

AVALIAÇÃO DE IMÓVEIS COM TÉCNICAS DE *MACHINE LEARNING*

Trabalho de Avaliação

ROGÉRIO TORESANI

João Pessoa - PB
2025

REALIZAÇÃO:



PATROCÍNIO:



AVALIAÇÃO DE IMÓVEIS COM TÉCNICAS DE *MACHINE LEARNING*

RESUMO

A avaliação de imóveis é atividade de grande relevância, visto que subsidia financiamentos imobiliários sendo que atualmente a carteira de crédito no Brasil supera um trilhão de reais. A presente pesquisa se propõe a elaborar e analisar modelos de predição para avaliação de imóveis utilizando técnicas de *machine learning*. Foram desenvolvidos, utilizando aplicativo Rstudio, modelos através de regressão linear múltipla, árvore de regressão, *random forest*, redes neurais artificiais e *deep learning*. Foi desenvolvida regressão linear também no aplicativo SisDEA. Os modelos foram analisados a luz do coeficiente de determinação, RMSE e MAE, ficando demonstrando que os modelos de *machine learning* tiveram melhor capacidade de predição para o caso estudado.

Palavras-chave: avaliação de imóveis; machine learning, random forest, regressão linear, redes neurais artificiais.

REALIZAÇÃO:



PATROCÍNIO:



SUMÁRIO

1. EXPOSIÇÃO	4
2. CONCLUSÃO	24
REFERÊNCIAS	24

REALIZAÇÃO:



PATROCÍNIO:



1. EXPOSIÇÃO

Introdução

A atividade de avaliação de imóveis é bastante relevante para a sociedade pois permite determinar o valor de bens imóveis a partir de métodos científicos. É de se esperar que em transações de compra e venda, aluguel ou dação em pagamento existam expectativas de valores diferentes e as vezes conflitantes, frutos das diferentes perspectivas de cada parte. O vendedor/locador tende a superestimar sua propriedade enquanto o comprador/locatário por sua vez pode ter dificuldades em reconhecer todas as vantagens de determinado imóvel. Quando a transação requer a participação de um Banco, por exemplo para financiamento ou *Home Equity*, é uma questão muito sensível para a instituição financeira saber o real valor do bem, visto que se adotar valor acima do real coloca em risco suas garantias e se adotar valor abaixo do real pode perder um negócio promissor.

Segundo Fiker (2017) existem diversos métodos para avaliação de imóveis, sendo que o mais recomendado é o Método Comparativo de Dados de Mercado onde, conforme ABNT (2019) se identifica o valor de mercado do bem por meio de tratamento dos atributos dos elementos comparáveis, constituintes da amostra. A utilização desse método a partir da regressão linear múltipla se apresenta como a solução mais comum adotada por avaliadores de imóvel atualmente (Duarte, Hochheim e Reginato, 2022). Nesse sentido, existem diversos aplicativos no mercado para avaliação de imóveis como por exemplo o SisDEA (Pelli Sistemas, 2024), o TS SISREG (Tecsyst Engenharia, 2024) e o CASTLER (Regression Engenharia, 2024) e todos eles utilizam a Regressão Linear ou não Linear como principal ferramenta.

Durante muitos anos as avaliações foram elaboradas preferencialmente com uso de fatores de homogeneização conforme pode ser verificado, por exemplo na Norma Para Avaliação de Imóveis Urbanos do IBAPE/SP (IBAPE/SP, 2005) até que por volta de 2010 começou-se a usar preferencialmente o método comparativo de dados de mercado com uso de regressão linear. Nessa época houve a revisão da Norma ABNT NBR 14653:2 e foi dada ênfase a esse último método de avaliação. Nesse contexto, ocorreu o advento de aplicativos dedicados a regressão linear a partir de dados de mercado coletados. Como as instituições financeiras, grandes contratantes do serviço de avaliação de imóveis, começaram a priorizar o uso de regressão linear. A Caixa Econômica Federal, por exemplo, determinou em seu último edital de credenciamento de avaliadores de imóveis que a regra é usar o Método Comparativo de Dados de Mercado com uso de regressão linear múltipla, sendo permitidas algumas exceções (Caixa Econômica Federal, 2019). Assim sendo, o perfil do profissional avaliador passou a se moldar no entorno do método comparativo de dados de mercado e uso de regressão linear com apoio dos aplicativos existentes.

Essa situação tem se mantido relativamente inalterada nos últimos anos, porém nesse período o campo da ciência de dados evoluiu rapidamente. O cenário tem mudado com o surgimento de novas técnicas, melhores equipamentos, computação em nuvem, entre outros fatores. O uso da ciência de dados tem se intensificado nos diversos setores econômicos e atividades humanas. Dessa forma, é verossímil que as novas técnicas tem potencial para contribuir com a atividade de avaliação de imóveis.

Atualmente a carteira de crédito imobiliário de pessoas físicas das instituições financeiras no Brasil supera 1 trilhão de reais, sendo que a maioria das garantias é tipo Alienação Fiduciária (BCB, 2024). Nesse tipo de financiamento, o imóvel é a garantia da operação, ou seja, caso o devedor não honre com os pagamentos o mesmo poderá ir a leilão

sendo que em geral as instituições financeiras não podem executar a dívida para além da garantia, que é o imóvel (Lei 9514/1997). O Banco Central do Brasil [BCB] (2012, 2022) estabeleceu que toda operação de financiamento imobiliário prescinde da respectiva avaliação do imóvel a valor de mercado, regra essa também recomendada pelo Acordo de Basiléia III (BIS, 2017). Diante da importância e relevância da atividade de avaliação de imóveis, a presente pesquisa se justifica enquanto busca por alternativas para realização da atividade utilizando métodos mais atuais que poderão se mostrar mais rápidos e seguros que os métodos atualmente utilizados.

Assim sendo, esse trabalho de conclusão tem como objetivo elaborar e analisar modelos preditivos para avaliação de imóveis a partir de técnicas de *machine learning*.

Material e Métodos

Instrumento de Coleta de Dados

A presente pesquisa é um Estudo de Caso. Por meio de um Estudo de Caso se investiga um fenômeno contemporâneo dentro de seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos (Yin, 2014). Além disso, o Estudo de Caso pressupõe profundidade e detalhamento de uma ou poucas unidades (Vergara, 2011), assim como intencionado nesse estudo. Os principais passos para desenvolvimento da pesquisa estão condensados na Figura 1.

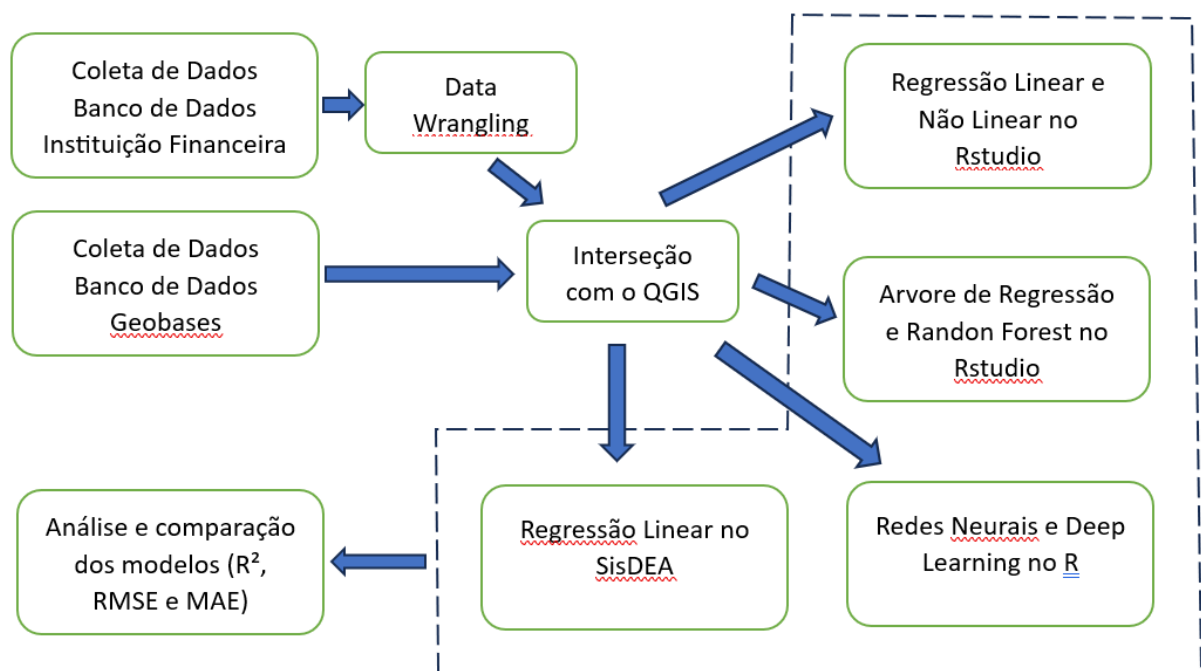


Figura 1. Principais passos para desenvolvimento da pesquisa
 Fonte: Dados originais da pesquisa

Para elaboração de modelos preditivos para avaliação de imóveis a partir de técnicas de *machine learning* foram utilizados dados do mercado imobiliário das cidades de Serra, Vitória e Villa Velha, todas localizadas do Estado do Espírito Santo. As observações utilizadas se referem a apartamentos que foram avaliados quanto ao valor de mercado para obtenção de financiamento em uma Instituição Financeira (IF). As observações se referem a avaliações ocorridas entre 2015 e 2024.

