INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À ENGENHARIA E GESTÃO

RICARDO DA SILVA TAVARES

FRAMEWORK AUTOMATIZADO PARA AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Campos dos Goytacazes/RJ

2021

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À ENGENHARIA E GESTÃO

RICARDO DA SILVA TAVARES

FRAMEWORK AUTOMATIZADO PARA AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Prof. Dr. Renato Gomes Sobral Barcellos
(Orientador)

Prof. Dr. Henrique Rego Monteiro da Hora
(Coorientador)

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação do Instituto Federal de Educação, Ciência e Tecnologia Fluminense, no Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão (MPSAEG), como parte dos requisitos necessários à obtenção do título de Mestre em Sistemas Aplicados à Engenharia e Gestão.

Campos dos Goytacazes/RJ
2021

Biblioteca Anton Dakitsch
CIP - Catalogação na Publicação

S586f

Silva Tavares, Ricardo
FRAMEWORK AUTOMATIZADO PARA AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS / Ricardo Silva Tavares - 2021.
69 f.: il. color.

Orientador: Renato Gomes Sobral Barcellos
Coorientador: Henrique Rego Monteiro da Hora

Dissertação (mestrado) -- Instituto Federal de Educação, Ciência e Tecnologia Fluminense, Campus Campos Centro, Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão, Campos dos Goytacazes, RJ, 2021.
Referências: f. .

1. Evaluation. 2. Real Estate. 3. Data Mining. 4. Machine Learning. 5. Framework. I. Gomes Sobral Barcellos, Renato, orient. II. Rego Monteiro da Hora, Henrique, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca Anton Dakitsch do IFF com os dados fornecidos pelo(a) autor(a).

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

IFFLUMINENSE *CAMPUS* CAMPOS CENTRO

PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À ENGENHARIA E GESTÃO

RICARDO DA SILVA TAVARES

FRAMEWORK AUTOMATIZADO PARA AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação do Instituto Federal de Educação, Ciência e Tecnologia Fluminense, no Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão (MPSAEG), como parte dos requisitos necessários à obtenção do título de Mestre em Sistemas Aplicados à Engenharia e Gestão.

Aprovado em 12 de julho de 2021.

BANCA EXAMINADORA:

Prof. Dr. Renato Gomes Sobral Barcellos
Doutor em Geociências - IFFluminense
(Orientador)

Prof. Dr. Henrique Rego Monteiro da Hora
Doutor em Engenharia de Produção - IFFluminense
(Coorientador)

Profa. Dra. Altina Silva Oliveira Dionisio
Doutora em Engenharia de Produção - UFF

Prof. Dr. Gustavo Saad Terra
Doutor em Engenharia Civil - IFFluminense

Dedico este trabalho à minha querida mãe Antônia Rita, que sempre esteve ao meu lado, me educando, incentivando e apoiando nos estudos, trabalhos e em todas as partes da minha vida.

Ela não estará fisicamente presente no término dessa etapa, mas sempre estará comigo espiritualmente e no meu coração. Te amo mãe.

AGRADECIMENTOS

Agradeço a Deus por todas as oportunidades colocadas na minha vida. Sua bondade é infinita, e sem ela essa caminhada não seria possível.

À minha mãe Antônia Rita e ao meu pai Antônio Carlos, que sempre me deram todo apoio e carinho para que eu seja uma pessoa melhor.

À minha irmã Simone, que está sempre comigo, nos momentos bons e ruins, não me deixando desistir quando as forças estão quase no fim.

Ao meu sobrinho Pedro Henrique, que trouxe muita alegria quando tudo parecia perdido. Sua gargalhada consegue alegrar meu coração mesmo quando os dias estão tortuosos.

Aos laços de amizades criados no mestrado, em especial ao meu grande amigo Renato Vale, meu irmão, que sempre terá um lugar especial no meu coração.

Ao meu orientador, Renato Barcellos, pela paciência e compreensão. Suas palavras de apoio e seus ensinamentos foram um dos pilares que me ajudaram chegar até esse momento.

Minha eterna gratidão ao meu coorientador, professor Henrique. Não existem palavras para agradecer todos seus ensinamentos e conhecimento que desenvolvi com suas orientações.

Aos docentes do Instituto Federal Fluminense campus Campos Centro (IFF campus Campos Centro), minha casa há mais de 20 anos. Muito obrigado a todos pela contribuição na ampliação dos conhecimentos.

LISTA DE FIGURAS

Figura 1.1: Unidades residenciais lançadas – Acumulado 12 meses – Fonte: (CBIC, 2020) ..11

Figura 2.1: Fluxograma de procedimentos para seleção de artigos para revisão sistemática. .20

Figura 2.2: Publicações de artigos selecionados para revisão distribuídos anualmente.....21

Figura 3.1: Crescimento do mercado imobiliário americano. Fonte: (BUREAU, 2021).....35

Figura 3.2: Expectativa de comercialização de unidade residenciais no Canadá. Fonte: (CREA, 2021).....36

Figura 3.3: Unidades residenciais vendidas por região do Brasil – Fonte: (CBIC, 2020) 36

Figura 3.4: Método DEA sob Dupla Ótica. Fonte: (LINS; NOVAES; LEGEY, 2005).....44

Figura 3.5: Etapas do DCBC de acordo com (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).....46

Figura 3.6: Processo de análise de eficiência de modelos através da validação cruzada. 55

Figura 3.7: Gráfico de correlação entre as variáveis. 59

Figura 3.8: Comparação de performance dos modelos: (a) RMSE, (b) MAPE e (c) R^2 60

LISTA DE TABELAS

Tabela 2.1: Query de consulta nas bases de dados 19

Tabela 2.2: Resumo das técnicas de aquisição de dados aplicados nos estudos selecionados. 22

Tabela 2.3: Quantidade de dados utilizados para treinamento de modelos de valoração de imóveis. 22

Tabela 2.4: Modelos utilizados em avaliações patrimoniais nos trabalhos revisados. 23

Tabela 3.1: Definição dos atributos do *dataset* 56

Tabela 3.2: Resumo de informações contidas no dataset 58

SUMÁRIO

1. INTRODUÇÃO 11

1.1. MOTIVAÇÃO 11

1.2. PROBLEMA 13

1.3. OBJETIVO 15

2. ARTIGO 01: AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS: REVISÃO SISTEMÁTICA 17

2.1. RESUMO 17

2.2. ABSTRACT 17

2.3. INTRODUÇÃO 18

2.4. METODOLOGIA 19

2.4.1. IDENTIFICAÇÃO E COLETA DOS TRABALHOS RELACIONADOS 19

2.4.2. PROCEDIMENTOS TÉCNICOS 20

2.5. RESULTADOS 21

2.5.1. EXTRAÇÃO E LIMPEZA DE DADOS 23

2.5.2. TÉCNICAS DE MACHINE LEARNING APLICADA A AVALIAÇÃO PATRIMONIAL 24

2.5.3. MÉTODOS DE AVALIAÇÃO DOS MODELOS 25

2.6. DISCUSSÃO 26

2.7. CONCLUSÃO 29

2.8. REFERÊNCIAS BIBLIOGRÁFICAS 29

3. ARTIGO 02: AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS: ESTUDO DE CASO 38

3.1. RESUMO 38

3.2. ABSTRACT 38

3.3. INTRODUÇÃO 39

3.4. REFERENCIAL TEÓRICO 44
3.4.1. VALOR DE MERCADO, PREÇO E CUSTO 44
3.4.2. METODOLOGIAS APLICÁVEIS PARA AVALIAÇÃO PATRIMONIAL 45
3.4.3. DESCOBERTA DO CONHECIMENTO EM BASE DE DADOS 49
3.4.4. TÉCNICAS PARA COLETA DE DADOS DISPERSOS NA WEB 51
3.4.5. ALGORITMOS DE MACHINE LEARNING 51
3.5. METODOLOGIA 54
3.5.1. FERRAMENTAL DE APOIO 54
3.5.2. DESENVOLVIMENTO DO FRAMEWORK 56
3.5.3. FASE 1: EXTRAÇÃO DE DADOS DE ANÚNCIO DE IMÓVEIS DISPERSOS EM SITES DA INTERNET 57
3.5.4. FASE 2: LIMPEZA, ORGANIZAÇÃO E TRANSFORMAÇÃO 58
3.5.5. FASE 3: TREINAMENTO E AVALIAÇÃO DA PERFORMANCE DOS ALGORITMOS DE MACHINE LEARNING 59
3.6. RESULTADOS 62
3.7. CONCLUSÕES 67
3.8. REFERÊNCIAS BIBLIOGRÁFICAS 68

1. INTRODUÇÃO

1.1. MOTIVAÇÃO

O mercado imobiliário e da construção civil constituem um dos principais setores industriais no Brasil, foram responsáveis por uma receita líquida aproximada de R$ 2.56 bilhões, correspondendo a 4,36% do PIB brasileiro.

Empregam quase 2 milhões de funcionários formais, de acordo com os resultados apresentados na última Pesquisa Anual da Indústria da Construção (IBGE, 2017). Em 2019, foi um dos principais setores empregadores no Brasil, ofertando mais de 71 mil novas vagas formais (MTE, 2020).

De acordo com os Indicadores Imobiliários Nacionais divulgado pela Câmara Brasileira da Indústria da Construção (CBIC), em 2019 foram lançadas mais de 185 mil unidades residenciais e em 2020, somados os primeiros 3 trimestres, são cerca de 87 mil novas residências (CBIC, 2020).

Mesmo com a crise pandémica provocada pelo COVID-19, o mercado imobiliário ainda apresenta desde a crise imobiliária brasileira que levou a sua queda em 2015, conforme é possível observar na Figura 1.1.

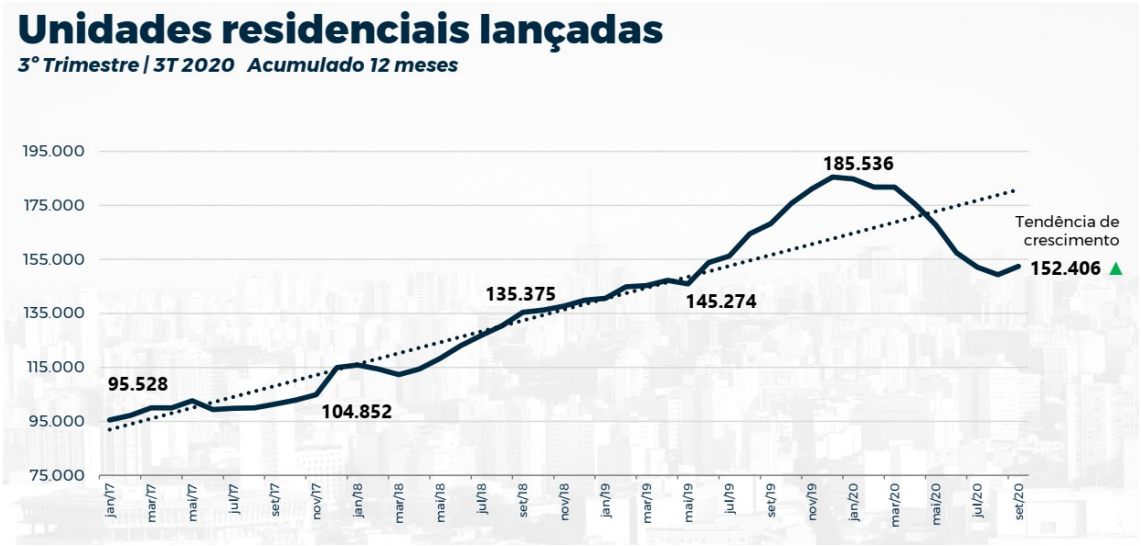

Figura 1.1: Unidades residenciais lançadas – Acumulado 12 meses – Fonte: (CBIC, 2020)

Segundo Pelli (2006), órgãos governamentais e privados utilizam o valor de mercado de imóveis para tomada de decisões, trazendo grande importância para a economia, como

arrecadação de tributos, impostos e taxas, sendo estas calculadas sobre o valor venal dos imóveis.

A avaliação de imóveis é uma análise complexa, que resulta na determinação do valor de mercado do bem avaliado, sendo este impactado por fatores como características físicas do imóvel, geolocalização e fatores político-econômicos (MELANDA; HUNTER; BARRY, 2016).

As avaliações imobiliárias servem como embasamento para as transações de financiamentos desses bens, ratificando a importância de avaliações precisas. Segundo dados apresentados pela ACEBIP (2020) e pelo BACEN (2020), o volume de crédito contratado para compra de imóveis no ano de 2019 chegou a R$ 58.6 bilhões.

Existem vários métodos para realização de avaliação patrimonial, entretanto a comparação de dados de mercado é o principal método utilizado atualmente.

A segunda parte da Norma Brasileira 14653 de 2011 (NBR 14653-2/2011) apresenta diversos métodos para avaliação patrimonial, enfatizando o método comparativo direto de dados de mercado para identificação de valores de um bem. A norma classifica a avaliação de imóveis pelo método comparativo em dois tipos:

1. O primeiro é o tratamento por homogeneidade de fatores;

2. O segundo é o tratamento científico, sendo este mais utilizado devido ao seu embasamento técnico científico (ABNT, 2011).

