application in a small municipality.** Urbe - Revista de Gestão Urbana, 2019, V. 11.

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

GRUS, Joel. **Data Science from scratch.** O'reilly books. United States of America, 2015.

GUJARATI, D. N; PORTER, D. C. **Econometria básica**. 5. ed. Porto Alegre: AMGH Bookman, 2018.

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.

HONG, J.; CHOI, H., KIM, W.S.; **A house price valuation based on the random forest approach:** the mass appraisal of residential property in south korea. International Journal of Strategic Property Management. V. 24 Ed. 3, p. 140–152. 2020. https://doi.org/10.3846/ijspm.2020.11544

HORNBURG, R.A., HOCHHEIM, N. **Avaliação em massa de imóveis usando geoestatística e krigagem bayesiana: um estudo de em Balneário Camboriú/SC**. REEC - Revista Eletrônica De Engenharia Civil, 13(1), 2017. https://doi.org/10.5216/reec.v13i1.42347

IAAO (International Association of Assessing Officers). **Standards on ratio studies**. Mirrouri: IAAO. 2013.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning:** with applications in R. 2013

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

MARSLAND, Stephen. **Machine Learning An Algorithmic Perspective.** 2nd Ed. Taylor & Francis Group. 2015.

MARTINS FILHO, C.; BIN, O. **Estimation of hedonic price functions via additive nonparametric regression.** Empirical Economics, 30, 2005, 93–114

MAYRINK, Victor Teixeira de Melo. **Avaliação do algoritmo Gradient Boosting em aplicações de previsão de carga elétrica a curto prazo.** 91 f. Dissertação Universidade Federal de Juiz de Fora. Programa de Pós-Graduação em Modelagem Computacional, 2016.

MCCLUSKEY, W. J., e ANAND, S. **The application of intelligent hybrid techniques for the mass appraisal of residential properties**. Journal of Property Investment and Finance, 17(3), 1999, 218–238

MITCHELL, Tom M. Machine Learning, por Tom M. Mitchell (Autor), Thomas Mitchell (Autor), Mitchell Thomas (Autor). 1997

MURPHY, Kevin P. **Machine learning: a probabilistic perspective**. Massachusetts Institute of Technology. 2012.

OLIVEIRA, Antônio Augusto Ferreira. **Avaliação em massa com modelos de aprendizado de máquina aplicados aos terrenos urbanos do município de Fortaleza**. 80 f. Dissertação

(mestrado) – Universidade Federal do Ceará, Mestrado Profissional em Economia do Setor Público, Fortaleza, 2020.

PELLI NETO, A. **Redes neurais artificiais aplicadas às avaliações em massa: estudo de caso para a cidade de Belo Horizonte / MG.** 2006. 111 f. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal de Minas Gerais, Belo Horizonte. 2006. http://hdl.handle.net/1843/AVFV-6W7R62pp. 2825-2830, 2011.

SAMUEL, Arthur L. **Some Studies in Machine Learning Using the Game of Checkers**. IBM Journal of Research and Development. 1959

SELIM, H., **Determinants of house prices in Turkey: Hedonic regression versus artificial neural network**. Expert Systems with Applications, 36, 2009**,** 2843–2852.

THEODORO, L.T.C.; UBERTI , M.S.; ANTUNES , M.A.H.; DEBIASI, P., 2019. **Avaliação em massa de imóveis rurais através da regressão clássica e da geoestatística.** Revista Brasileira de Cartografia, v. 71, n. 2, p. 459-485, 24 jun. 2019. https://doi.org/10.14393/rbcv71n2-47458

UBERTI, M.S., ANTUNES, M.A.H., DEBIASI, P., TASSINARI, W. **Mass appraisal of farmland using classical econometrics and spatial modeling**. Land Use Policy 72, 2018, 161-170. https://doi.org/10.1016/j.landusepol.2017.12.044

VERIKAS, A., LIPNICKAS, A., & MALMQVIST, K., **Selecting neural networks for a committee decision.** International Journal of Neural Systems, 12(5), 2002, 351–362

YILMAZER, S., KOCAMAN, S., **A mass appraisal assessment study using machine learning based on multiple regression and random forest.** Land Use Policy 99, 2020**.** https://doi.org/10.1016/j.landusepol.2020.104889

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.

ZILLI, C. A. **Regressão geograficamente ponderada aplicada na avaliação em massa de imóveis urbanos.** (Dissertação de Mestrado). Programa de Pós Graduação em Engenharia de Transportes e Gestão Territorial, Universidade Federal de Santa Catarina, Florianópolis (2020).