Os dados foram obtidos (i) no banco de dados do Sistema de Informações do Mercado Imobiliário da IF e (ii) junto a Plataforma Geobases (Geobases, 2018). Ambas as fontes forneceram dados estruturados em forma de tabelas sendo que no sistema da IF foi gerado um arquivo de colunas separadas por vírgulas ou .csv e a Plataforma Geobases forneceu um arquivo do tipo *shapefile*.

O banco de dados fornecidos pela IF é composto por 57.639 observações e 337 variáveis, abrangendo imóveis de todas as tipologias (apartamentos, casas, lojas, terrenos, salas comerciais entre outros) localizados no Estado do Espírito Santo. Cada observação equivale a uma avaliação feita para o imóvel com o objetivo de subsidiar algum tipo de financiamento imobiliário. As variáveis de cada observação se referem a características dos imóveis (por exemplo: quantidade de quartos, área e estado de conservação) e características da região do imóvel (por exemplo: coordenadas geográficas do imóvel, vocação da região e padrão de acabamento do edifício).

O banco de dados obtido junto a Plataforma Geobases é composto por 3598 observações e 25 variáveis. Cada observação se refere a um setor censitário dos municípios da Grande Vitória (Cariacica, Fundão, Guarapari, Serra, Viana, Vila Velha e Vitória). As variáveis do banco de dados se referem a características do setor censitário tais como Bairro, Cidade, Código do Setor Censitário, Renda, entre outros. Como se trata de arquivo do tipo *shapefile*, o banco de dados permite a visualização dos dados em mapa.

Forma de Abordagem

Para análise dos dados foram utilizados os aplicativos RStudio v. 2022.11.0 Build 148 (Posit PBC, Boston, MA, USA), SisDEA v. Home Version 1.25 (Envolvimento de Dados, Redes Neurais Artificiais, Regressão Linear e Não Linear, Belo Horizonte, BH, Brasil) e QGIS v. 3.22.14 (Quantum Geographic Information System, Boston, MA, USA). O principal aplicativo utilizado foi o RStudio, que concentrou as atividades de *Data Wrangling* e construção de modelos. O aplicativo SisDEA foi utilizado para geração de modelos de regressão linear e não linear enquanto o aplicativo QGIS foi utilizado para realizar a interseção entre o banco de dados fornecido pela IF e o banco de dados obtido no Portal Geobases.

Análise de Dados

Inicialmente realizou-se o processo de *Data Wrangling* do banco de dados fornecido pela IF. O *Data Wrangling* pode ser compreendido como um processo iterativo de processamento e enriquecimento dos dados para posterior análise e visualização (Vos et al., 2022). Nesse processo se busca dados duplicados, erros de Tabela, incorreta imputação de valores ou ausência de imputação de valores (Azeroual, 2020). O *Data Wrangling* foi realizado no aplicativo Rstudio utilizando os pacotes dplyr e Tidyverse. Para retirada de observações com dados inconsistentes foi realizada a seguinte filtragem descrita na Tabela 1.

Tabela 1. Critérios de filtragem de observações do banco de dados fornecido pela Instituição Financeira [IF]

Variável	Critérios de Restrição
Total de Vagas	Menor ou igual a 5 vagas
Número de Quartos	Entre 1 e 5 quartos
Área do imóvel	Entre 30 e 400 m ²
Valor Unitário (R\$/m ²)	Maior que R\$2000/m ²
Padrão das Edificações na Região	Normal (forte predominância) ou Normal (c/ aspectos de alto) ou Normal (c/ aspectos de baixo) ou Alto (por predominância) ou Baixo

Acabamento Fachada Principal da Edificação	Pintura/Textura ou Pastilha ou Cerâmica de Boa Qualidade ou Granito ou Compatível ou Porcelanato
Esquadrias Fachada Principal da Edificação	Alumínio ou Ferro ou Madeira ou Vidro Temperado
Padrão de acabamento da Unidade	Normal (forte predominância) ou Normal (c/ aspectos de baixo) ou Normal (c/ aspectos de baixo) ou "Normal (c/ aspectos de alto) ou Regular (reparos simples) ou Alto (superior, luxo)
Estado de Conservação da Unidade	Bom (aparência de novo) ou Bom (aparência de usado) ou Regular (reparos importantes) ou Regular (reparos simples)
Idade Estimada da edificação	<= 5 ou > 5 <= 10 ou > 10 <= 20 ou > 20 <= 50 ou > 50 <= 100
Nº de Unidades no Andar	Entre 1 e 12 unidades
Nº de pavimentos do Prédio	entre 2 e 30
Estado de Conservação do prédio	Bom (aparência de novo) ou Bom (aparência de usado) ou Regular (reparos importantes) ou Regular (reparos simples)

Fonte: Dados originais da pesquisa

Foi necessário realizar filtragem dos dados fornecidos pela IF para adequação ao caso a ser estudado. Nesse sentido, a Tabela 2 contém as restrições estabelecidas.

Tabela 2. Filtragem para adequação ao estudo de caso pesquisado

Variável	Atributo
Município	Vitoria, Serra, Vila Velha
Tipo de Peça Técnica	Laudo de Avaliação
Categoria do Imóvel	Apartamento

Fonte: Dados originais da pesquisa

O banco de dados da IF conta com 337 variáveis, porém a grande maioria não se mostrou relevante para a presente pesquisa pois se referem, por exemplo, ao endereço dos imóveis, ao nome do cliente, à agência concessora do crédito, ao tipo de produto ofertado ao cliente e à origem de recursos, entre outros. Dessa forma, foram mantidas 16482 observações e 26 variáveis. A Tabela 3 contém a caracterização do banco de dados fornecido pela IF ao final do processo de *Data Wrangling* e demais filtragens necessárias.

Tabela 3. Características do banco de dados fornecido pela Instituição Financeira após processo de *Data Wrangling* e demais filtragens

Nº	Variável	Descrição	Descrição
1	ID	Length:16482 Class :character	Codigo identificador de cada observação
2	Latitude	Length:16482 Class :character	Latitude geográfica do imóvel avaliado no Datum WGS84
3	Longitude	Length:16482 Class :character	Longitude geográfica do imóvel avaliado no Datum WGS84

4	Quantidade de Vagas	Geral	Min. :0.000 1st Qu.:1.000 Median :1.000 Mean :1.096 3rd Qu.:1.000 Max. :4.000	Quantidade de vagas de garagem pertencentes ao imóvel avaliando
5	Quantidade Quartos	de	Min. :1.0 1st Qu.:2.0 Median :2.0 Mean :2.3 3rd Qu.:3.0 Max. :5.0	Quantidade de quartos do imóvel avaliando
6	Área privativa		Min. :30.00 1st Qu.:47.78 Median :59.20 Mean :68.84 3rd Qu.:75.96 Max. :400.00	Área privativa do imóvel avaliando em metros quadrados
7	Valor Unitário		Min. :2000 1st Qu.:2799 Median :3490 Mean :3965 3rd Qu.:4676 Max. :15909	Valor Unitário do imóvel avaliando em R\$/m²
8	Data de Referência da Avaliação		Min. :16881 1st Qu.:17809 Median :18663 Mean :18514 3rd Qu.:19244 Max. :19746	Data da avaliação do Imóvel avaliando
9	Equipamentos de Segurança	de	NÃO:8243 SIM:8239	Existência (Sim ou Não) de equipamentos de segurança no condomínio do imóvel avaliando
10	Espaço Churrasqueira	com	NÃO:4418 SIM:12064	Existência (Sim ou Não) de espaço para churrasqueira no condomínio do imóvel avaliando
11	Mini Esportiva	Quadra	NÃO:14348 SIM:2134	Existência (Sim ou Não) de Mini Quadra Esportiva no condomínio do imóvel avaliando
12	Piscina		NÃO:7035 SIM:9447	Existência (Sim ou Não) de Piscina no condomínio do imóvel avaliando
13	Portaria e ou Guarita		NÃO:3293 SIM:13189	Existência (Sim ou Não) de Portaria e ou Guarita do imóvel avaliando
14	Salão de Festas		NÃO:3101 SIM:13381	Existência (Sim ou Não) de Salão de Festas no condomínio do imóvel avaliando
15	Outros Salões de Lazer	de	NÃO:14281 SIM:2201	Existência (Sim ou Não) de Outros Salões de Lazer do imóvel avaliando
16	Sauna, Ofurô ou Hidromassagem		NÃO:12265 SIM:4217	Existência (Sim ou Não) de Sauna, Ofurô ou Hidromassagem no condomínio do imóvel avaliando