É possível utilizar outros modelos científicos, como Regressão Espacial, Análise Envoltória de Dados sob Dupla Ótica (DEA) e Redes Neurais Artificiais (RNA). Entretanto, a norma brasileira faz ressalvas que para utilização destas metodologias deve-se haver uma justificativa plausível sob o ponto de vista teórico e prático.

Os modelos de Regressão Linear Múltipla (RLM) utilizados na Engenharia de Avaliações podem ser considerados hedônicos, pois a formulação do preço de um imóvel está diretamente associada às variáveis independentes, sendo estas representativas de características intrínsecas e extrínsecas associadas a cada bem.

Conforme definido por Rosen (1974), modelos hedônicos para precificação de produtos é um equilíbrio, significando que as vantagens observadas pelo comprador e vendedor orientam as decisões destes em um espaço de características analisadas de um produto, assim como o equilíbrio de mercado.

As avaliações patrimoniais têm se apoiado em técnicas de econometria, como a Regressão Linear, com o cálculo de coeficientes através do Método dos Mínimos Quadrados (MMQ), baseando-se nos modelos de preços hedônicos, para expressar monetariamente os atributos que caracterizam um imóvel avaliando através de dados do mercado, conforme apresentados na Equação 1.1.

$$\boldsymbol{Y = a_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i}$$

Equação 1.1: Equação de Regressão Linear Múltipla

Onde a variável dependente é definida por Y; $\beta_1 \ldots \beta_n$ são os regressores ou coeficientes da regressão; $x_{i1}, \ldots, x_{in}$ representam os valores das variáveis independentes ou explicativas; e ε_i representa os erros aleatórios do modelo.

1.2. PROBLEMA

Devido a heterogeneidade dos imóveis diversas características importantes devem ser analisadas simultaneamente e a modelagem hedônica para obtenção de preços de imóveis baseado em amostras do mercado imobiliário é o principal método utilizado mundialmente, tendo várias pesquisas demonstrando o potencial da técnica (DO; GRUDNITSKI, 1992).

Por meio da Econometria Tradicional os preços implícitos são estimados pela análise de regressão na construção de índices de preços hedônicos (ROSEN, 1974). Esse processo faz uma regressão do preço de um imóvel baseado em suas características físicas, geolocalização e econômicas, chamadas de variáveis independentes.

Entretanto, a utilização da regressão proposta por modelos hedônicos, como a Regressão Linear Múltipla (RLM), vem sendo criticada pela ausência de precisão nos valores obtidos nas avaliações patrimoniais. Tais erros inclusive são aceitos como normais em diversos tribunais do Brasil e do mundo.

Conforme apresentado por Crosby, Lavers e Murdoch (1998), os tribunais ingleses consideram até 15% com uma margem de erro aceitável, mas sem nenhuma base empírica para balizar tal decisão.

A própria norma brasileira 14653-2 adota o "campo de arbítrio" de até 15%, semelhante aos tribunais ingleses, que pode ser utilizado quando variáveis relevantes para avaliação de um imóvel não tiverem sido elencadas na obtenção do modelo devido à escassez de amostras de mercado ou devido a essas variáveis não se apresentaram estatisticamente significantes em modelos de regressão (ABNT, 2011).

Problemas metodológicos associados a RLM são conhecidos há algum tempo e incluem não linearidade, multicolinearidade, má especificação da forma de função e heterocedasticidade (PETERSON; FLANAGAN, 2009).

A preocupação com os pressupostos básicos é considerável, uma vez que o não atendimento deles inviabiliza os modelos hedônicos de regressão.

A NBR 14653-2/2011 deixa claro que as avaliações patrimoniais que utilizam modelagem hedônica para inferir o comportamento do mercado e formação de valores devem ter seus pressupostos explicitados e testados, além de adotar medidas corretivas, quando necessário.

Os pressupostos apresentados na NBR 14653-2/2011 são:

a) Evitar micronumerosidade;

b) Equilíbrio da amostra;

c) Erros devem ser homocedásticos;

d) Erros devem ter distribuição normal;

e) Erros devem ser independentes sob condição de normalidade;

f) Variáveis importantes devem ser incorporadas ao modelo e as variáveis irrelevantes devem ser descartadas;

g) Deve-se evitar multicolinearidade entre as variáveis, devendo-se analisar a coerência das características do imóvel avaliando com a estrutura de multicolinearidade inferida, sendo vedada a utilização do modelo em caso de incoerência;

h) Não pode haver correlações entre o erro aleatório (εi) e as variáveis independentes do modelo;

i) Possíveis pontos influenciantes (outliers) devem ser avaliados e retirados do modelo quando justificados.

Além dos problemas apontados, a avaliação patrimonial ainda conta com o fator subjetividade. Para Steiner *et al.* (2008), a avaliação de imóveis geralmente é realizada de forma subjetiva, com base na experiência pessoal dos profissionais avaliadores, comparando dados do imóvel a ser avaliado com imóveis semelhantes já negociados.

Em González (2002), o autor indica que parte da subjetividade está atrelada à falta de experiência e de entendimento do mercado pelo profissional avaliador, pois este consulta uma amostra de dados de maior interesse, sem analisar o todo. O autor ainda afirma que a coleta de dados dar-se-á informalmente, pois no Brasil não existem bases de dados de transações imobiliárias homologadas por organismos de classe ou empresas.

O problema na coleta de dados é semelhante ao apresentado por Abidoye e Chan (2018) tendo os autores que restringir as análises de seu estudo devido a insuficiência de dados.

Em Zurada, Levitan e Guan (2011), os autores também sofreram com restrições na comparação de Modelos de Regressão devido à falta de dados.

1.3. OBJETIVO

Tendo em vista a importância da correta estimativa de valores das avaliações patrimoniais e as dificuldades elencadas, principalmente na obtenção dos dados necessários para realização da escorreita comparação de mercado, o objetivo geral deste trabalho é desenvolver um framework de avaliação patrimonial automatizado.

O framework passará por todas as fases da descoberta de conhecimento em bases de dados (KDD) baseado em técnicas de Mineração de Dados. Iniciando pela coleta de dados em bases de dados dispersas na web, como *webscraping* e *ETL* (*Extract, Transform and Load*), criando um *Big Data* capaz de permitir uma visão mais holística do mercado imobiliário.

Esse *Big Data* tem a intenção de reduzir erros nas avaliações patrimoniais através de 02 fatores preponderantes. O primeiro é a subjetividade na coleta de amostras de mercado. Já o segundo

fator é a obtenção de quantidade de dados capaz de aprimorr o treinamento de algoritmos de *Machine Leaning* (ML), resultando em aumento da precisão das avaliações baseadas no método comparativo de dados de mercado.

O objetivo geral do presente estudo desdobra-se nos seguintes objetivos específicos:

a) Desenvolver uma revisão sistemática utilizando os estudos científicos encontrados conforme as abordagens distintas de técnicas de Mineração de Dados aplicadas a avaliações patrimoniais;

b) Realizar um estudo de caso para averiguar a performance do framework proposto utilizando dados de anúncios de vendas de apartamentos localizados no bairro de Botafogo, Rio de Janeiro/BR.

2. ARTIGO 01: AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS: REVISÃO SISTEMÁTICA

2.1. RESUMO

Com o crescente volume de dados de transações imobiliárias, as avaliações imobiliárias automatizadas têm sido amplamente estudadas em muitos países para diferentes fins, entretanto, os estudos geralmente baseiam-se em um pequeno conjunto de dados fornecido por empresas imobiliárias. Nesta pesquisa, foi realizada uma revisão sistemática utilizando a metodologia PRISMA proposta por Moher et al. (2009), baseada em quatro conceitos. Primeiramente, a coleta automatizada de dados em bancos de dados dispersos na web; em segundo lugar, a avaliação automatizada de imóveis; terceiro, aplicação de algoritmos de ML; quarto, a previsão de valores imobiliários. As bases de conhecimento utilizadas foram Scopus, Web of Science e Science Direct. Foram encontrados apenas 8 artigos pertinentes ao escopo da pesquisa. Por fim, observou-se que algoritmos de Aprendizado de Máquina aplicados à avaliação patrimonial apresentaram desempenho superior aos modelos hedônicos.

Palavras-chave: Avaliação; Imóvel; Mineração de Dados; Revisão Sistemática

2.2. ABSTRACT

With the increasing volume of real estate transaction data, automated real estate appraisals have been widely studied in many countries for different purposes. Most of them compare the effectiveness of hedonic models with the ML algorithms application, however, the studies are based on a small dataset usually provided by real estate companies. In this research, a systematic review was carried out using on the PRISMA methodology proposed by Moher et al. (2009), based in four concepts. First, automated data collection in dispersed databases on the web; Second, real estate automated appraisal; Third, ML algorithms application; Fourth, real estate values prediction. The knowledge bases used were Scopus, Web of Science, Science Direct and Taylor & Francis. Only 4 papers pertinent to the research scope were found. The results demonstrate that ML models are more accurate than RLM analysis in their ability to predict value. Finally, it was observed that Machine Learning algorithms applied to patrimonial appraisal had a superior performance than hedonic models.

Keywords: Real Estate; Property; Appraisal; Valuation; Data Mining; Systematic Review

2.3. INTRODUÇÃO

O mercado imobiliário proporciona o desenvolvimento urbano, provocando uma grande movimentação financeira e promovendo o desenvolvimento de diversos serviços (NUNES et al., 2019).

Muitos stakeholders tem interesses nas avaliações patrimoniais, empresas imobiliárias, governo, bancos, agências de seguros, os proprietários de imóveis e os compradores interessados. As avaliações são indiscutivelmente mais importantes para o investimento em imóveis do que qualquer outra classe de ativo principal (RICS, 2017).

O RICS Red Book define o termo ‘avaliação’ como uma opinião sobre o valor de mercado de um imóvel, em uma data especificada (ROYAL INSTITUTION OF CHARTERED SURVEYORS, 2017).

O valor de mercado de uma mercadoria é obtido através da comparação de várias outras disponíveis, ou seja, cujas cuja escolhas de produtos semelhantes são possíveis. Quando essas mercadorias são trocadas livremente no mercado, os valores obtidos nessas transações definem o valor de cada uma. Nesse sentido, os modelos de avaliações de imóveis utilizados podem ser considerados hedônicos. A fundamentação teórica do modelo hedônico é baseada na teoria da demanda do consumidor de Lancaster (LANCASTER, 1966).

Portanto, o valor de mercado de um imóvel pode ser identificado como o valor médio ou preço mais provável a ser atingido em transações normais, em um determinado momento (GONZÁLEZ, 2002), no qual comprador e vendedor tem o desejo em realizar a negociação, ambos não compelidos (ABUNAHMAN, 2008), observando as condições de mercado, as características do imóvel e em um determinado tempo.

Tradicionalmente, as avaliações imobiliárias são realizadas por profissionais especialmente treinados. Para os compradores de imóveis, um sistema automatizado de estimativa de preços pode ser útil para estimar os preços dos imóveis atualmente no mercado (KUMKAR et al., 2018b). O uso dessas novas técnicas requer a inclusão, nos grupos tradicionais de avaliação, de novas figuras profissionais, como os analistas de dados (VALIER, 2020).

A assertividade de avaliações tem sido um tópico de pesquisa por muitos anos (KLAMER; BAKKER; GRUIS, 2017) e a inteligência artificial tem provocado uma grande discussão sobre

a temática. A inovação afeta a natureza das avaliações, procedimentos operacionais e as habilidades exigidas do setor profissional (RICS, 2017).

O uso bem-sucedido de modelos de aprendizado de máquina nas estimativas já é percebido, como é caso mais conhecido é o modelo de avaliação de imóveis Zestimate da agência americana Zillow (VALIER, 2020).

A literatura científica também tem lidado extensivamente com o uso de algoritmos de aprendizado de máquina em modelos de previsão automática de valor, entretanto, quase em sua totalidade os estudos restringem-se a um pequeno *dataset* fornecido por agências imobiliárias.

O objetivo deste trabalho é realizar uma revisão sistemática dos modelos de avaliação baseados em técnicas de Mineração de Dados, utilizados para avaliação de bens imóveis, no qual a coleta de dados dar-se-á de forma automatizada, a partir de dados de anúncios de imóveis localizados na internet.

2.4. METODOLOGIA

2.4.1. IDENTIFICAÇÃO E COLETA DOS TRABALHOS RELACIONADOS

Uma revisão sistemática trata-se de uma análise utilizando métodos sistemáticos e explícitos cujo objetivo é identificar, selecionar e avaliar pesquisas relevantes sobre uma temática que está sendo estudada pelo pesquisador (MOHER et al., 2009).

Utilizando os padrões da metodologia PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-analyses*), proposta por (MOHER et al., 2009), a revisão organizada faz uso de métodos coordenados para identificar, selecionar e incluir estudos relacionados a determinado conteúdo.

As bases SCOPUS®, Web of Science® e Science Direct® foram utilizadas para buscar estudos relacionados à *machine learning*, técnicas de coletas de dados em bases dispersas na web e sua aplicação nas avaliações de imóveis em janeiro de 2021.