17	Quadra Poliesportiva	NÃO:11179 SIM:5303	Existência (Sim ou Não) de Quadra Poliesportiva no condomínio do imóvel avaliando
18	Pilotis	NÃO:14921 SIM:1561	Existência (Sim ou Não) de Pilotis no condomínio do imóvel avaliando
19	Acabamento Fachada Principal da Edificação	Cerâmica de Boa Qualidade:647 Granito ou Compatível:342 Pastilha: 2314 Pintura/Textura: 13008 Porcelanato:171	Tipo de material empregado nas paredes da fachada principal do Edifício do imóvel avaliando
20	Esquadrias Fachada Principal da Edificação	Alumínio: 15516 Ferro: 69 Madeira: 116 Vidro Temperado: 781	Tipo de material empregado nas esquadrias da fachada principal do Edifício do imóvel avaliando
21	Padrão de acabamento da Unidade	Alto (superior, luxo): 23 Normal (c/ aspectos de alto): 2294 Normal (c/ aspectos de baixo): 2597 Normal (forte predominância): 11568	Padrão de acabamento constatado no imóvel avaliando
22	Estado de Conservação da Unidade	Bom (aparência de novo): 2520 Bom (aparência de usado): 11174 Regular (reparos importantes): 93 Regular (reparos simples): 2695	Estado de conservação do imóvel avaliando
23	Idade Estimada da edificação	<= 5: 3651 > 10 <= 20: 4578 > 20 <= 50: 3269 > 5 <= 10: 4961 > 50 <= 100: 23	Idade estimada em anos da edificação onde se localiza o imóvel avaliando
24	Nº de Unidades no Andar	Min. :1.00 1st Qu.:4.00 Median :6.00 Mean :5.87 3rd Qu.:8.00 Max. :12.00	Quantidade de unidades habitacionais no mesmo pavimento que o imóvel avaliando
25	Nº de pavimentos do Prédio	Min. :2.000 1st Qu.:4.000 Median :8.000 Mean :9.029 3rd Qu.:12.000 Max. :30.000	Quantidade de pavimentos existentes no edifício onde se localiza o imóvel avaliando.
26	Estado de Conservação do prédio	Bom (aparência de novo): 2152 Bom (aparência de usado): 10900 Regular (reparos importantes): 82 Regular (reparos simples): 3348	Estado de conservação do edifício onde se localiza o imóvel avaliando

Fonte: Dados originais da pesquisa

Após a realização do *Data Wrangling* na base obtida no Portal Geobases constatou-se que todas as observações e variáveis poderiam ser mantidas devido (i) à inexistência de inconsistências e (ii) a total adequação ao objetivo da pesquisa.

A próxima etapa foi realizada com auxílio do aplicativo QGIS. Ambas as bases foram carregadas no aplicativo com objetivo de realizar a interseção dos dados tendo como

referência as coordenadas geográficas das observações do banco de dados da IF e as informações de geolocalização do polígono referente a cada setor censitário da base obtida no Portal Geobases. A Figura 2 contém um fragmento do mapa contendo os polígonos dos setores censitários e os pontos referentes as observações e foi obtida no aplicativo QGIS. Essa operação também auxiliou na retirada de observações com informações incorretas de latitude e longitude.



Figura 2. Fragmento do mapa gerado no QGIS setores censitários e observações do banco de dados fornecido pela Instituição Financeira
Fonte: Dados originais da pesquisa

A interseção resultou em um novo banco de dados com todas as variáveis dos dois bancos de dados. Esse novo banco de dados foi exportado no formato .csv e importado no aplicativo RStudio para continuidade da análise. Em seguida foi mantida apenas a variável Renda entre as que vieram do banco de dados obtido no Portal Geobases bem como foram retiradas as variáveis Latitude e Longitude vindas do banco de dados da IF.

Como pode ser visto, a operação de interseção dos bancos de dados ocorreu para obtenção da variável Renda. O valor dessa variável em cada observação significa a renda do setor censitário ao qual o imóvel da observação está contido. Segundo Dantas, Magalhães e Vergolino (2007), a localização de um imóvel é informação primordial em uma avaliação de imóveis e conforme Doszyń (2022) são utilizadas variáveis do tipo proxy para considerar os efeitos de localização do imóvel em modelo. Nesse sentido, a variável renda foi utilizada como proxy para retratar o efeito da localização de cada imóvel nos modelos desenvolvidos. Ao final dessa etapa foi realizado novo saneamento com retirada de observações cuja renda era igual a zero.

O banco de dados resultante possui algumas variáveis categóricas e para utilizá-las nos modelos sem risco de cometer ponderação arbitrária foi realizado o processo de conversão de cada variável categórica em n-1 variáveis *dummy*, onde n é a quantidade de categorias da variável. A conversão foi realizada no Rstudio com auxílio da ferramenta `dummy_columns` do pacote `fastDummies`. Dessa forma, obteve-se um bando de dados resultante com 13556 observações e 43 variáveis.

Regressões lineares e não lineares

Foi realizada regressão linear e não linear dos dados utilizando os aplicativos Rstudio e SisDEA. Inicialmente serão definidos os procedimentos realizados no Rstudio para na sequência descrever os procedimentos realizados no SisDEA.

Iniciou-se no Rstudio com a regressão linear utilizando a base de dados resultante com auxílio da ferramenta lm do pacote stats. Com a aplicação do procedimento *Stepwise* foi possível eliminar as variáveis explicativas cujos parâmetros não se mostraram estatisticamente diferentes de zero (Favero e Belfiore, 2022). Para isso foi utilizada a ferramenta step do pacote stats. As principais características do modelo resultante foram condensadas na tabela 4 com auxílio da ferramenta summary e a eq. (1) representa o modelo resultante.

Tabela 4. Principais características do modelo elaborado com regressão linear

Coefficients:	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1,07E+07	2,05E+05	-51.985	<2e-16***
Renda	2,13E+02	4,89E+00	43.522	<2e-16***
Quantidade Geral de Vagas	4,93E+05	2,03E+04	24.285	<2e-16***
Quantidade de Quartos	1,94E+05	2,13E+04	9.102	<2e-16***
Área privativa	-1,28E+04	4,53E+02	-28.234	<2e-16***
Data de Referência da Avaliação	7,17E+02	1,06E+01	67.976	<2e-16***
Nº de Unidades no Andar	-2,18E+04	3,99E+03	-5.475	4.46e-08***
Nº de pavimentos do Prédio	4,56E+04	2,33E+03	19.577	<2e-16***
Padrão das Edificações na Região_Alto(por predominância)	1,58E+05	5,95E+04	2.650	0.00807**
Padrão das Edificações na Região_Baixo(por predominância)	-2,68E+05	1,31E+05	-2.056	0.03983*
Padrão das Edificações na Região_Alto(por predominância)_Normal(c/aspectosdealto)	1,70E+05	3,09E+04	5.487	4.17e-08***
Padrão das Edificações na Região_Alto(por predominância)_Normal(c/aspectosdebaixo)	-1,89E+05	2,90E+04	-6.508	7.90e-11***
Estado de Conservação do prédio_Bom(aparênciadenovo)	2,10E+05	3,18E+04	6.601	4.23e-11***
Estado de Conservação do prédio_Regular(reparossimples)	5,90E+04	2,74E+04	2.158	0.03095*
Idade Estimada da edificação_<=5	3,54E+05	2,76E+04	12.799	<2e-16***
Idade Estimada da edificação_>20<=50`	-6,58E+04	2,63E+04	-2.501	0.01240*
Idade Estimada da edificação_>50<=100	-6,26E+05	2,25E+05	-2.781	0.00542**
Equipamentos de Segurança_SIM	-7,64E+04	1,78E+04	-4.299	1.73e-05***
Espaço com Churrasqueira_NÃO	4,88E+04	2,32E+04	2.101	0.03568*
Mini Quadra Esportiva_SIM	-1,28E+05	2,75E+04	-4.656	3.25e-06***
Piscina_NÃO	-3,23E+05	2,59E+04	-12.484	<2e-16***
Portaria e ou Guarita_NÃO	6,34E+04	2,29E+04	2.774	0.00554**
Outros Salões de Lazer_SIM	6,95E+04	2,65E+04	2.629	0.00858**
Sauna, Ofurô ou Hidromassagem_SIM	1,79E+05	2,39E+04	7.496	6.99e-14***
Quadra Poliesportiva_SIM	-1,31E+05	2,07E+04	-6.313	2.82e-10***
Pilotis_SIM	1,26E+05	2,91E+04	4.320	1.57e-05***

Acabamento Fachada Principal da Edificação_Cerâmica de Boa Qualidade	4,76E+05	4,55E+04	10.460	<2e-16***
Acabamento Fachada Principal da Edificação_Granito ou Compatível	6,91E+05	6,20E+04	11.142	<2e-16***
Acabamento Fachada Principal da Edificação_Pastilha	4,47E+05	2,89E+04	15.496	<2e-16***
Acabamento Fachada Principal da Edificação_Porcelanato	3,57E+05	8,37E+04	4.267	1.99e-05***
Padrão de acabamento da Unidade_Alto (superior, luxo)	2,63E+06	2,24E+05	11.726	<2e-16***
Padrão de acabamento da Unidade_Normal (c/ aspectos de alto)	5,81E+05	2,91E+04	19.943	<2e-16***
Padrão de acabamento da Unidade_Normal (c/ aspectos de baixo)	-2,41E+05	2,81E+04	-8.545	<2e-16***
Estado de Conservação da Unidade_Regular (reparos importantes)	-2,77E+05	1,08E+05	-2.577	0.00999**
Estado de Conservação da Unidade_Regular (reparos simples)	-9,09E+04	2,87E+04	-3.170	0.00153**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 972.1 on 13521 degrees of freedom

Multiple R-squared: 0.6445, Adjusted R-squared: 0.6436

F-statistic: 721 on 34 and 13521 DF, p-value: < 2.2e-16

Fonte: Dados originais da pesquisa

(1)

$$\text{Valor unitário} = -1,070185e04 + 2,133041e(-1) * \text{Renda} + 4,900252e02 * \text{Vagas} +$$

$$1,927157e02 * \text{Quartos} - 1,274853e01 * \text{Área} + 7,204777e(-1) * \text{Data} - 2,298303e01 * \text{Unidades. no. Andar} +$$

$$4,514653e01 * \text{Nº. de pavimentos} + 1,626565e02 * \text{Padrão. das. Edificações. na. Região_Alto} - 2,719046e02 * \text{Padrão. das. Edificações. na. Região_baixo} +$$

$$1,713092e02 * \text{Padrão Edificações Região_Normal (aspectos de alto)} - 1,890581e02 * \text{Padrão Edificações Região_Normal (aspectos de baixo)} +$$

$$1,750781e02 * \text{Estado Conservação prédio_Bom (aparência de novo)} - 1,615463e02 * \text{Estado Conservação prédio_Regular (reparos importantes)} +$$

$$5,692096e01 * \text{Estado Conservação prédio_Regular (reparos simples)} + 3,332346e02 * \text{Idade. Estimada. edificação} < 5 -$$

$$4,069044e01 * \text{Idade. Estimada. edificação} > 10 < 20 - 9,315002e01 * \text{Idade. Estimada. da. edificação. como. um. todo} > 20 <= 50 -$$

$$6,473217e02 * \text{Idade. Estimada. da. edificação} > 50 <= 100 - 7,781677e01 * \text{Equipamentos. de. Segurança_SIM} +$$

$$5,823306e01 * \text{Espaço. com. Churrasqueira_NÃO} - 1,295609e02 * \text{Mini. Quadra. Esportiva_SIM} -$$

$$3,135604e02 * \text{Piscina_NÃO} + 7,071176e01 * \text{Portaria. e. ou. Guarita_NÃO} - 3,248473e01 * \text{Salão. de. Festas_NÃO} +$$

$$6,921836e01 * \text{Outros. Salões. de. Lazer_SIM} + 1,781118e02 * \text{Sauna... Ofurô... Hidromassagem_SIM} -$$

$$1,308829e02 * \text{Quadra. Poliesportiva_SIM} + 1,261900e02 * \text{Pilotis_SIM} + 4,766777e02 * \text{Acabamento. Fachada. Principal_Cerâmica de Boa Qualidade} +$$

$$6,938193e02 * \text{Acabamento. Fachada. Principal. _Granito ou Compatível} + 4,490277e02 * \text{Acabamento. Fachada. Principal_Pastilha} +$$

$$3,567225e02 * \text{Acabamento. Fachada. Principal_Porcelanato} - 1,174471e01 *$$

Esquadrias. Fachada. Principal_Ferro — 2.604852e01 *
 Esquadrias. Fachada. Principal. da. Edificação_Madeira — 1.919532e01 *
 Esquadrias. Fachada. Principal_Vidro Temperado` + 2.607904e03 *
 Padrão. de. acabamento. da. Unidade_Alto (superior, luxo)` + 5.778520e02 *
 Padrão. de. acabamento. da. Unidade_Normal (c/ aspectos de alto) — 2.394693e02 *
 Padrão. de. acabamento. da. Unidade_Normal (c/ aspectos de baixo) + 4.669540e01 *
 Estado. de. Conservação. da. Unidade_Bom (aparência de novo)` — 2.477866e02 *
 Estado. de. Conservação. da. Unidade_Regular (reparos importantes) — 8.348691e01 *
 Estado. de. Conservação. da. Unidade_Regular (reparos simples)`

Para verificação da aderência dos resíduos à normalidade não foi possível utilizar o teste de Shapiro-Francia pois a ferramenta sf.test do pacote nortest só consegue processar até 5000 observações, quantidade inferior ao do banco de dados em estudo. Dessa forma, procedeu-se a plotagem do histograma de distribuição dos resíduos para verificação visual de sua aderência. Como pode ser verificado na Figura 3, existe aderência da distribuição de resíduos à curva normal pois o histograma construído possui forma que se aproxima da curva normal.

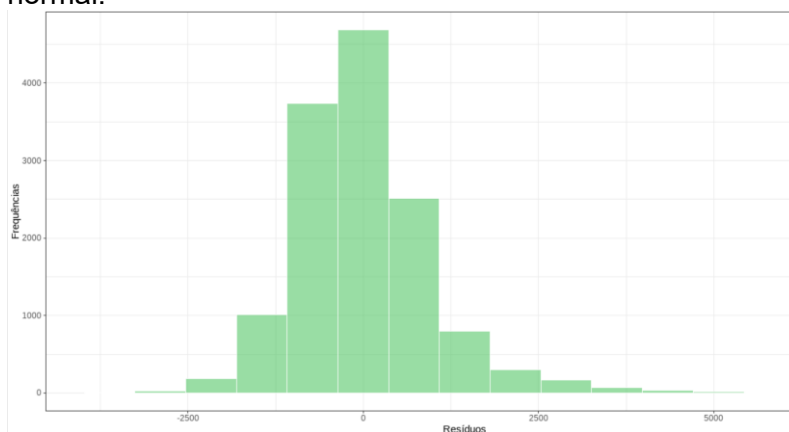


Figura 3. Histograma de distribuição dos resíduos da regressão linear
 Fonte: Dados originais da pesquisa

Para verificação da heterocedasticidade no modelo foi utilizado o teste de Breusch-Pagan através da ferramenta ols_test_breusch_pagan do pacote olsrr, que indicou ausência de heterodasticidade conforme Figura 4.

```

Breusch Pagan Test for Heteroskedasticity
-----
Ho: the variance is constant
Ha: the variance is not constant

Data
-----
Response : R..m2
Variables: fitted values of R..m2

Test Summary
-----
DF          =    1
Chi2        = 3926.2625
Prob > Chi2 = 0.0000
  
```

Figura 4. Resultado do teste de Breusch-Pagan para a regressão linear
 Fonte: Dados originais da pesquisa

Buscando um modelo mais ajustado aos dados e ao mercado imobiliário que se tenta retratar, foi realizada regressão não linear a partir da transformação de Box-Cox. Conforme Favero e Belfiore (2022), a transformação de Box-Cox pode auxiliar na definição de outras formas funcionais, diferentes da linear. Assim sendo, foi calculado o parâmetro lambda ($\lambda = -0.7199462$) através da ferramenta powerTransform do pacote car. Em seguida foi feita a transformação da variável dependente a partir do lambda calculado e foi realizada nova regressão, dessa vez utilizando a variável dependente transformada. O modelo obtido da variável transformada passou também pelo procedimento *Stepwise* e o resultado pode ser visualizado na Tabela 5 e na eq. (2).