Os conceitos e os correspondentes tesauros utilizados foram: a) Imóveis – Property, Properties, House, Apartment e Real Estate; b) Avaliação – Valuation, Prediction, Evaluation, Appraisal, Sales, Price, Costs e Forecasting; c) Mineração de Dados – Data Mining, Data Analytic, Data Science, Machine Learning, Big Data, Artificial Intelligence, Computational Intelligence,

KDD; d) Dados da Web – Webscraping, Web Harvest, Web Crawler e Web Mining; e e) Excluir Conferências.

Utilizando os conceitos e os correspondentes tesauros foram criadas as estratégias de buscas, conforme apresentada na Tabela 2.1 abaixo:

Tabela 2.1: Query de consulta nas bases de dados

TITLE-ABS-KEY ("hous pric*" OR "property pric*" OR "real estate" OR "habitation pric*")*	Conceito A
AND	
TITLE-ABS-KEY ("data mining" OR "datamining" OR data-mining OR c45 OR j48 OR "Random forest" OR "decision tree" OR clustering OR knn OR kdd OR "artificial intelligence" OR "neural network" OR ann OR "machine learn" OR "computational intelligence" OR weka)*	Conceito B
AND	
TITLE-ABS-KEY ("valuat" OR "evaluat*" OR prediction OR "apprais*" OR prices OR pricing OR "evaluation engineering")*	Conceito C
AND	
TITLE-ABS-KEY ("web scrap" OR "web harvest*" OR "web crawler" OR scraping OR "web mining")*	Conceito D
AND	
(LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ar"))	Corte de Tipo de Documentos

2.4.2. PROCEDIMENTOS TÉCNICOS

2.4.2.1. SELEÇÃO DOS ESTUDOS E CRITÉRIOS DE INCLUSÃO

Foram adotados os seguintes critérios para encontrar os referenciais teóricos com melhor conexão com o tema proposto:

a) Serem estudos científicos;

b) Estarem disponíveis em inglês;

c) Apresentarem inovação baseado nas técnicas de mineração de dados aplicadas a avaliação patrimonial, apresentando resultados de performance dos modelos e ganhos na assertividade das avaliações.

Foram rejeitados da seleção os trabalhos não pertinentes ao tema, tendo por base os títulos e o resumo dos artigos, e os estudos encontrados em duplicidade nas bases de pesquisa.

2.4.2.2. PROCEDIMENTOS DE ANÁLISE

Para análise dos estudos classificados mediantes aos critérios de inclusão, elencados no tópico anterior, foram seguidas as etapas:

a) Leitura dos títulos e exclusão dos não pertinentes ao tema;

b) Leitura e análise dos resumos;

c) Das pesquisas com resumos pertinentes, leitura dos documentos na íntegra;

2.5. RESULTADOS

A Figura 2.1 demostra o processo de identificação, rastreamento, elegibilidade e seleção dos artigos que serão revisados.

A estratégia de busca nas bases de dados cientificas retornaram um total de 52 artigos, sendo 42 para a base Scopus e 10 para Web of Science. Cabe salientar que não havia resultados na base de dados Science Direct.

Assim, dos 52 estudos que foram encontrados nas demais bases de dados, após a remoção das duplicações e feita uma análise dos títulos, 45 documentos foram analisados. Desses, 8 artigos foram considerados pertinentes ao tema e selecionados para leitura dos documentos na íntegra.

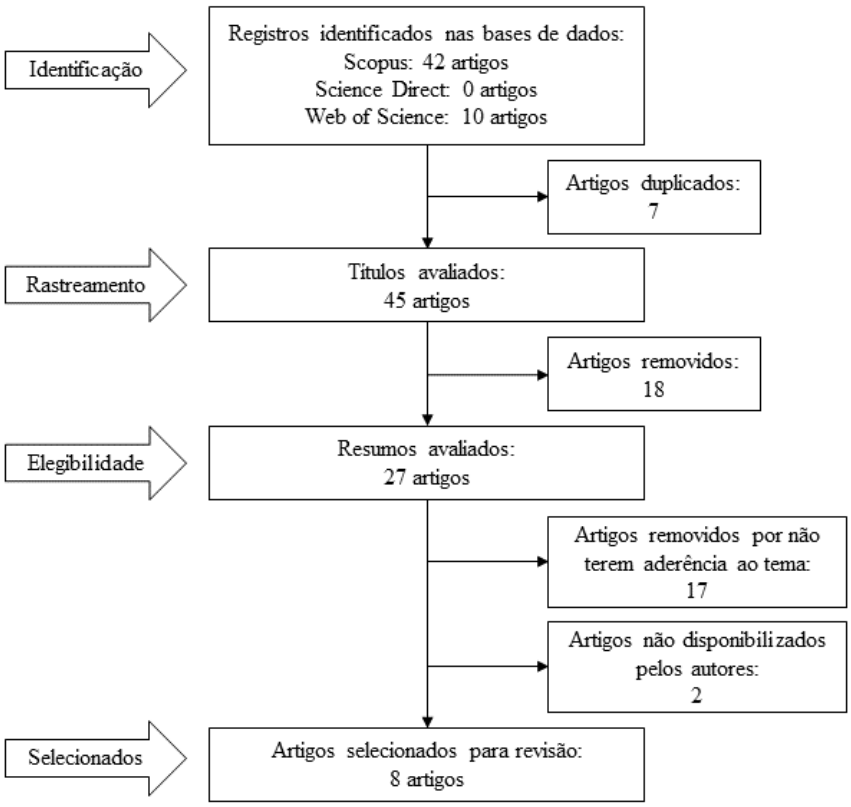

Figura 2.1: Fluxograma de procedimentos para seleção de artigos para revisão sistemática.

Fazendo uma análise da evolução das publicações ao longo do tempo, de acordo com os artigos selecionados para revisão, pode-se afirmar que a área de estudo é bem recente, sendo o artigo mais antigo datado do ano de 2018. Entretanto, é uma área que está em evolução, conforme apresentado na Figura 2.2.

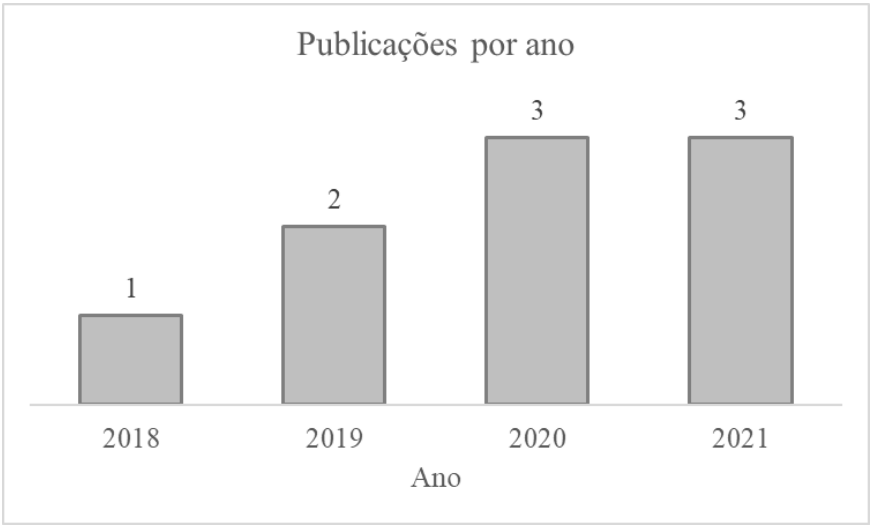

Figura 2.2: Publicações de artigos selecionados para revisão distribuídos anualmente.

Os principais autores e artigos tiveram 1 citação cada. Sendo eles: Kumkar *et al.* (2018b), Sharma *et al.* (2021), Annamoradnejad *et al.* (2019b) e Ahmed *et al.* (2020a).

Após a análise crítica dos trabalhos os principais métodos utilizados para avaliações patrimoniais utilizando técnicas de mineração de dados foram identificados. As formas de coletas, as técnicas de limpeza e transformação dos dados e a utilização de algoritmos de *Machine Learning*, além da avaliação dos algoritmos maiores assertividades.

Embora tenham em comum o objetivo de utilização de técnicas computacionais para avaliação patrimonial, cada um dos trabalhos difere quanto a forma em que realizam a coleta de dados e os tipos de algoritmos de ML utilizados, bem como o objetivo do estudo.

Na sequência são apresentados os resultados para cada fase da descoberta de informações baseadas em dados aplicados nos trabalhados selecionados.

2.5.1. EXTRAÇÃO E LIMPEZA DE DADOS

Referente as técnicas de extração de extração de dados nem todos os artigos selecionados basearam-se em frameworks de *webscraping*. Alguns autores utilizaram dados cedidos por empresas imobiliárias para realização dos estudos. O resumo das técnicas utilizadas e quantidade de dados adquiridos estão resumidos na Tabela 2.2.

Tabela 2.2: Resumo das técnicas de aquisição de dados aplicados nos estudos selecionados.

Artigo	Método de Aquisição	Total de Dados	Local
Kumkar et. al 2018	Webscaping	45 mil	Mumbai, Índia
Annamoradnejad et al., 2019	Webscaping	139 mil	Teerã, Irã
Niu e Niu, 2019	Base de Dados Imobiliária	44 mil	Distrito de Xihu, Hangzhou, China
Ahmed et. al., 2020	Webscaping	Não informado	28 cidades do Paquistão
Berawi et. al., 2020	Webscaping	1,751 mil	Sul de Jacarta, Indonésia
Salem e Mazzara, 2020	Webscaping	Não informado	Amã, Jordânia
Harten et. al., 2021	Webscaping	33 mil	Xangai, China
Sharma et. al., 2021	Webscaping	Não informado	Califórnia, EUA

Todos os autores utilizaram técnicas de limpeza da base de dados baseado na remoção de valores nulos e *outliers*, entretanto, não aprofundaram nas informações referentes a identificação desses valores discrepantes.

Em Annamoradnejad *et al.* (2019b), os autores removeram dados que eram datados do ano de 2017, além da realização da limpeza dos dados conforme citado anteriormente.

A tabela abaixo apresenta o dataset com a quantidade de dados selecionados para realização dos treinamentos de algoritmos.

Tabela 2.3: Quantidade de dados utilizados para treinamento de modelos de valoração de imóveis.

Artigo	**Total de Dados**	**Dataset de Treinamento**	**%**
Kumkar et. al 2018	45.297	16.194	35,8%
Annamoradnejad et al., 2019	139.751	39.257	28,1%
Niu e Niu, 2019	44.113	35.291	80,0%
Ahmed et. al., 2020	Não informado	Não informado	-
Berawi et. al., 2020	1.751	1.237	70,6%
Salem e Mazzara, 2020	Não informado	Não informado	-
Harten et. al., 2021	33.084	3.450	10,4%
Sharma et. al., 2021	Não informado	Não informado	-

2.5.2. TÉCNICAS DE MACHINE LEARNING APLICADA A AVALIAÇÃO PATRIMONIAL

Nos trabalhos selecionados para revisão sistemática os autores selecionaram diversos modelos de aprendizados diferentes, sendo os principais: o modelo hedônico RLM e o algoritmo de ML Random Forest. A distribuição dos modelos utilizados pode ser observada na tabela

Tabela 2.4: Modelos utilizados em avaliações patrimoniais nos trabalhos revisados.

	Modelos							
Artigo	**Random Forest**	**Gradient Boosting**	**XGBoost**	**RNA**	**RLM**	**CART**	**KNN**	**RP**
Kumkar et. al 2018	X	X	X	O	O	O	O	O
Annamoradnejad et al., 2019	O	O	O	O	X	O	O	O
Niu e Niu, 2019	X	O	X	X	X	O	O	O
Ahmed et. al., 2020	O	O	O	O	X	X	O	O
Berawi et. al., 2020	O	O	O	O	O	O	O	O
Salem e Mazzara, 2020	O	O	O	O	X	X	X	O
Harten et. al., 2021	O	O	O	O	X	O	O	O
Sharma et. al., 2021	X	O	O	O	X	X	O	X

Sendo:

X: Selecionado

O: Não selecionado

XGBoost: *Extreme Gradient Boosting*

RNA: Redes Neurais Artificiais

RLM: Regressão Linear Múltipla

CART: *Classification and Regression Tree*

KNN: *K-Nearest Neighbors*

RP: Regressão Polinomial.

2.5.3. MÉTODOS DE AVALIAÇÃO DOS MODELOS

Em Kumkar *et al.* (2018a), os autores avaliaram a performance dos modelos através da Porcentagem da Média do Erro Absoluto (*Mean Absolute Percentage Error – MAPE*).

O trabalho apresentado por Niu e Niu (2019) utilizaram como medidas de avaliação de performance dos modelos a raiz quadrada do erro médio quadrático (Root Mean Square Error – RMSE) e a Porcentagem da Média do Erro Absoluto (*Mean Absolute Percentage Error – MAPE*).

Em Ahmed *et al.* (2020b), os autores utilizaram RLM não para avaliações de imóveis, mas para prever o crescimento médio do preço de imóveis.

No trabalho Annamoradnejad *et al.* (2019a), os autores utilizaram o estudo para avaliar o impacto de cada variável utilizada na RLM, determinando quais as variáveis mais impactantes.