Tabela 5. Principais características do modelo elaborado com regressão com uso de transformação de Box-Cox

Coefficients:	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1,38E+03	1,13E-01	12.177.343	< 2e-16 ***
Renda	1,18E-04	2,65E-06	44.350	< 2e-16 ***
Quantidade Geral de Vagas	2,59E-01	1,10E-02	23.560	< 2e-16 ***
Quantidade de Quartos	1,21E-01	1,17E-02	10.257	< 2e-16 ***
Área privativa	-6,52E-03	2,48E-04	-26.320	< 2e-16 ***
Data de Referência da Avaliação	3,97E-04	5,83E-06	68.154	< 2e-16 ***
Nº de pavimentos do Prédio	2,58E-02	1,28E-03	20.106	< 2e-16 ***
				0.001368
Padrão das Edificações na Região_Baixo	-2,32E-01	7,25E-02	-3.202	**
Padrão das Edificações na Região_Normal(c/aspectosdealto)	1,11E-01	1,65E-02	6.744	1.60e-11 ***
Padrão das Edificações na Região_Normal(c/aspectosdebaixo)	-1,71E-01	1,60E-02	-10.649	< 2e-16 ***
Estado de Conservação do prédio_Bom(aparênciadenovo)	1,17E-01	1,76E-02	6.665	2.75e-11 ***
Estado de Conservação do prédio_Regular(reparosimportantes)	-1,48E-01	6,93E-02	-2.134	0.032833 *
Idade Estimada da edificação_<=5	1,90E-01	1,53E-02	12.405	< 2e-16 ***
				0.000130
Idade Estimada da edificação_>50<=100	-4,79E-01	1,25E-01	-3.828	***
				3.44e-05
Espaço com Churrasqueira_NÃO	5,51E-02	1,33E-02	4.144	***
				8.25e-14
Mini Quadra Esportiva_SIM	-1,14E-01	1,53E-02	-7.474	***
Piscina_NÃO	-3,08E-01	1,38E-02	-22.339	< 2e-16 ***
				4.15e-07
Portaria e ou Guarita_NÃO	6,67E-02	1,32E-02	5.065	***
				8.65e-05
Salão de Festas_NÃO	-5,71E-02	1,45E-02	-3.927	***
				0.000869
Outros Salões de Lazer_SIM	4,89E-02	1,47E-02	3.331	***

				3.98e-11	
Sauna, Ofurô ou Hidromassagem_SIM	8,80E-02	1,33E-02	6.610	***	
Quadra Poliesportiva_SIM	-1,30E-01	1,15E-02	-11.336	< 2e-16	***
				1.23e-08	
Pilotis_SIM	9,20E-02	1,61E-02	5.699	***	
Acabamento Fachada Principal da Edificação_CerâmicodeBoaQualidade	2,25E-01	2,52E-02	8.902	< 2e-16	***
Acabamento Fachada Principal da Edificação_GranitoouCompatível	3,08E-01	3,43E-02	8.986	< 2e-16	***
Acabamento Fachada Principal da Edificação_Pastilha	2,63E-01	1,60E-02	16.483	< 2e-16	***
Acabamento Fachada Principal da Edificação_Porcelanato	1,54E-01	4,64E-02	3.312		0.000930 ***
Esquadrias Fachada Principal da Edificação_Ferro	-1,57E-01	7,26E-02	-2.161	0.030675	*
Esquadrias Fachada Principal da Edificação_Madeira	-1,08E-01	5,24E-02	-2.060	0.039388	*
Padrão de acabamento da Unidade_Alto(superior,luxo)	6,31E-01	1,23E-01	5.128		2.96e-07 ***
Padrão de acabamento da Unidade_Normal(c/aspectosdealto)	2,06E-01	1,59E-02	12.955	< 2e-16	***
Padrão de acabamento da Unidade_Normal(c/aspectosdebaixo)	-2,20E-01	1,56E-02	-14.071	< 2e-16	***
Estado de Conservação da Unidade_Regular(reparosimportantes)	-1,55E-01	6,02E-02	-2.578		0.009947 **
Estado de Conservação da Unidade_Regular(reparossimples)	-7,81E-02	1,33E-02	-5.868		4.51e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0005406 on 13522 degrees of freedom

Multiple R-squared: 0.6702, Adjusted R-squared: 0.6694

F-statistic: 832.7 on 33 and 13522 DF, p-value: < 2.2e-16

Fonte: Dados originais da pesquisa

(2)

$$\text{Valor unitário} = \{[1.376994 + 1.171051e(-7) * \text{Renda} + 2.549384e(-4) * \text{Vagas} + 1.191090e(-4) * \text{Quartos} - 6.545645e(-6) * \text{Área} + 4.005336e(-7) * \text{Data.de.Refência.da.Avaliação} - 4.951252e(-6) * \text{Unidades.no.Andar} + 2.576988e(-5) * \text{Nº.de.pavimentos} + 4.724179e(-5) * \text{Padrão.das.Edificações_Alto (por predominância)} - 2.402170e(-4) * \text{Padrão.das.Edificações_Baixo} + 1.195131e(-4) * \text{Padrão.das.Edificações_Normal (c/ aspectos de alto)} - 1.756066e(-4) * \text{Padrão.das.Edificações._Normal (c/ aspectos de baixo)} + 1.052301e(-4) * \text{Estado.de.Conservação.do.prédioBom (aparência de novo)} - 1.388644e(-4) * \text{Estado.de.Conservação.do.prédio_Regular (reparos importantes)} + 2.108606e(-6) * \text{Estado.de.Conservação.do.prédio._Regular (reparos simples)} + 1.822399e(-4) * \text{Idade.Estimada.da.edificação_} \leq 5 - 1.967548e(-5) * \text{Idade.Estimada.da.edificação_} > 10 < 20 - 4.119837e(-5) * \text{Idade.Estimada.da.edificação_} > 20 \leq 50 - 5.089637e(-4) * \text{Idade.Estimada.da.edificação_} > 50 \leq 100 - 1.736681e(-5) * \text{Equipamentos.de.Segurança_SIM} + 5.576315e(-5) * \text{Espaçocom.Churrasqueira_NÃO} -$$

$1.103170e(-4) * \text{Mini. Quadra. Esportiva_SIM} - 3.066303e(-4) * \text{Piscina_NÃO} +$
 $6.408578e(-5) * \text{Portaria. e. ou. Guarita_NÃO} - 5.860057e(-5) * \text{Salãode. Festas_NÃO} +$
 $5.008358e(-5) * \text{Outros. Salões. de. Lazer_SIM} + 8.892471e(-5) *$
 $\text{Sauna... Ofurô... Hidromassagem_SIM} - 1.267005e(-4) * \text{Quadra. Poliesportiva_SIM} +$
 $9.446337e(-5) * \text{Pilotis_SIM} + 2.192256e(-4) *$
 $\text{Acabamento. Fachada. Principal_Cerâmica de Boa Qualidade} + 2.982842e(-4) *$
 $\text{Acabamento. Fachada. Principal_Granito ou Compatível} + 2.596152e(-4) *$
 $\text{Acabamento. Fachada. Principal. da. Edificação_Pastilha} + 1.462342e(-4) *$
 $\text{Acabamento. Fachada. Principal_Porcelanato} - 1.588913e(-4) * \text{Esquadrias. Fachada_Ferro} -$
 $1.071817e(-4) * \text{Esquadrias. Fachada. Principal_Madeira} + 1.186416e(-5) *$
 $\text{Esquadrias. Fachada. Principal_Vidro Temperado} + 6.027505e(-4) *$
 $\text{Padrão. de. acabamento. da. Unidade_Alto (superior, luxo)} + 2.001837e(-4) *$
 $\text{Padrão. de. acabamento. da. Unidade_Normal (c/ aspectos de alto)} - 2.189269e(-4) *$
 $\text{Padrão. de. acabamento. da. Unidade_Normal (c/ aspectos de baixo)} + 1.217484e(-5) *$
 $\text{Estado. de. Conservação. da. Unidade_Bom (aparência de novo)} - 1.531842e(-4) *$
 $\text{Estado. de. Conservação. da. Unidade_Regular (reparos importantes)} - 7.454317e(-5) *$
 $\text{Estado. de. Conservação. da. Unidade_Regular (reparos simples)] * 0.7199462 + 1}^{(1/-0.7199462)}$

Foi elaborado novo histograma de distribuição de resíduos devido a impossibilidade de executar o teste de Shapiro-Francia e o resultado está na Figura 5. Foi realizado também o teste de Breusch-Pagan e ficou esclarecido que a distribuição de resíduos é heterocedástica conforme Figura 6.

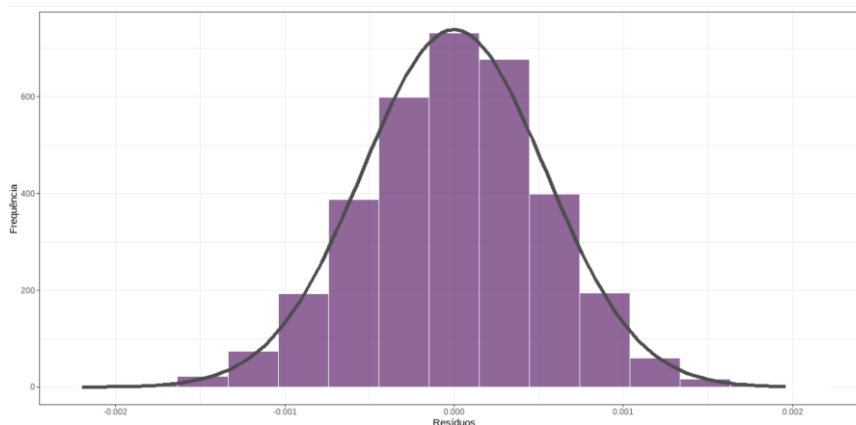


Figura 5. Histograma de distribuição dos resíduos da regressão com transformação de Box-Cox
 Fonte: Dados originais da pesquisa

Breusch Pagan Test for Heteroskedasticity

Ho: the variance is constant
 Ha: the variance is not constant

Data

Response : bcRm2
 Variables: fitted values of bcRm2

Test Summary

DF = 1
 Chi2 = 22.19162
 Prob > Chi2 = 2.467479e-06

Figura 6. Resultado do teste de Breusch-Pagan para a regressão com transformação de Box-Cox
 Fonte: Dados originais da pesquisa

Dados	13556
Dados Considerados	13556
Variáveis	42
Variáveis Consideradas	33
Coefficientes	
Correlação	51 - 0.7712 / -1.#IND
Determinação	51 - 0.5948 / -1.#IND
R2 Ajustado	51 - 0.5938 / -1.#IND
Testes de Hipóteses	
F Calculado	620.2
Significância Modelo	0.01
D Calculado	
Durbin Watson	
Normalidade dos Resíduos	
-1 e +1 desvios padrões	71%
-1,64 e +1,64 desvios padrões	90%
-1,96 e +1,96 desvios padrões	94%
Diversos	
Desvio Padrão	3.5535E-008
Outliers do Modelo	692 (5.10%)
Método de Cálculo	Geral

Figura 9. Principais estatísticas do modelo elaborado no SisDEA
 Fonte: Dados originais da pesquisa

A seguir serão descritos os procedimentos para obtenção de modelos de predição a partir de técnicas de *machine learning*, sendo elas: *Árvore de Regressão*, *Random Forest* e *Deep Learning*. Para desenvolver os modelos optou-se por dividir o banco de dados em dois novos sendo um para treino e outro para teste. O banco de dados de treino ficou com 75% das observações (10191) e o banco de dados de teste ficou com 25% das observações (3365), totalizando 13556 observações. A definição de quais observações fazem parte de cada banco de dados foi feita de forma aleatória, a partir da ferramenta *runif* do pacote *stats*. Para garantir a reprodutibilidade do processo utilizou o recurso *setseed* com parâmetro 123.

Árvore de Regressão e *Random Forest*

Os modelos com árvore de regressão foram construídos no Rstudio a partir da ferramenta *rpart* do pacote *rpart*. Inicialmente foi construída um modelo utilizando a base de treino com árvore de profundidade 10 e custo de complexidade (CP) igual a zero. A segunda árvore foi construída com profundidade 30, custo de complexidade (CP) igual a zero e 10 *cross validation* (*xval* = 10). Foram aplicadas as ferramentas *printcp* e *plotcp* do pacote *rpart* para construir o gráfico da Figura 10 contendo os custos de complexidade (CP) e os erros em validação cruzada (do inglês *cross validation*).

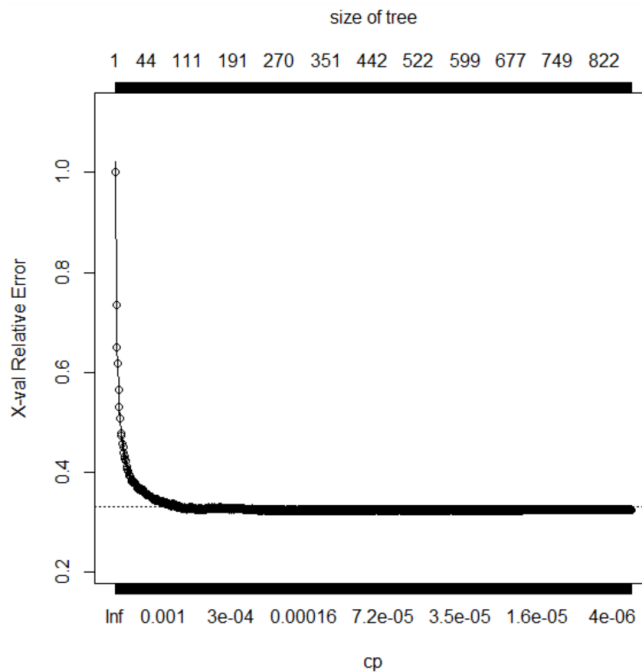


Figura 10. Erros de *cross validation* em função do custo de complexidade

Fonte: Dados originais da pesquisa

O próximo passo foi obter o melhor CP para validação cruzada (CPmin) para a segunda árvore, cujo valor foi CPmin = 0.0001836985. A segunda árvore foi então recalculada com CP igual ao CPmin, chegando-se então a um modelo mais ajustado.

Construiu-se também um modelo com a técnica *Random Forest* com o objetivo de alcançar um poder de predição maior. Foi utilizada a ferramenta randomForest do pacote randomForest. O modelo foi moldado utilizando a base de treino e o parâmetro ntree = 10. Em seguida foi realizado o processo de *tuning* do modelo *Random Forest* utilizando a ferramenta train do pacote caret, adotando os parâmetros method = cv (k-fold cross validation) com 3 folds e tuneGrid = (3, 4, 5, 6, 7, 8). A Figura 11 contém a plotagem do RMSE em função do preditor selecionado, demonstrando que existe possibilidade de existirem bons preditores fora do intervalo selecionado. Dessa forma, realizamos novo processo de *tuning* utilizando o tuneGrid = (15, 20, 25, 30, 35, 40) e 15 folds. Nesse sentido, a Figura 12 contém a plotagem do RMSE em função do preditor para a nova configuração e demonstra que a melhor configuração parece estar próxima do preditor 20.