No trabalho de Salem e Mazzara (2020) foi utilizada a técnica de *train-test split* para avaliar a acurácia dos modelos de ML.

Os autores de Harten, Kim e Brazier (2021) utilizaram a modelagem hedônica e averiguaram a sua eficiência baseado no coeficiente de determinação (R^2).

2.6. DISCUSSÃO

Nessa seção serão discutidos as metodologias, resultados e conclusões dos artigos selecionados que se relacionam com o escopo desta pesquisa.

Na extração dos dados, os autores de Kumkar *et al.* (2018a) utilizaram técnicas de coleta de dados na web, como por exemplo *Web scraping*, mas não detalharam este processo. O *dataset* obtido possui 45.297 dados de imóveis residenciais de cidade de Mumbai, na Índia, contendo 13 variáveis: Valor de venda, área útil, área construída, área total, localização, quantidade de quartos, quantidade de banheiros, quantidade de varandas, quantidade de vagas de garagem, número do andar, tipo de mobília (com ou sem mobília), tipo de piso (cerâmico, madeira, granito etc.) e o tipo da propriedade (apartamento, casa, casa em condomínio, dentre outros).

Em Niu e Niu (2019), os autores criaram uma plataforma web, baseado no framework Django, para que usuários pudessem inserir os dados dos imóveis. Para o estudo foram utilizados dados cedidos por uma empresa imobiliária contendo 44113 registros de anúncios de imóveis no distrito de Xihu, em Hangzhou, na China. O dataset contém informações do tipo do imóvel, localização, andar, área útil, área construída, valor, condição do imóvel, latitude, longitude, dentre outras variáveis utilizadas.

Diferentemente do trabalho anterior, Ahmed *et al.* (2020b) e Annamoradnejad *et al.* (2019a) utilizaram técnicas de mineração de dados da web. Ambos construíram robôs do tipo web crawlers para coletar dados de sites de ofertas de imóveis.

Ahmed *et al.* (2020b) utilizou o HTML Agility-pack que se trata de uma biblioteca escrita em C# que automatiza a verificação de código HTML capaz de coletar dados dessa fonte, como título, subseções, dentre outros. Os autores utilizaram a ferramenta para extrair dados de imóveis de websites de ofertas de imóveis do Paquistão.

Já Annamoradnejad *et al.* (2019a), utilizaram a mesma técnica de web scraping para coletar dados de anúncios de imóveis em websites da cidade de Teerã, no Irã. O *dataset* criado contém dados como: valor, idade do imóvel, andar, área, distância até o centro da cidade, dentre outros, de aproximadamente 139 mil imóveis.

Referente a limpeza e transformação dos dados os artigos não foram precisos nas metodologias utilizadas, entretanto em Kumkar *et al.* (2018a), os autores informaram que as observações que possuíam dados faltando foram removidas, bem como os outliers. Já Niu e Niu (2019), os autores utilizaram redes neurais para identificar e remover dados de imóveis em duplicidade através de algoritmo de similaridade semântico e criou uma camada intermediaria entre a entrada de dados e o modulo de avaliação patrimonial. A camada intermediaria analisa as variáveis de entrada, estabelece uma hierarquia dos fatores que influenciam no sistema de avaliação e normaliza os valores das variáveis de acordo com seu peso.

Em Kumkar *et al.* (2018a), os autores realizaram um comparativo entre várias técnicas de *Machine Learning* para realizar avaliações de imóveis. A averiguação da eficácia de cada modelo foi estimada pela Porcentagem Absoluta do Erro Médio (Mean Absolute Percentage Error – MAPE). Os autores encontraram que o modelo XBoost obteve o menor erro percentual, com 13,63%, seguido do modelo Gradient Boosting (GBDT), Random Forest (RF) e Bagging, com 14,00%, 14,96% e 15,36% respectivamente.

No estudo desenvolvido por Niu e Niu (2019), os autores também fizeram comparativo de modelos de *Machine Learning*. Os resultados obtidos apontaram que as *Redes Neurais Artificiais* tiveram uma maior precisão nas avaliações, seguido do modelo *Random Forest* e *Gradient Boosting*. Ainda nesse estudo, os autores utilizaram modelagem hedônica com os mesmos dados dos outros algoritmos para verificar a precisão da Regressão Linear, o resultado demonstrou que essa modelagem tem a menor precisão. Por fim, os autores realizaram um

método Ensemble utilizando os algoritmos de *Machine Leaning* testados anteriormente. A ideia é maximizar as vantagens e minimizar as desvantagens que cada modelo de *Machine Learning* possui. O resultado baseou-se na média ponderada das avaliações resultantes de cada algoritmo de *Machine Leaning* testado. As avaliações de imóveis obtidas pelo método Ensemble associando RNA + GBDT + RF foram as mais precisas dentre todos os modelos testados.

Em Ahmed *et al.* (2020b), os autores utilizaram o *dataset* de ofertas de imóveis obtido através de *web scraping* em sites de anúncios para desenvolver um sistema capaz de analisar se um imóvel terá valorização nos próximos 5 anos. Primeiramente os autores separaram os dados dos imóveis por região e obtiveram as médias dos valores dos imóveis para os anos de 2014 à 2015, 2016 à 2017 e 2018 à 2019. Posteriormente, utilizando a Regressão Linear, estimaram os valores médios para os anos de 2020 à 2021, 2022 à 2023 e 2024 à 2025, obtendo a estimativa média de crescimento de valorização das regiões analisadas.

A partir dos cálculos realizados, os autores categorizaram os dados de valor atual e porcentagem média de crescimento em baixo, médio e alto. Com os dados categorizados e classificados os autores utilizaram o algoritmo de arvore de decisão J4.8 desenvolvendo um sistema de suporte a decisão para aquisição de imóveis de acordo com as regiões do Paquistão.

Annamoradnejad *et al.* (2019a) utilizaram quase 140 mil dados de ofertas de imóveis obtidos através de Web scraping de sites de anúncios de imóveis dos distritos da cidade de Teerã para averiguar quais variáveis são influenciadoras nos valores das propriedades.

As variáveis foram divididas em 3 categorias: (i) estrutural: Área; idade da construção e andar; (ii) localização: distância até o principal distrito de negócios (CBD); valor da terra; densidade do distrito; (iii) ambiental: qualidade do ar; área verde per capita;

Os resultados apresentados pelos autores foram que a área do apartamento é o fator mais determinante na construção do valor de imóveis com uma correlação positiva de 0,89. O andar onde está localizado o imóvel também é um fator inversamente determinante, possuindo uma correlação negativa de -0,68.

(Salem e Mazzara (2020) criaram um *bot* capaz de predizer valores de imóveis através do aplicativo Telegram. Os autores realizaram um comparativo entre 04 modelos de aprendizado: RLM, *CART, KNN-R e KNN-C*. A eficácia de cada modelo foi estimada pela acurácia. Os

autores observaram que o modelo *K-Neighbors Regressor* obteve a melhor métrica com 0.50, seguido por RLM, *CART* e *K-Neighbors Classifier*, com 0.34, 0.14 e 0.10, respectivamente.

2.7. CONCLUSÃO

Baseado no grupo de estudos obtidos a partir da análise da produção científica utilizando todos os conceitos, observa-se que é um tema muito atual, pois os artigos encontrados começaram a serem publicados a partir do final de 2018.

Ademais, somente 08 estudos foram encontrados, permitindo inferir que se trata de uma área cuja técnicas estatísticas e computacionais para aprimorar os resultados das avaliações patrimoniais vem sendo exploradas sem buscar uma solução para um ponto primordial, a obtenção de dados de forma automática, conforme visto nos trabalhos de (KUMKAR et al., 2018a) e (NIU; NIU, 2019a).

A obtenção de dados de fontes difusas é importante para criação de um Big Data em quantidade suficiente, conforme observado nos estudos de (AHMED et al., 2020b) e (ANNAMORADNEJAD et al., 2019a).

Entretanto, nesses 02 últimos trabalhos citados, os autores tiveram foco em outros resultados, não explorando a possibilidade de avaliar os resultados de algoritmos de *Machine Learning* com uma grande quantidade de dados provenientes de sites de anúncios de ofertas de imóveis, o que permitiria uma visão holística do mercado, evitando enviesamento e aprimorando a precisão dos resultados.

Por fim, observa-se que em todos os trabalhos nos quais os algoritmos de *Machine Learning* foram aplicados para realização da avaliação patrimonial, esses tiveram uma performance muito superior aos modelos hedônicos.

2.8. REFERÊNCIAS BIBLIOGRÁFICAS

ABIDOYE, R. B.; CHAN, A. P. C. Artificial neural network in property valuation: application framework and research trend. **Property Management**, 16 out. 2017.

ABIDOYE, R. B.; CHAN, A. P. C. Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network. **Pacific Rim Property Research Journal**, v. 24, n. 1, p. 71–83, 2 jan. 2018a.

ABIDOYE, R. B.; CHAN, A. P. C. Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network. **Pacific Rim Property Research Journal**, v. 24, n. 1, p. 71–83, 2 jan. 2018b.

ABIDOYE, R. B.; CHAN, A. P. C. Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network. **Pacific Rim Property Research Journal**, v. 24, n. 1, p. 71–83, 2 jan. 2018c.

ABNT. **ABNT NBR 14653-2 - Avaliação de bens Parte 2: Imóveis Urbanos**. [s.l.] ABNT, 2011.

ABUNAHMAN, S. A. **Curso básico de engenharia legal e de avaliações**. 4. ed. São Paulo: PINI, 2008.

ACEBIP. **Associação Brasileira das Entidades de Crédito Imobiliário e Poupança**. Disponível em: <https://www.abecip.org.br/>. Acesso em: 27 mar. 2020.

AHMED, H. et al. Producing Standard Rules for Smart Real Estate Property Buying Decisions based on Web Scraping Technology and Machine Learning Techniques. **International Journal of Advanced Computer Science and Applications (IJACSA)**, v. 11, n. 3, 40/30 2020a.

AHMED, H. et al. Producing standard rules for smart real estate property buying decisions based on web scraping technology and machine learning techniques. **International Journal of Advanced Computer Science and Applications**, v. 11, n. 3, p. 498–505, 2020b.

AHN, T.; CHARNES, A.; COOPER, W. W. Some statistical and DEA evaluations of relative efficiencies of public and private institutions of higher learning. **Socio-Economic Planning Sciences**, v. 22, n. 6, p. 259–269, jan. 1988.

ANNAMORADNEJAD, R. et al. **Using Web Mining in the Analysis of Housing Prices: A Case study of Tehran**. 2019 5th International Conference on Web Research, ICWR 2019. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2019a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069907225&doi=10.1109%2fICWR.2019.8765250&partnerID=40&md5=6faec40df83f2429b4c6357594c31367>

ANNAMORADNEJAD, R. et al. **Using Web Mining in the Analysis of Housing Prices: A Case study of Tehran**. 2019 5th International Conference on Web Research (ICWR). **Anais**... In: 2019 5TH INTERNATIONAL CONFERENCE ON WEB RESEARCH (ICWR). abr. 2019b.

BACEN. **Banco Central do Brasil**. Disponível em: <https://www.bcb.gov.br/estatisticas/mercadoimobiliario>. Acesso em: 27 mar. 2020.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1 ago. 1996.

BREIMAN, L. et al. **Classification And Regression Trees**. Boca Raton: Routledge, 2017.

BUITRAGO-SUESCÚN, O. Y. et al. Data Envelopment Analysis for Efficiency Measurement on Higher Education Institutions: a State of the Art Review. **Revista Científica General José María Córdova**, v. 15, n. 19, p. 147–173, jun. 2017.

BUREAU, U. C. **Census.gov**. Disponível em: <https://www.census.gov/en.html>. Acesso em: 29 maio. 2021.

BUSSAB, W. DE O.; MORETTIN, P. A. **Estatística Básica**. [s.l.] SARAIVA EDITORA, 2017.

CASADO, F. L. ANÁLISE ENVOLTÓRIA DE DADOS: CONCEITOS, METODOLOGIA E ESTUDO DA ARTE NA EDUCAÇÃO SUPERIOR. v. 20, n. 01, p. 13, 2007.

CBIC. **Indicadores Imobiliários Nacionais**. Disponível em: <https://cbic.org.br/estudos/>. Acesso em: 13 dez. 2020.

ČEH, M. et al. Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. **ISPRS International Journal of Geo-Information**, v. 7, n. 5, p. 168, maio 2018.

CHARNES, A.; COOPER, W. W.; RHODES, E. Measuring the efficiency of decision making units. **European Journal of Operational Research**, v. 2, n. 6, p. 429–444, nov. 1978.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 785–794, 13 ago. 2016.

CREA. **CREA – Canadian Real Estate Association**, 2021. Disponível em: <https://www.crea.ca/>. Acesso em: 29 maio. 2021

CROSBY, N.; LAVERS, A.; MURDOCH, J. Property valuation variation and the "margin of error" in the UK. **Journal of Property Research**, v. 15, n. 4, p. 305–330, 1 jan. 1998.

DANTAS, R. A.; MAGALHÃES, A. M.; VERGOLINO, J. R. DE O. Avaliação de imóveis: a importância dos vizinhos no caso de Recife. **Economia Aplicada**, v. 11, n. 2, p. 231–251, jun. 2007.

DANTAS, RUBENS ALVES, R. **UMA NOVA METODOLOGIA PARA AVALIAÇÃO DE IMÓVEIS UTILIZANDO REGRESSÃO ESPACIAL**. . In: XXI COBREAP. ESPIRITO SANTO: 2001.

DE'ATH, G.; FABRICIUS, K. E. Classification and regression trees: a powerful yet simple technique for ecological data analysis. **Ecology**, v. 81, n. 11, p. 3178–3192, 1 nov. 2000.

DO, A. Q.; GRUDNITSKI, G. A neural network approach to residential property appraisal. **The Real Estate Appraiser**, v. 58, p. 38–45, 1 jan. 1992.

DRUCKER, H. Improving Regressors Using Boosting Techniques. **Proceedings of the 14th International Conference on Machine Learning**, 17 ago. 1997.

FARIAS, A. M. L. DE; LAURENCEL, L. DA C. **Estátistica Descritiva**. UFF: [s.n.].

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–37, 15 mar. 1996.

FEURER, M. et al. Auto-sklearn: Efficient and Robust Automated Machine Learning. In: HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. (Eds.). . **Automated Machine Learning**. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2019a. p. 113–134.

FEURER, M. et al. Efficient and Robust Automated Machine Learning. p. 9, 2019b.

GONZÁLEZ, M. A. S. **Aplicação de Técnicas de Descobrimento de Conhecimento em Bases de Dados e de Inteligência Artificial em Avaliação de Imóveis**. Rio Grande do Sul: UFRGS, dez. 2002.

GOVERNMENT OF CANADA, S. C. **The Daily — New Housing Price Index, April 2021**. Disponível em: <https://www150.statcan.gc.ca/n1/daily-quotidien/210520/dq210520d-eng.htm>. Acesso em: 29 maio. 2021.

GROVER, R. Mass valuations. **Journal of Property Investment & Finance**, 7 mar. 2016.

HALLAK, R.; PEREIRA FILHO, A. J. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS: sensibilidade a variações com os esquemas de advecção e assimilação de dados. **Revista Brasileira de Meteorologia**, v. 26, p. 591–608, dez. 2011.

HARTEN, J. G.; KIM, A. M.; BRAZIER, J. C. Real and fake data in Shanghai's informal rental housing market: Groundtruthing data scraped from the internet. **Urban Studies**, v. 58, n. 9, p. 1831–1845, 1 jul. 2021.

HO, T. K. The random subspace method for constructing decision forests. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 20, n. 8, p. 832–844, ago. 1998.

IBGE. **Pesquisa Anual da Indústria da Construção - PAIC | IBGE**. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/industria/9018-pesquisa-anual-da-industria-da-construcao.html?=&t=destaques>. Acesso em: 27 mar. 2020.

JIANG, H.; JIN, X.-H.; LIU, C. The effects of the late 2000s global financial crisis on Australia's construction demand. **Construction Economics and Building**, v. 13, n. 3, p. 65–79, 18 set. 2013.

JOVIC, A.; BRKIC, K.; BOGUNOVIC, N. **An overview of free software tools for general data mining**. 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). **Anais**... In: 2014 37TH INTERNATIONAL CONVENTION ON INFORMATION AND COMMUNICATION TECHNOLOGY, ELECTRONICS AND MICROELECTRONICS (MIPRO). maio 2014.

KLAMER, P.; BAKKER, C.; GRUIS, V. Research bias in judgement bias studies – a systematic review of valuation judgement literature. **Journal of Property Research**, v. 34, n. 4, p. 285–304, 2 out. 2017.

KUMKAR, P. et al. **Comparison of Ensemble Methods for Real Estate Appraisal**. Proceedings of the 3rd International Conference on Inventive Computation Technologies, ICICT 2018. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2018a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082649654&doi=10.1109%2fICICT43934.2018.9034449&partnerID=40&md5=ec62aa51060a948ee7a4f759733ffb59>

KUMKAR, P. et al. **Comparison of Ensemble Methods for Real Estate Appraisal**. 2018 3rd International Conference on Inventive Computation Technologies (ICICT). **Anais**... In: 2018 3RD INTERNATIONAL CONFERENCE ON INVENTIVE COMPUTATION TECHNOLOGIES (ICICT). nov. 2018b.

LANCASTER, K. J. A New Approach to Consumer Theory. **Journal of Political Economy**, v. 74, n. 2, p. 132–157, 1966.

LASOTA, T. et al. **Comparison of Ensemble Approaches: Mixture of Experts and AdaBoost for a Regression Problem**. (N. T. Nguyen et al., Eds.)Intelligent Information and Database Systems. **Anais**...: Lecture Notes in Computer Science.Cham: Springer International Publishing, 2014.

LINS, M. P. E.; NOVAES, L. F. DE L.; LEGEY, L. F. L. Real Estate Appraisal: A Double Perspective Data Envelopment Analysis Approach. **Annals of Operations Research**, v. 138, n. 1, p. 79–96, 1 set. 2005.

MCCLUSKEY, W. J. et al. Prediction accuracy in mass appraisal: a comparison of modern approaches. **Journal of Property Research**, v. 30, n. 4, p. 239–265, 1 dez. 2013.

MELANDA, E.; HUNTER, A.; BARRY, M. Identification of locational influence on real property values using data mining methods. **Cybergeo : European Journal of Geography**, 4 fev. 2016.

MING-SYAN CHEN; JIAWEI HAN; YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 866–883, dez. 1996.

MOHER, D. et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. **PLOS Medicine**, v. 6, n. 7, p. e1000097, 21 jul. 2009.

MTE. **CAGED**. Disponível em: <http://pdet.mte.gov.br/caged>. Acesso em: 27 mar. 2020.

NAKAGAWA, S.; SCHIELZETH, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. **Methods in Ecology and Evolution**, v. 4, n. 2, p. 133–142, 1 fev. 2013.

NEDER, H. D. et al. Índice de defasagem do imposto predial e territorial urbano (IPTU) dos Municípios de Minas Gerais: um estudo de caso para Uberlândia (MG). **Revista ESPACIOS**, v. 38, n. 46, 6 out. 2017.

NIU, J.; NIU, P. **An intelligent automatic valuation system for real estate based on machine learning**. (T. J.M.R.S, Ed.)ACM International Conference Proceeding Series. **Anais**...Association for Computing Machinery, 2019a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077820799&doi=10.1145%2f3371425.3371454&partnerID=40&md5=25936a4abf77f70deb07c0299ab0c874>

NIU, J.; NIU, P. **An intelligent automatic valuation system for real estate based on machine learning**. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing. **Anais**...: AIIPCC '19.New York, NY, USA: Association for Computing Machinery, 19 dez. 2019b. Disponível em: <http://doi.org/10.1145/3371425.3371454>. Acesso em: 14 dez. 2020

NUNES, D. B. et al. Modelo de regressão linear múltipla para avaliação do valor de mercado de apartamentos residenciais em Fortaleza, CE. **Ambiente Construído**, v. 19, n. 1, p. 89–104, mar. 2019.

PATEL, J. M. Web Scraping in Python Using Beautiful Soup Library. In: PATEL, J. M. (Ed.). . **Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale**. Berkeley, CA: Apress, 2020. p. 31–84.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825–2830, 2011.

PELLI, A. N. **Redes neurais artificiais aplicadas às avaliações em massa estudo de caso para a cidade de Belo Horizonte / MG**. [s.l.] Universidade Federal de Minas Gerais, 10 mar. 2006.

PETERSON, S.; FLANAGAN, A. Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. **Journal of Real Estate Research**, American Real Estate Society. v. 31(2), p. 147–164, 2009.

RICS. **The future of valuations**. Disponível em: <https://www.rics.org/globalassets/rics-website/media/knowledge/research/insights/future-of-valuations-insights-paper-rics.pdf>. Acesso em: 5 jun. 2021.

ROSEN, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. **Journal of Political Economy**, v. 82, n. 1, p. 34–55, jan. 1974.

ROYAL INSTITUTION OF CHARTERED SURVEYORS (ED.). **RICS valuation - Global standards: incorporating the IVSC International Valuation Standards**. London: RICS, 2017.

SALEM, H.; MAZZARA, M. ML-based Telegram bot for real estate price prediction. **Journal of Physics: Conference Series**, v. 1694, n. 1, p. 012010, 1 dez. 2020.

SENI, G.; ELDER, J. F. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. **Synthesis Lectures on Data Mining and Knowledge Discovery**, v. 2, n. 1, p. 1–126, 1 jan. 2010.

SHARMA, N. et al. Real Estate Price's Forecasting Through Predictive Modelling. In: JOSHI, A.; KHOSRAVY, M.; GUPTA, N. (Eds.). . **Machine Learning for Predictive Analysis**. Lecture Notes in Networks and Systems. Singapore: Springer Singapore, 2021. v. 141p. 589–597.

STEINER, M. T. A. et al. Métodos estatísticos multivariados aplicados à engenharia de avaliações. **Gestão & Produção**, v. 15, n. 1, p. 23–32, abr. 2008.

VALIER, A. Who performs better? AVMs vs hedonic models. **Journal of Property Investment & Finance**, 27 mar. 2020.

WANG, F. et al. **House Price Prediction Approach based on Deep Learning and ARIMA Model**. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). **Anais**... In: 2019 IEEE 7TH INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND NETWORK TECHNOLOGY (ICCSNT). out. 2019.

WEINSTOCK, L. R. Introduction to U.S. Economy: Housing Market. p. 3, 2021.

WINARNO, E.; HADIKURNIAWATI, W.; ROSSO, R. N. **Location based service for presence system using haversine method**. 2017 International Conference on Innovative and Creative Information Technology (ICITech). **Anais**... In: 2017 INTERNATIONAL CONFERENCE ON INNOVATIVE AND CREATIVE INFORMATION TECHNOLOGY (ICITECH). nov. 2017.

WU, X. et al. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p. 1–37, 1 jan. 2008.

ZHOU, G. et al. Artificial Neural Networks and the Mass Appraisal of Real Estate. **International Journal of Online and Biomedical Engineering (iJOE)**, v. 14, n. 03, p. 180–187, 30 mar. 2018.

ZURADA, J.; LEVITAN, A. S.; GUAN, J. A comparison of regression and artificial intelligence methods in a mass appraisal context. **Journal of Real Estate Research**, v. 33, n. 3, p. 349–387, 2011.

3. ARTIGO 02: AVALIAÇÃO DE IMÓVEIS URBANOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS: ESTUDO DE CASO

3.1. RESUMO

O mercado imobiliário tem dimensões econômicas importantes no PIB dos países. A avaliação imobiliária é necessária para várias atividades, incluindo análise de investimento, tributação e indenizações. Junto com os compradores e vendedores em potencial, estimar os preços dos imóveis também pode beneficiar os investidores. Neste trabalho, foi desenvolvido um framework para avaliação patrimonial de forma automatizada, utilizados métodos de mineração na web para gerar um conjunto de dados limpo e organizado, a partir de um popular site de anúncios de imóveis, e utilizando quatro algoritmos de aprendizado de máquina: *RF*, *XGBoost*, *AdaBoost* e *CART*, além do modelo hedônico, *RLM* para seleção daquele com melhor performance. A comparação é baseada na estimativa dos preços dos imóveis em Botafogo, Rio de Janeiro/BR. O uso de técnicas de Mineração de Dados para estimar os preços dos imóveis mostraram superioridade perante o modelo hedônico e o modelo proposto foi capaz de coletar os dados diretamente da web, remover pontos discrepantes, treinar e identificar o modelo com maior acurácia. O algoritmo RF foi superior aos demais modelos testados, tendo uma acurácia maior que a RLM em quase 5%, sem a utilização de técnicas de *tuning*.

Palavras-chave: Avaliação; Imóvel; Patrimônio; Mineração de Dados; Aprendizado de Máquina

3.2. ABSTRACT

The real estate market has important economic dimensions in countries' GDP. Real estate valuation is required for a variety of activities, including investment analysis, taxation and court indemnity. Along with potential buyers and sellers, estimating real estate prices can also benefit investors. Many researches have been seeking to improve real estate valuation techniques using ML, but they are based on a relatively small set of data, mostly collected with the help of real estate companies. In this work, web mining techniques were used to generate an organized dataset from a popular real estate advertisement website and make a comparative analysis of four ML algorithms: RF, XGBoost, AdaBoost and CART, in addition to hedonic model, for real estate valuation. The comparison is based on estimated property prices in Botafogo' district, Rio de Janeiro/BR. The Data Mining techniques applied to estimate real estate prices showed superiority over the hedonic model. The framework proposed was able to collect data

directly from the real estate advertising websites, remove outliers, train and identify the best accuracy model. The RF algorithm was superior to the other models tested, having an accuracy greater than the RLM by almost 5%, without the use of tuning techniques.

Keywords: Real Estate; Appraisal; Valuation; Property; Data Mining; Machine Learning

3.3. INTRODUÇÃO

O mercado imobiliário tem dimensões econômicas importantes no Produto Interno Bruto (PIB) dos países em função do crescimento populacional e do aumento do desenvolvimento urbano.