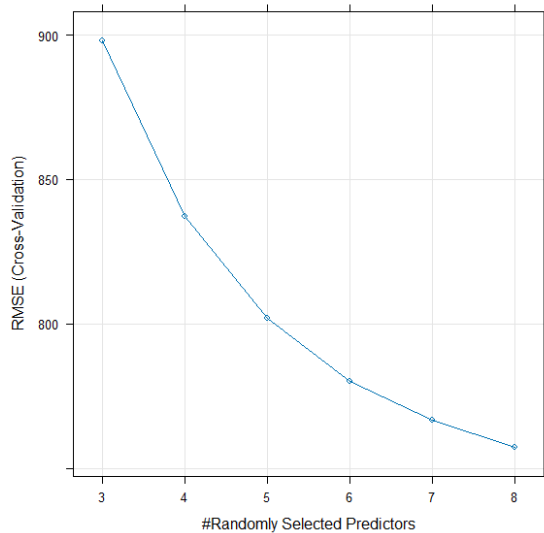


Figura 11 – Root Mean Square Error [RMSE] em função dos preditores 3, 4, 5, 6, 7 e 8
 Fonte: Dados originais da pesquisa

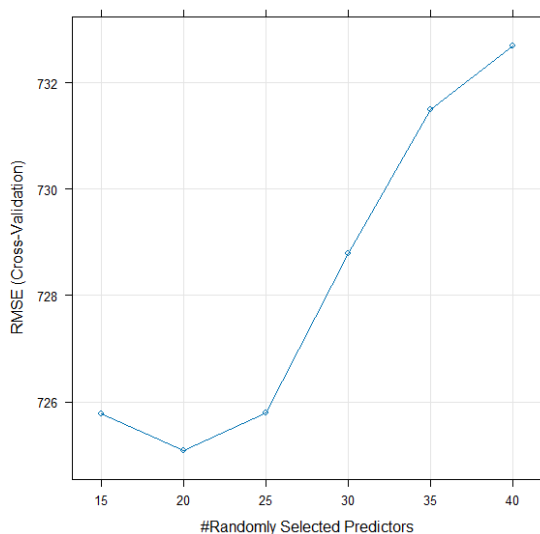


Figura 12 – Root Mean Square Error [RMSE] em função dos preditores 15, 20, 25, 30, 35 e 40
 Fonte: Dados originais da pesquisa

Para avaliação do coeficiente de determinação (R^2), Raiz do Erro Quadrático Médio (em inglês *root mean square error* [RMSE]) e Erro absoluto médio (em inglês *mean absolute error* [MAE]) de cada modelo de árvore de regressão e *Random Forest* foi utilizada a ferramenta *postResample* do pacote *caret*. A Tabela 6 contém o resultado da avaliação utilizando a base de treino e base de teste. As estatísticas R^2 , RMSE e MAE são largamente utilizadas como medidas do poder de predição de modelos (Chicco et al., 2021). O R^2 demonstra o grau de ajuste do modelo, o MAE representa a diferença em módulo entre valores calculados e observados e o RMSE foca nos valores extremos das diferenças calculadas no MAE (Alsayadi et al., 2022). Dessa forma, a presente pesquisa comparou os diferentes modelos desenvolvidos a partir das três estatísticas.

Tabela 6. Coeficiente de determinação, *Root Mean Square Error* [RMSE] e *Mean Absolute Error* [MAE] de cada modelo de Árvore de Regressão e *Random Forest*.

Modelo	Base de Treino			Base de Teste		
	R ²	RMSE	MAE	R ²	RMSE	MAE
Arvore de Regressão com xval = 1, maxdepth=10 e CP=0	0,8030	728	503	0,7063	872	597
Arvore de Regressão com xval = 10, maxdepth=30 e CP=0	0,8421	652	432	0,7117	869	588
Arvore de Regressão com xval = 0, maxdepth=30 e CP=0.0001836985	0,8088	718	507	0,7040	876	601
Randon Forest com ntree=10	0,9472	391	251	0,7886	731	498
Randon Forest com tuning com grid de hiperparâmetros c(3,4,5,6,7,8), method = cv (kfold com cross validation) e 5 folds	0,9578	372	254	0,8129	698	478
Randon Forest com tuning com grid de hiperparâmetros c(15, 20, 25, 30, 35, 40), method = cv (kfold com cross validation) e 15 folds	0,9671	318	209	0,8211	674	459

Fonte: Dados originais da pesquisa

Deep Learning

Como forma de buscar modelos mais ajustados ao mercado imobiliário local, foram elaborados modelos usando a técnica de *deep learning*, ou seja, Redes Neurais Artificiais com camadas ocultas. Inicialmente foi realizada a padronização dos valores das variáveis, de forma que todas assumam valores entre zero e um.

Os modelos foram desenvolvidos utilizando a ferramenta neuralnet do pacote neuralnet. Foram elaborados modelos a partir de redes neurais com variadas quantidades de camadas ocultas e de neurônios por camada. Por fim, tendo como objetivo melhor o tuning, foi utilizada a ferramenta train para treinar e otimizar algumas configurações de Redes Neurais. Os parâmetros utilizados em cada modelo bem como suas avaliações estão condensados na Tabela 7.

Tabela 7. Coeficiente de determinação, *Root Mean Square Error* [RMSE] e *Mean Absolute Error* [MAE] de cada modelo de Rede Neural Artificial e *Deep Learning*

Modelo	Base de Treino			Base de Teste		
	R ²	RMSE	MAE	R ²	RMSE	MAE
Rede Neural Artificial simples com linear.output = True e threshold=0,01	0,6855	2920	1996	0,6825	2896	1906
Rede Neural Artificial com duas camadas ocultas com 3 e 4 neurônios, linear.output = True e threshold=0,01	0,7653	2795	1996	0,7497	2798	1918

Rede Neural Artificial com uma camada oculta com 15 neurônios, linear.output = True e threshold=0,1	0,8092	2717	1996	0,7318	2832	1915
Rede Neural Artificial com duas camadas ocultas com 7 e 3 neurônios, linear.output = True e threshold=0,1	0,7871	2757	1996	0,7684	2767	1922
Rede Neural Artificial com uma camada oculta com 7 neurônios, linear.output = True e threshold=0,1	0,7589	2782	1996	0,7828	2765	1921
Rede Neural Artificial com três camadas ocultas com 7, 12 e 15 neurônios, linear.output = True e threshold=0,1	0,7755	2778	1996	0,7471	2801	1917
Rede Neural Artificial com tuning utilizando .size=7, .decay = c(0, .005, .01, 0.015, 0.02, 0.025), method=nnet e maxit=1000	0,7823	2766	1996	0,7600	2780	1919
Rede Neural Artificial com tuning utilizando .size=15, .decay = 15 c(0, .005, .01, 0.015, 0.02, 0.025), method=nnet e maxit=1000	0,8041	2727	1996	0,7548	2789	1917

Fonte: Dados originais da pesquisa

Por fim, a Tabela 8 contém as características de todos os modelos desenvolvidos. Para melhor visualização os modelos foram ordenados conforme desempenho tendo como critérios (i) maior coeficiente de determinação, (ii) menor RMSE e (iii) menor MAE.

Tabela 8. Coeficiente de determinação, *Root Mean Square Error* [RMSE] e *Mean Absolute Error* [MAE] de cada modelo desenvolvido na pesquisa

Modelo	Base de Treino			Base de Teste		
	R ²	RMSE	MAE	R ²	RMSE	MAE
Regressão SisDEA ^{(1) (3)}				0,5938	1175	746
Regressão com transformação de Box-Cox e procedimento stepwise ^{(1) (2)}				0,6126	1033	688
Regressão Linear Rstudio com procedimento stepwise ⁽¹⁾				0,6445	971	713
Rede Neural Artificial simples com linear.output = True e threshold=0,01	0,6855	2920	1996	0,6825	2896	1906
Arvore de Regressão com xval = 0, maxdepth=30 e CP=0.0001836985	0,8088	718	507	0,7040	876	601
Arvore de Regressão com xval = 1, maxdepth=10 e CP=0	0,8030	728	503	0,7063	872	597
Arvore de Regressão com xval = 10, maxdepth=30 e CP=0	0,8421	652	432	0,7117	869	588

Rede Neural Artificial com uma camada oculta com 15 neurônios, linear.output = True e threshold=0,1	0,8092	2717	1996	0,7318	2832	1915
Rede Neural Artificial com três camadas ocultas com 7, 12 e 15 neurônios, linear.output = True e threshold=0,1	0,7755	2778	1996	0,7471	2801	1917
Rede Neural Artificial com duas camadas ocultas com 3 e 4 neurônios, linear.output = True e threshold=0,01	0,7653	2795	1996	0,7497	2798	1918
Rede Neural Artificial com tuning utilizando .size=15, .decay = 15 c(0, .005, .01, 0.015, 0.02, 0.025), method=nnet e maxit=1000	0,8041	2727	1996	0,7548	2789	1917
Rede Neural Artificial com tuning utilizando .size=7, .decay = c(0, .005, .01, 0.015, 0.02, 0.025), method=nnet e maxit=1000	0,7823	2766	1996	0,7600	2780	1919
Rede Neural Artificial com duas camadas ocultas com 7 e 3 neurônios, linear.output = True e threshold=0,1	0,7871	2757	1996	0,7684	2767	1922
Rede Neural Artificial com uma camada oculta com 7 neurônios, linear.output = True e threshold=0,1	0,7589	2782	1996	0,7828	2765	1921
Randon Forest com ntree=10	0,9472	391	251	0,7886	731	498
Randon Forest com tuning com grid de hiperparâmetros c(3,4,5,6,7,8), method = cv (kfold com cross validation) e 5 folds	0,9578	372	254	0,8129	698	478
Randon Forest com tuning com grid de hiperparâmetros c(15, 20, 25, 30, 35, 40), method = cv (kfold com cross validation) e 15 folds	0,9671	318	209	0,8211	674	459

Fonte: Dados originais da pesquisa

Nota: ⁽¹⁾ Não foi realizada divisão do Banco de Dados em treino e teste para essas regressões;

⁽²⁾ Os fitted values foram transformados para a forma inicial e comparados com os valores sem transformação de Box-Cox da variável dependente para cálculo do R², RMSE e MAE; ⁽³⁾ O RMSE e o MAE foram calculados no Rstudio porque o SisDEA não apresenta essas estatísticas para o modelo.