No relatório publicado em maio de 2021, pelo *Congressional Research Service* (CRS) dos Estados Unidos, no ano anterior os gastos com investimentos residenciais fixos eram de cerca de US $ 885 bilhões, representando cerca de 4,2% do PIB. Se considerar os gastos com serviços de habitação, o que inclui os aluguéis dos locatários e serviços públicos e o aluguel imputado dos proprietários e pagamentos de serviços públicos, em conjunto, os gastos com o mercado imobiliário representaram 17,5% do PIB em 2020 (WEINSTOCK, 2021).

Já no Brasil, o mercado imobiliário e da construção civil foi responsável por 4,36% do PIB brasileiro no ano de 2017, bem próximo ao percentual do mesmo setor no mercado americano (IBGE, 2017).

No relatório apresentado em maio de 2021, pelo *U.S. Census Bureau e U.S. Department of Housing and Urban Development* (BUREAU, 2021), cerca de 863 mil novas unidades residenciais foram vendidas em abril/2021, correspondendo a um crescimento de 48,3%.

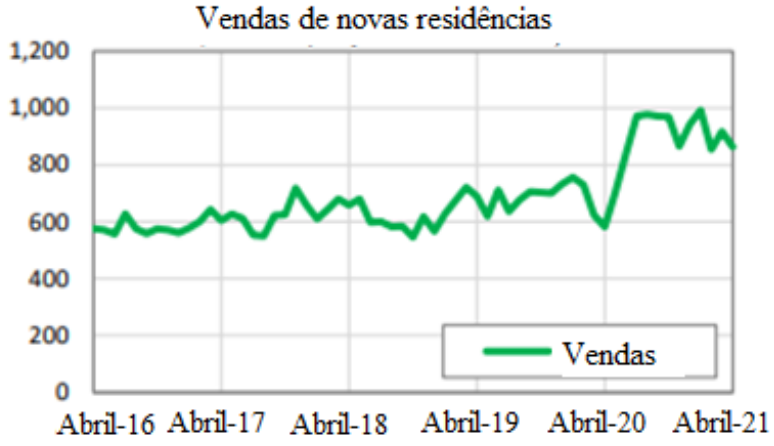

Figura 3.1: Crescimento do mercado imobiliário americano. Fonte: (BUREAU, 2021)

O *Canadian Housing Statistics Program* (CHSP), através do *National Statistic Office,* divulgou um relatório informando que a venda de unidades residenciais cresceu 256% em abril de 2021, comparado ao mesmo período do ano anterior. A estimativa é de que mais de 700 mil unidades residenciais sejam vendidas no mesmo ano, conforme apresentado na Figura 3.2 (GOVERNMENT OF CANADA, 2021).

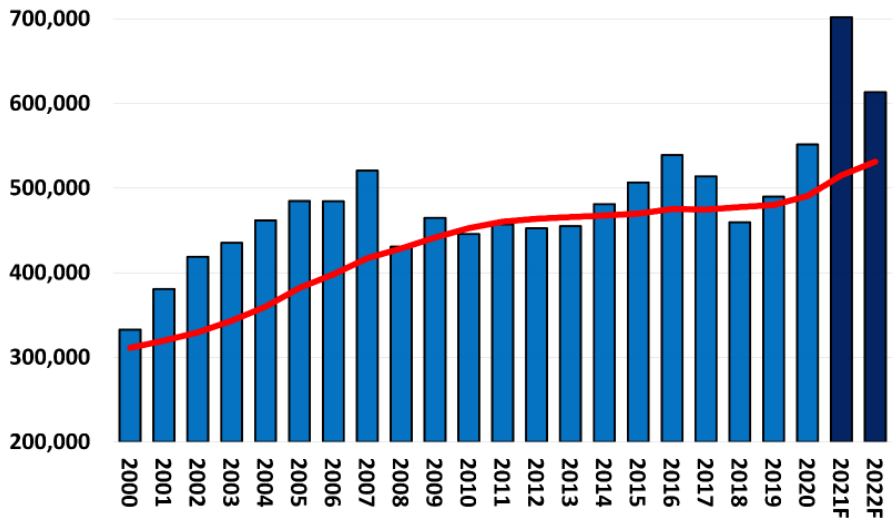

Figura 3.2: Expectativa de comercialização de unidade residenciais no Canadá. Fonte: (CREA, 2021)

No Brasil as vendas de unidades residenciais também tiveram aumento considerável. O primeiro e terceiro trimestres de 2020 superaram em 27,7% e 27,3%, respectivamente os mesmos períodos do ano anterior (CBIC, 2020).

Em todas as regiões do Brasil observa-se aumento de vendas no mercado imobiliário, conforme apresentado na Figura 2.

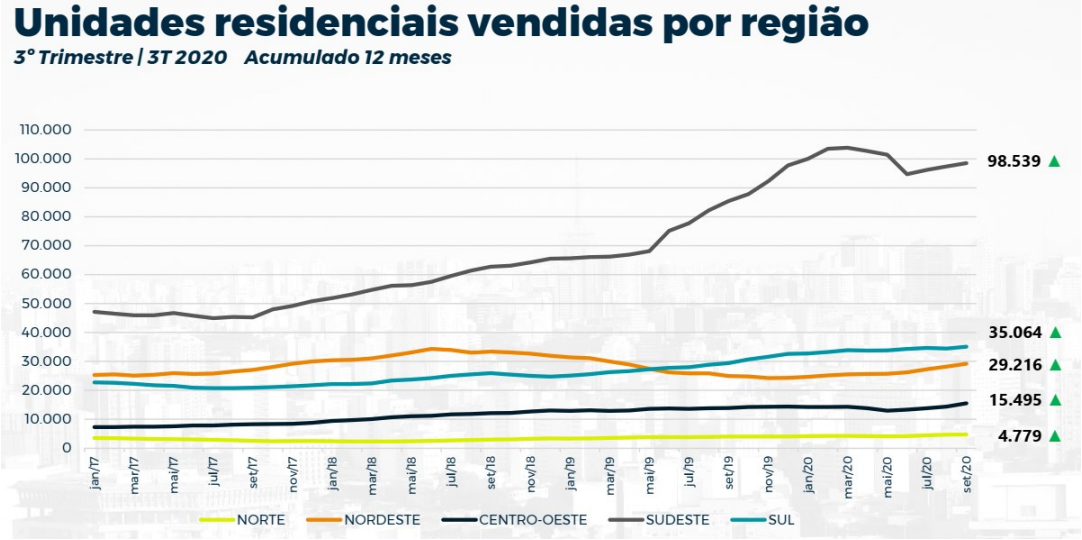

Figura 3.3: Unidades residenciais vendidas por região do Brasil – Fonte: (CBIC, 2020)

A construção civil é uma indústria muito importante em todo o mundo, não só pelo seu papel no mercado financeiro, mas também socioeconômico, devido ao alto número de pessoas empregadas por esse setor.

Antes da bolha imobiliária em 2006, o setor da construção civil empregava mais de 1 milhão de pessoas nos Estados Unidos. Entretanto, como resultado do estouro da bolha e da recessão, o número de empregados reduziu em quase 50%. Com a retomada do setor o número de empregados voltou a crescer, chegando à 872 mil funcionários empregados em março de 2021, segundo o *Bureau of Labor Statistics* (WEINSTOCK, 2021).

De acordo com os dados apresentados pelo Ministério do Trabalho, em 2019 foram empregados mais de 71 mil novos funcionários no setor da construção civil no Brasil (MTE, 2020). No censo realizado em 2017, pelo IBGE, cerca de 2 milhões de funcionários formais estão empregados no setor (IBGE, 2017).

O preço da habitação é um reflexo importante da economia e um indicador importante para o desenvolvimento saudável e estável do setor imobiliário. A avaliação do preço da habitação desempenha um papel significativo na formação da economia devido a importância do setor para a economia global e dos países tendo um papel fundamental nas decisões estratégicas relacionado aos investimentos imobiliários.

Uma avaliação patrimonial acurada permite os governos dos países regularem melhor o mercado imobiliário e manter seu desenvolvimento estável, saudável e ordenado, permitindo

os incorporadores imobiliários a tomar decisões de investimento com antecedência, além de ser uma forma de investimento, a moradia é um abrigo para atender às necessidades fundamentais das pessoas (WANG et al., 2019).

A inexatidão das avaliações de imóveis pode causar efeitos adversos nos investimentos dos stakeholders do setor imobiliário, o que pode afetar a economia de um país, como por exemplo, a crise financeira global de 2007 (JIANG; JIN; LIU, 2013).

Os bancos têm sido usuários importantes de avaliações imobiliárias para garantias de hipotecas. A precisão das avaliações feitas por avaliadores individuais, avaliação dos riscos atuais apresentados por uma carteira de hipotecas, identificação de fraude são requisitos para revisões periódicas de garantias (GROVER, 2016).

Os esforços de avaliação imobiliária introduzidos nas áreas de atividade de várias instituições públicas e privadas em todo o mundo e devem ser realizados através do emprego de métodos imparciais, objetivos e científicos com o propósito de determinar o estado real, direitos e obrigações dos valores de propriedade (ABIDOYE; CHAN, 2018b).

O modelo de precificação hedônico é um método de avaliação avançado que tem sido amplamente utilizado em todo o mundo. Este modelo baseia-se na teoria da demanda do consumidor criada por Lancaster (LANCASTER, 1966).

Sherwin Rosen aplicou a utilidade hedônica na precificação de um bem. Os consumidores tomam suas decisões de compra com base no número de boas características, bem como no custo por unidade de cada característica, na qual, o preço total de um produto pode ser considerado como uma soma do preço de cada um de seus atributos homogêneos, onde cada variável possui um preço implícito único, considerando o equilíbrio de mercado. Isso implica que o preço de um produto pode ser regredido com base nas suas características quem contribuem exclusivamente para o preço unitário composto geral (ROSEN, 1974).

Pesquisadores tem buscado soluções para aprimorar a precisão das avaliações imobiliárias, entretanto, a maioria dos trabalhos tem se baseado em conjuntos de dados relativamente pequenos (ANNAMORADNEJAD et al., 2019b).

A avaliação de imóveis baseada em modelagem hedônica, apesar de sua simplicidade e objetividade na abordagem, pode falhar na tentativa de capturar efetivamente a relação não linear que existe entre valores de propriedade e atributos de propriedade, é subjetivo por

natureza, impreciso e marcado com erros de especificação de forma funcional, entre outras deficiências (ABIDOYE; CHAN, 2018c).

Problemas metodológicos associados a RLM são conhecidos há algum tempo e incluem não linearidade, multicolinearidade, má especificação da forma de função e heterocedasticidade (PETERSON; FLANAGAN, 2009).

Ademais, a eficiência dos métodos hedônicos está diretamente atrelada aos dados utilizados para construção dos modelos de predição.

Conforme (GROVER, 2016), grande parte da literatura se preocupa em como melhorar a modelagem estatística dos preços de mercado, porém existem questões significativas relacionadas com o tipo e a qualidade dos dados usados nos modelos de avaliação e os requisitos para o uso bem-sucedido das avaliações.

A limitação da quantidade e da qualidade de dados amostrais sempre foi um dos principais problemas enfrentados na avaliação patrimonial, devido à dificuldade de coletar dados em grande quantidade por sua dispersão em sistemas de empresas privadas, bem como em páginas na web.

No Brasil não existem órgãos reguladores responsáveis por manter uma base de dados sólida e correta dos dados de vendas de imóveis. Portanto, as amostras coletadas nas avaliações patrimoniais são advindas, em sua grande maioria, de sites de anúncios de ofertas de vendas ou locação de imóveis. Esses dados encontram-se dispersos, sem consistência, possuindo erros ou omissão de informações importantes devido a participação humana nesse processo.

Conforme relatado no estudo de (ABIDOYE; CHAN, 2018c):

> “[...] se as pré-condições para modelagem de valor de propriedade (banco de dados robusto e de qualidade, treinamento adequado de avaliadores e mercado de propriedade transparente, entre outros) estiverem em vigor, a imprecisão da avaliação de propriedade pode ser reduzida ao mínimo no domínio da avaliação imobiliária.”

Com o aumento das bases de dados de ofertas de imóveis e do poder computacional, as técnicas de Mineração de Dados tornam-se interessantes, possibilitando ao avaliador ter uma visão mais holística do mercado imobiliário, permitindo-o chegar a conclusões mais assertivas.

Neste artigo será proposto um sistema inteligente e automatizado para realização de avaliações patrimoniais, desde a coleta dos dados até identificar melhores algoritmos de *ML* capazes de predizer valores de imóveis com maior acurácia.

O trabalho irá se basear nas técnicas de descoberta de conhecimento em base de dados (KDD) apresentadas por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Na etapa de ETL serão utilizados o webscraping, para coletar de dados dispersos em páginas da internet, a limpeza será baseada em eliminação de dados nulos e discrepantes, para transformação será utilizado o *StandardScaler,* que se baseia no escore padronizado, com a criação do Big Data serão utilizados os algoritmos de Machine Learning: CART, Random Forest, Extreme Gradient Boost e Adaptive Boosting para realização de avaliações de imóveis. O estudo de caso adotará o bairro de Botafogo, na cidade do Rio de Janeiro, Brasil. Além dos algoritmos citados, será utilizada também a modelagem hedônica através da Regressão Linear Múltipla.

O objetivo do presente trabalho é desenvolver um framework para avaliação patrimonial utilizando webscraping para coleta de dados em sites de anúncios de imóveis, limpeza e tratamento dos dados e treinamento de diferentes modelos de Machine Learning do tipo bagging e boosting para aprimorar a precisão das avaliações patrimoniais.

3.4. REFERENCIAL TEÓRICO

3.4.1. VALOR DE MERCADO, PREÇO E CUSTO

Avaliações patrimoniais precisas possuem grande importância para diversos setores da economia (DO; GRUDNITSKI, 1992).

Para Dantas (2001), a engenharia de avaliações se torna uma ciência importante no mercado imobiliário por determinar o valor, custos ou direitos sobre um imóvel.

O valor de mercado de uma mercadoria é obtido através da comparação de várias outras disponíveis, ou seja, cuja escolhas de produtos semelhantes são possíveis. Quando essas mercadorias são trocadas livremente no mercado, os valores obtidos nessas transações definem o valor de cada uma.

Portanto, o valor de mercado de um imóvel pode ser identificado como o valor médio ou preço mais provável a ser atingido em transações normais, em um determinado momento (GONZÁLEZ, 2002), no qual comprador e vendedor tem o desejo em realizar a negociação,

ambos não compelidos (ABUNAHMAN, 2008), observando as condições de mercado, as características do imóvel e em um determinado tempo.

3.4.2. METODOLOGIAS APLICÁVEIS PARA AVALIAÇÃO PATRIMONIAL

Existem vários métodos para calcular o valor de mercado de uma propriedade, sendo que o principal é o método de comparação de dados de mercado. Entretanto, nem sempre é possível realizar tal modelo avaliatório, sendo necessário aplicar outras metodologias para se atingir o objetivo necessário.

3.4.2.1. MÉTODO EVOLUTIVO

É um método composto do cálculo do custo de aquisição do terreno e do custo de construção da edificação subtraído do Fator de Obsolescência, mais comumente conhecido como Depreciação (ABUNAHMAN, 2008).

Este método tem como fundamento principal a premissa de que um comprador não pagará mais que o necessário para construir um imóvel semelhante àquele que está adquirindo (GONZÁLEZ, 2002).

A NBR 14653-2/2011 indica que este método pode ser considerado como método eletivo na inexistência de dados de mercado para realização do método comparativo de mercado (ABNT, 2011).

3.4.2.2. MÉTODO DA RENDA

O método da renda é mais comumente utilizado para avaliações de valores locatícios em Ações Renovatórias, como postos de combustíveis, hotéis, cinemas, dentre outros comércios, pois depende essencialmente da capacidade de gerar lucros. Sendo necessário o cálculo de despesas e receitas, montagem do fluxo de caixa e estabelecimento de uma taxa de juros compatível do mercado.

Este método possui várias fórmulas de cálculos consagradas na literatura, sendo aplicadas para cada tipo de negócio.

3.4.2.3. MÉTODO INVOLUTIVO

Dentre os métodos de avaliação patrimonial, o método involutivo, também chamado como método do máximo aproveitamento é sem dúvida o mais subjetivo. Este baseia-se na concepção de um projeto hipotético que busca o máximo aproveitamento para o imóvel avaliando (ABNT, 2011).

O método involutivo busca identificar qual o melhor uso, em quantidade e qualidade. Portanto, todos os tipos de projetos permitidos para o imóvel avaliando deve ser investigado e realizados os cálculos orçamentários, despesas, taxas e verificação de viabilidade. Por fim, o projeto que demonstrar maior eficiência é definido como o valor de mercado do imóvel avaliando.

3.4.2.4. O MÉTODO COMPARATIVO DE MERCADO

O método comparativo de dados do mercado é o mais utilizado na avaliação patrimonial. Para se obter o valor de um imóvel através desse método é necessário coletar dados de ofertas e transações com imóveis semelhantes àquele que se deseja avaliar.

A NBR 14653-2/2011 apresenta 02 técnicas que podem ser adotadas: (i) tratamento de fatores; (ii) tratamento científico.

3.4.2.4.1. HOMOGENEIZAÇÃO DE FATORES

A avaliação de imóveis baseada no tratamento de fatores é uma das mais utilizadas no Brasil devido a sua simplicidade. Está técnica consiste em utilizar fatores ponderadores a diversas características dos imóveis e estatística descritiva, fazendo com que o rol de amostras seja homogeneizado, ou seja, os elementos das amostras são alterados por coeficientes corretivos, de modo a torná-los mais semelhantes, como por exemplo:

- Amostras localizadas em logradouros mais ou menos valorizados daquele em que o imóvel a ser avaliado se localiza;
- Amostras com áreas distintas;
- Amostras com acabamento superior ou inferior, mais novos ou antigos;
- Amostras baseadas em ofertas de vendas, ou seja, que a transação ainda não ocorreu efetivamente, podem ter um fator majorante do preço;

Os coeficientes devem ser calculados mediante a metodologias científicas, sendo necessário, justifica-los do ponto de vista teórico e prático (ABNT, 2011).

É necessário utilizar amostras mais semelhantes possíveis ao imóvel avaliando para evitar distorções, portanto, a NBR 14653-2/2011 determina o uso da Tabela 3 – Grau de fundamentação no caso de utilização do tratamento por fatores, que em seu item 4 determina o intervalo admissível de ajuste para o conjunto de fatores. A norma ainda enquadra o laudo avaliatório em Grau de Fundamentação e Precisão, como forma de minimizar os erros em avaliações patrimoniais.

3.4.2.4.2. TRATAMENTO CIENTÍFICO

Esta linha de avaliação trata o processo de ajustamento das diferenças de uma forma mais tecnicista, em que o embasamento teórico se baseia na inferência estatística para inferir o comportamento do mercado e a formação de valores.

A principal técnica adotada nesse tratamento é a Regressão Linear Múltipla.

3.4.2.4.2.1. REGRESSÃO LINEAR MÚLTIPLA

Baseando-se na regressão das variáveis independentes representativas de diversas características das propriedades, como: área do terreno, área construída, amenidades, localização, dentre várias outras, possuindo efeitos acumulativos no preço, é possível obter a equação de regressão linear múltipla, cuja forma generalizada dar-se-á por:

$$\boldsymbol{Y = a_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i}$$

Equação 3.1: Equação de regressão múltipla para estimativa de valores de bens.

Onde o preço do imóvel avaliando é definido por Y; $\beta_1 \ldots \beta_n$ são os regressores ou coeficientes da regressão; $x_{i1}, \ldots, x_{in}$ representam os valores das características do imóvel avaliando, ou seja, são as variáveis independentes ou explicativas; e ε_i representa os erros aleatórios do modelo.

Além da RLM, modelos mais como Regressão Espacial, DEA e RNA também são admitidas pela NBR 14653-2/2011.

3.4.2.4.2.2. REGRESSÃO ESPACIAL

O mercado imobiliário é uma excelente área para aplicação e desenvolvimento de técnicas de análises especiais

Essa metodologia utiliza modelos estimados pela Econometria Espacial, com base na ferramenta estatística Regressão Espacial, em busca da correção da autocorrelação espacial que pode apresentar problemas de tendenciosidade, inconsistência ou ineficiência em modelos de RLN (DANTAS; MAGALHÃES; VERGOLINO, 2007).

3.4.2.4.2.3. DEA

O DEA é uma metodologia desenvolvida por Charnes, Copper e Rhodes, baseando-se no trabalho de M. J. Farrell (Charnes *et al.*, 1978) que propõe uma programação matemática onde se mede a eficiência de diferentes unidades avaliadas (Buitrago-Suescún *et al*., 2017), mais conhecidas como unidades tomadoras de decisão (DMUs, do inglês *Decision Making Units*), através de *inputs* e *outputs* comuns, sem necessidade de especificação prévia de pesos ou relações entre essas variáveis (Ahn *et al*., 1988; Casado, 2007).

Nesta metodologia, ao invés de dar enfoque nas medidas de tendência central, como média ou a mediana, comumente observado na estatística tradicional, o enfoque se dá na comparação com a unidade mais eficiente. A Análise Envoltória de Dados sob Dupla Ótica avalia a eficiência de uma DMU por duas perspectivas, a maximização dos outputs e a minimização dos inputs, conforme é possível observar na Figura 3.4.

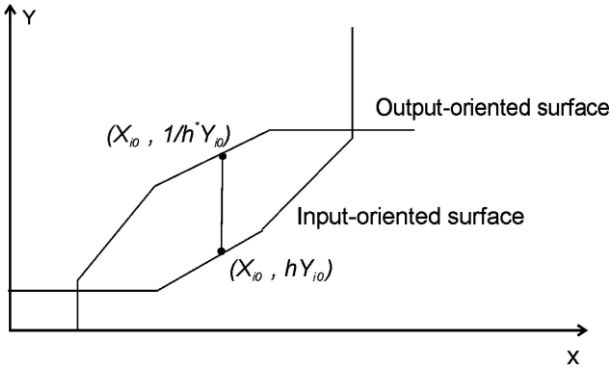

Figura 3.4: Método DEA sob Dupla Ótica. Fonte: (LINS; NOVAES; LEGEY, 2005)

Portanto, é possível determinar o valor de um imóvel sendo este considerado “eficiente” do ponto de vista do comprador e do vendedor, estabelecendo-se um intervalo de confiança para a negociação (LINS; NOVAES; LEGEY, 2005)

3.4.2.4.2.4. RNA

Redes Neurais Artificiais são sistemas de processamento de informação capazes de imitar a estrutura do cérebro humano e suas funções (ZHOU et al., 2018).

Com o avanço da tecnologia e sua aplicação ao mercado imobiliário associado ao avanço do aprimoramento de algoritmos de aprendizado e ao Big Data, a utilização de RNA na avaliação patrimonial tem se tornado bastante atrativa devido aos resultados de pesquisas realizadas por vários autores.

A NBR 14653-2/2011 cita a utilização de RNA para realização de avaliação de imóveis em seu tópico que trata sobre tratamentos científicos para induzir o valor a partir de amostras de mercado. No Anexo E, a norma brasileira traz uma série de recomendações para tratamento de dados por RNA.

O modelo mais utilizado modelo de RNA de três camadas *back-propagation* é o mais comumente utilizado na avaliação patrimonial (MCCLUSKEY et al., 2013). Este modelo consiste em três camadas (ABIDOYE; CHAN, 2017):

(i) Camada de entrada: onde as variáveis de entrada que representam atributos das propriedades são alimentadas na rede;

(ii) Camada oculta: local onde ocorre o processamento matemático;

(iii) Camada de saída: onde obtém-se o valor predito do imóvel;

3.4.3. DESCOBERTA DO CONHECIMENTO EM BASE DE DADOS

Tradicionalmente a forma de obtenção de informações a partir de dados baseia-se em uma interpretação pessoal associada em um conjunto de técnicas estatísticas. Nesse contexto são necessárias diversas condições para utilização dos dados, como o conhecimento prévio do modelo a ser estimado e da distribuição de probabilidades dos erros produzidos pelo modelo.

Com o aumento do volume de dados digitais associado a dispersão desses em diversos locais e formatos, surge a necessidade da utilização e construção de técnicas computacionais e ferramentas para auxiliar as pessoas a obter informações importantes a partir de uma massiva quantidade de dados de uma forma automatizada e inteligente buscando informações úteis.

Portanto, pode-se dizer que o Descobrimento de Conhecimento em Base de Dados (DCBD), proveniente do inglês *Knowledge-Discovery in Databases* (KDD), é o desenvolvimento de métodos e técnicas utilizadas para dar sentido aos dados brutos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

O DCBD e a Mineração de Dados em grandes bancos de dados tem sido tópico chave de pesquisa em vários setores da economia, como o setor imobiliário apresentado em várias pesquisas presentes nesse trabalho, podendo ser aplicado ao gerenciamento de informações e sistemas de apoio de decisões (MING-SYAN CHEN; JIAWEI HAN; YU, 1996).

Existe uma certa confusão sobre a utilização do termo DCBC e Mineração de Dados. Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), se refere ao processo geral da descoberta do conhecimento útil a partir de dados, enquanto a Mineração de Dados refere-se a uma etapa desse processo.

A Mineração de Dados é a aplicação de algoritmos específicos para extração de padrões de dados, entretanto, a aplicação indistinta de métodos de mineração de dados pode levar à descoberta de padrões sem sentido e inválidos.

Para fazer sentido e os padrões ou *insights* extraídos a partir de métodos de mineração de dados é fundamental a aplicação das etapas adicionais do DCBC, como preparação e seleção dos dados, limpeza dos dados e interpretação adequada dos resultados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). As etapas propostas podem ser diagramadas conforme a Figura 3.5 .