Resultados e Discussão

Conforme pontuado, grande parte das avaliações de imóveis é realizada com auxílio de aplicativos que são dedicados a essa atividade e que calculam os modelos a partir de regressões lineares e não lineares. A Tabela 8 apresenta a capacidade de predição de cada modelo desenvolvido para essa pesquisa a partir das estatísticas R², RMSE e MAE. Ficou demonstrado que os modelos com pior desempenho, ou seja, com pior capacidade de predição, são os tipos de modelo tradicionalmente utilizados pelos avaliadores de imóveis.

A regressão calculada pelo aplicativo SisDEA foi a que apresentou menor capacidade preditiva e as regressões calculadas no Rstudio tiveram segundo e terceiro pior desempenho.

Por outro lado, modelos obtidos através de Redes Neurais Artificiais com camadas ocultas (*deep learning*) ou Árvore de Regressão demonstraram desempenho intermediário, porém superiores as regressões lineares ou não lineares. Por fim, os modelos obtidos através de *Randon Forest* apresentaram os melhores resultados, chegando o modelo mais ajustado a um R^2 de 0,8211, RMSE de 674 e MAE de 459 na base de teste. Considerando que a média dos valores observados na variável dependente é de R\$3990/m², um MAE de R\$459/m² significa variação média de 11,5% entre valores observados e valores calculados.

2. CONCLUSÃO

O uso de técnicas de *machine learning* para avaliação de imóveis se demonstrou uma consistente alternativa. A pesquisa utilizou as técnicas de Redes Neurais Artificiais com Camadas Ocultas (*deep learning*), Árvores de Regressão e *Randon Forest*. A capacidade de predição dos modelos foi avaliada a partir das estatísticas R^2 , RMSE e MAE. Todos os modelos de *machine learning* testados para a base de dados demonstraram capacidade de predição superior aos modelos de regressão linear e não linear, que são os comumente utilizados pelos avaliadores de imóveis. O modelo de regressão linear ou não linear com melhor capacidade de predição apresentou R^2 de 0,6445, RMSE de 971 e MAE de 971 enquanto o modelo de *machine learning* com melhor capacidade de predição (*Randon Forest*) apresentou R^2 de 0,8211, RMSE de 674 e MAE de 459 na base de teste.

Sugerimos que estudos futuros verifiquem a superioridade das técnicas de *machine learning* para avaliação de imóveis a partir de outras bases de dados com quantidades de observações diferentes, principalmente para amostras menores e diferentes mercados.

REFERÊNCIAS

Alsayadi, H.A.; Abdelhamid, A.A.; El-Kenawy, E.S.; Ibrahim, A.; Eid, M.M. 2022. Ensemble of Machine Learning Fusion Models for Breast Cancer Detection Based on the Regression Model. *Fusion: Practice and Applications*. 9: 19-26.

Associação Brasileira de Normas Técnica [ABNT]. 2019. Avaliação de bens Parte 1: Procedimentos gerais. Rio de Janeiro, RJ, Brasil.

Azeroual, O. 2020. DataWrangling in Database Systems: Purging of Dirty Data. *Data*. 5: 1-9.

Banco Central do Brasil [BCB]. 2012. Resolução Nº 4.088, de 24 de maio de 2012. Dispõe acerca do registro de informações referentes às garantias constituídas. *Diário Oficial da União*, Brasília, 28 mai. 2012. Seção 1, p.29.

Banco Central do Brasil [BCB]. 2022. Resolução BCB Nº 217, de 30 de março de 2022. Dispõe sobre as condições para o registro das informações relativas às garantias constituídas sobre veículos automotores e imóveis, bem como das informações sobre a propriedade de veículos automotores, relacionadas às operações que especifica. *Diário Oficial da União*, Brasília, 1 abr. 2022. Seção 1, p.375.

Banco Central do Brasil [BCB]. 2024. Informações do Mercado Imobiliário. Disponível em <<https://www.bcb.gov.br/estatisticas/mercadoimobiliario>>. Acesso em: 31 mar. 2024.

Bank for International Settlements [BIS]. 2017. Basel III: Finalising post-crisis reforms. Disponível em < <https://www.bis.org/bcbs/publ/d424.htm> >. Acesso em 31 mar. 2024.

Brasil. 1997. Lei 9.514, de 20 de novembro de 1997. Dispõe sobre o Sistema de Financiamento Imobiliário, institui a alienação fiduciária de coisa imóvel e dá outras providências. Diário Oficial da União, Brasília, 21 nov. 1997. Seção 1, p. 27182.

Caixa Econômica Federal. 2019. Edital de Convocação Nº 0003/2019. Brasília, DF, Brasil.

Chicco, D.; Warrens, M.J.; Jurman, G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science. 7: 1-24.

Dantas, R.A.; Magalhães, A.M.; Vergolino, J.R.O. 2007. Avaliação de imóveis: a importância dos vizinhos no caso de Recife. Economia Aplicada. 11: 231-251.

Doszyń, M. 2022. Econometric Models of Real Estate Prices with Prior Information. Mixed Estimation. Real Estate Management and Valuation. 30: 61-72.

Duarte, L.S.; Hochheim, N.; Reginato, V.S.C. Regressão linear múltipla aplicada em avaliação imobiliária - Estudo de caso na área central de Florianópolis. Revista Concilium 22:880-889.

Fávero, L.P.; Belfiore, P. 2022. Manual de análise de dados. 1 ed. LTC, Rio de Janeiro, RJ, Brasil.

Fiker, J. 2017. Manual de avaliações e perícias em imóveis urbanos. 5ed. Oficina de Textos, São Paulo, SP, Brasil.

Instituto Brasileiro de Avaliações e Perícias de Engenharia de São Paulo [IBAPE/SP]. 2005. Norma para Avaliação de Imóveis Urbanos. São Paulo, SP, Brasil.

Pelli Sistemas. 2024. SisDEA Avaliação de Imóveis Urbanos e Rurais. Disponível em <<https://pellisistemas.com/software/sisdea-avaliacao-de-imoveis/>>. Acesso em: 31 mar. 2024.

Regression Engenharia. CASTLER. Disponível em <<http://www.regression.com.br/index.php/castler>>. Acesso em: 31 mar. 2024.

Sistema Integrado de Bases Geoespaciais do Estado do Espírito Santo [Geobases]. 2018. PDUI - Renda Média dos Domicílios. Disponível em <https://geobases.es.gov.br/https://ide.geobases.es.gov.br/layers/geonode_data:geonode:pdui_media_renda_mensal>. Acesso em 15 fev. 2024.

TecSys Engenharia. 2024. TS-SISREG. Disponível em <<https://tecsys.eng.br/>>. Acesso em: 31 mar. 2024.

Vergara, S.C. 2016. Projetos e Relatórios de Pesquisa em Administração. 16 ed. Atlas. São Paulo, SP, Brasil.

Vos, D.; Döhmen, T.; Schelter, S. 2022. Towards Parameter-Efficient Automation of Data Wrangling Tasks with Prefix-Tuning. Table Representation Learning Workshop at NeurIPS.

Disponível em <<https://openreview.net/forum?id=8kyYJs2YkFH>>. Acessado em 22 mar. 2024.

Yin, R.K. 2014. Estudo de Caso: Planejamento e Métodos. 5ed. Bookman. Porto Alegre, RS, Brasil.