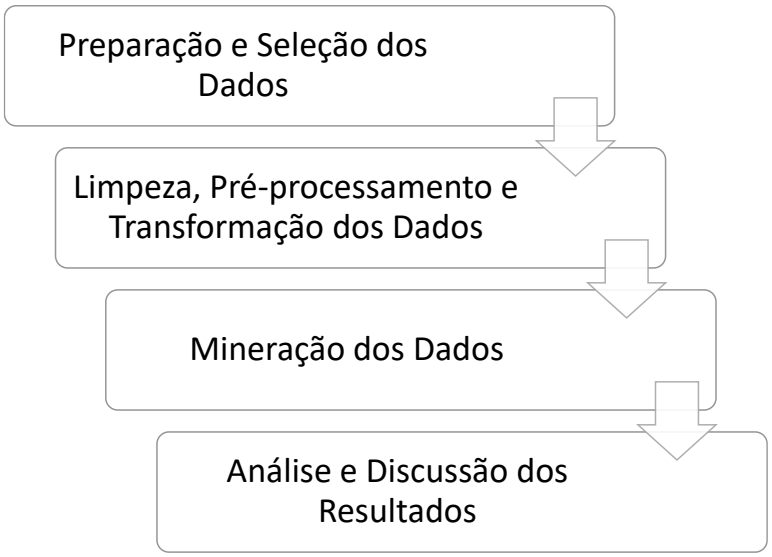

Figura 3.5: Etapas do DCBC de acordo com (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

3.4.4. TÉCNICAS PARA COLETA DE DADOS DISPERSOS NA WEB

A internet é uma fonte rica para anúncios e divulgação de imóveis tanto por imobiliárias quanto pelos proprietários, todavia, é praticamente impossível para um ser humano ser capaz de coletar e organizar uma quantidade de dados de fontes tão dispersas e inconsistentes. Logo, a utilização do *web scraping* para obtenção desses dados de anúncios se mostra bastante promissora (NEDER et al., 2017).

O *Web scraping*, também conhecido como *web harvesting ou web crawling*, é uma técnica de mineração de dados capaz de extrair dados de sites de forma automatizada. Isso pode ser feito usando programação, um robô, através de scripts de automação, ou até mesmo um software (PATEL, 2020). A aplicação dessa técnica em escala possibilita a extração de uma grande quantidade de dados, sem necessidade de uma supervisão direta de um ser humano.

3.4.5. ALGORITMOS DE MACHINE LEARNING

Além da avaliação do modelo hedônico, Regressão Linear Múltipla, foram utilizados os seguintes modelos de aprendizado de máquina

3.4.5.1. CART

O modelo CART (*Classification and Regression Trees*) é definido definidos por uma regra simples, dividir para conquistar. Segundo (BREIMAN et al., 2017), é um procedimento de

particionamento recursivo binário capaz de processar atributos contínuos e nominais como alvos e preditores.

Em cada divisão, os dados são divididos em dois grupos mutuamente exclusivos. O procedimento de divisão é então aplicado a cada grupo separadamente. A divisão é continua até obter uma árvore excessivamente ramificada, que é então podada de volta ao tamanho desejado (DE'ATH; FABRICIUS, 2000).

Em problemas onde deseja-se classificar os resultados é tipicamente caracterizado pela distribuição, em contrapartida, onde as variáveis targets são contínuas, ou seja, utiliza-se a regressão para encontrar os resultados, o valor médio da variável de resposta é empregado.

A CART não oferece medidas de desempenho interno para a seleção de árvores com base nos dados de treinamento, pois tais medidas são consideradas suspeitas. Em vez disso, o desempenho da árvore é sempre medido em dados de teste independentes (ou por meio de validação cruzada) e a seleção da árvore prossegue somente após a avaliação baseada em dados de teste (WU et al., 2008).

3.4.5.2. RANDOM FOREST

O *Random Forest* (RF)*,* ou Floresta Aleatória em português, é um algoritmo de classificação e regressão baseado nos métodos de bagging. O bagging constrói um conjunto de árvores, cada uma treinada a partir de um subconjunto (Db) obtida do conjunto de dados original (D) usando o seguinte procedimento de amostragem: dado um D com N dados, cria-se um subconjunto Db escolhendo aleatoriamente k linhas de dados de D com substituição, ou seja, após selecionar um subconjunto ele é imediatamente retornado para D e pode ser selecionado novamente. Depois de remover duplicatas, se N for grande e k = N, espera-se que Db contenha aproximadamente dois terços dos dados de D (ČEH et al., 2018).

De uma forma mais simplificada, os preditores utilizando o método bagging são capazes de gerar várias versões de um preditor e usá-los para obter um preditor agregado. A média da agregação sobre as versões de preditores retornam um resultado numérico para um caso de regressão e faz uma votação plural para casos ao prever uma classe (BREIMAN, 1996).

3.4.5.3. EXTREME GRADIENT BOOST

Boosting é um grupo de algoritmos em estratégia de conjunto que se baseia consecutivamente em estimadores fracos para gerar um estimador final forte. Um estimador fraco é um modelo que pode não ser muito preciso ou pode não levar muitos estimadores em consideração (LASOTA et al., 2014).

Ao contrário do Bagging que constrói cada modelo de forma independente e, em seguida, agrega as previsões dos modelos no final sem dar preferência a nenhum modelo, no Boosting, cada modelo construído dita quais recursos o próximo modelo enfocará. O reforço pode efetivamente converter estimadores fracos em fortes, construindo um modelo fraco, tirando conclusões sobre a importância de vários recursos e parâmetros e, em seguida, usando essas conclusões para construir um modelo novo e mais forte.

O Gradient Boosting (GB) é uma generalização do Boosting e visa reduzir o erro com cada modelo consecutivo até que um modelo final seja produzido (HO, 1998). No GB, um novo modelo de aprendizado fraco é adicionado em cada fase, e os modelos anteriores também permanecem inalterados. GB tem três componentes principais

Segundo Chen e Guestrin (2016), Extreme Gradiente Boosting (XGBoost) visa melhorar a escalabilidade e precisão do Gradient Boosting. Ele aproveita ao máximo a capacidade de computação por meio de processamento paralelo e validação cruzada integrada. Ele é projetado de forma a reduzir o tempo de computação e garantir o uso ideal dos recursos de memória e hardware. A principal característica que torna o XGBoost superior e único em comparação com outros algoritmos é a inclusão de regularização.

3.4.5.4. ADAPTIVE BOOSTING

O algoritmo AdaBoost é um dos métodos de *ensemble* mais importantes, por conta de sua previsão muito precisa, grande simplicidade e poder ser utilizado para uma ampla resolução de problemas (WU et al., 2008).

Um regressor AdaBoost é um meta estimador que começa ajustando um regressor no conjunto de dados original e, em seguida, ajusta cópias adicionais do regressor no mesmo conjunto de dados, mas onde os pesos das instâncias são ajustados de acordo com o erro da previsão atual. Como tal, os regressores subsequentes focam mais nos casos difíceis.

No boost, as máquinas são treinadas sequencialmente. O primeiro regressor é treinado em exemplos escolhidos com substituição (de tamanho N1) do conjunto de treinamento original. Em seguida, passamos todos os padrões de treinamento por esta primeira máquina e observamos quais estão mais errados. Para problemas de regressão, aqueles padrões cujos valores preditos diferem mais de seus valores observados são definidos como "mais" errados. Para os padrões mais errados, suas probabilidades de amostragem são ajustadas para que sejam mais propensos a serem escolhidos como membros do conjunto de treinamento para a segunda máquina. Portanto, à medida que avançamos na construção de máquinas, padrões difíceis têm maior probabilidade de aparecer nos conjuntos de treinamento. Assim, diferentes máquinas são melhores em diferentes partes do espaço de observação. Os regressores são combinados usando a mediana ponderada, por meio da qual os preditores que estão mais "confiantes" sobre suas previsões são ponderados com mais peso (DRUCKER, 1997).

3.5. METODOLOGIA

3.5.1. FERRAMENTAL DE APOIO

Foram utilizadas várias ferramentas para o atingimento do objetivo do presente estudo.

A linguagem de programação adotada foi o Python, principalmente devido a sua extensibilidade por conta da integração de diversas bibliotecas e frameworks construído pela comunidade que utiliza a linguagem para desenvolvimento de aplicações.

3.5.1.1. LINGUAGEM DE PROGRAMAÇÃO PYTHON

A linguagem de programação Python devido a sua natureza interativa de alto nível e seu ecossistema em desenvolvimento de bibliotecas científicas, é considerada uma das principais linguagens para computação científica, desenvolvimento algorítmico e análise exploratória de dados (PEDREGOSA et al., 2011).

Python é uma linguagem dinâmica e interpretada, ou seja, não precisa ser compilada para uma linguagem de baixo nível para que seja executada por um computador. O código fonte é lido pelo interpretador no momento da execução, como uma grande vantagem desse método o software desenvolvido pode ser utilizado em diversos sistemas operacionais sem a necessidade de recompilação do código, ou seja, é multiplataforma. Como desvantagem pode-se citar a perda de velocidade na execução do software desenvolvido nessa linguagem em comparação com uma linguagem compilada, como por exemplo o C/C++.

Esta linguagem foi criada por Guido van Rossum em 1991, tendo suporte para vários paradigmas de programação, com o objetivo de promover uma produção de programas mais simples e rápida, com sintaxe legível, porém permitindo também o desenvolvimento de algoritmos mais complexos. A linguagem Python apresenta uma biblioteca padrão que contém diversas classes e funções para realizar tarefas, além de ser possível utilizar bibliotecas externas, de acordo com a necessidade do usuário.

3.5.1.2. FRAMEWORK SCRAPY

Para extração de dados dos sites de anúncios de vendas de imóveis foi utilizado o framework Scrapy.

O Scrapy é um framework capaz de criar aplicações para realização de webscraping, coletando dados não estruturados em páginas da internet e tornando-os em dados estruturados ou semi-estrurados. O framework é composto de 5 partes:

Motor: responsável por controlar o fluxo de dados entre todos os componentes do sistema e disparar eventos quando determinadas ações ocorrem.

Agendador: recebe requisições e as enfileira para alimentar posteriormente quando o motor os solicitar.

***Downloader*:** responsável por buscar páginas da web e alimentá-las para o motor que, por sua vez, as alimenta para os *Spiders*.

***Spiders*:** são classes escritas pelos usuários do framework para analisar as respostas e extrair itens das páginas buscadas.

***Pipeline*:** O *Pipeline* é responsável por processar os itens uma vez que eles tenham sido extraídos pelos Spiders. O Pipeline realiza inúmeras tarefas de ETL e cria o banco de dados com as informações extraídas das páginas da web.

3.5.1.3. SCIKIT-LEARN

Recentemente a evolução do hardware de computadores permitiu um grande avanço na área de aprendizado de máquinas e de aplicação dessa tecnologia em diversos setores (FEURER et al., 2019a).

O Scikit-learn é um pacote gratuito em Python que estende a funcionalidade dos pacotes NumPy e SciPy para fornecer implementações de última geração de diversos algoritmos de Mineração de Dados, incluindo com isso, aqueles responsáveis pelo ML (PEDREGOSA et al., 2011).

A biblioteca mantém uma interface fácil de usar totalmente integrada com a linguagem Python e também é bastante rápido, apesar de ser escrito em uma linguagem interpretada, pois todos os colaboradores do ecossistema Scikit-learn, NumPy, SciPy, Matplotlib, dentre outras bibliotecas, são solicitados a otimizar o código em vários aspectos (JOVIC; BRKIC; BOGUNOVIC, 2014).

Segundo (PEDREGOSA et al., 2011), o Scikit-learn difere de outras caixas de ferramentas de ML implementadas em Python por vários motivos:

- A biblioteca é distribuída sob a licença BSD, sendo, portanto, de domínio público e podendo ser modificado sem nenhuma restrição;

- Incorpora código compilado para eficiência, como para implementações C / C ++ existentes no Cython (JOVIC; BRKIC; BOGUNOVIC, 2014);

- Utiliza NumPy e SciPy para facilitar a distribuição, ao contrário de PyMvPa (Hanke et al., 2009) que possui dependências opcionais, como R e Shogun;

- Pode utilizar diversos paradigmas de programação, inclusive a Programação Imperativa, ao contrário do PyBrain que usa uma estrutura de fluxo de dados.

3.5.2. DESENVOLVIMENTO DO FRAMEWORK

A metodologia para construção do framework dividiu-se em 4 etapas, sendo realizadas em três fases distintas:

Fase 01:

1. Extração de dados de anúncio de imóveis em sites da internet;

Fase 02:

2. Limpeza, organização e transformação dos dados coletados;

3. Enriquecimento da base de dados coletando informações de distâncias dos imóveis para pontes valorizantes e desvalorizantes;

Fase 03:

4. Treinamento e verificação da acurácia dos algoritmos de *Machine Learning*.

3.5.3. FASE 1: EXTRAÇÃO DE DADOS DE ANÚNCIO DE IMÓVEIS DISPERSOS EM SITES DA INTERNET

Esta pesquisa utilizou o framework Scrapy, para coleta de dados dispersos em páginas da internet de anúncio de imóveis.

O processo de coleta de dados está descrito no fluxo abaixo:

1. O script inicia buscando a URL do website de anúncio de imóveis, são passados como parâmetros de consulta: o bairro, cidade, o tipo de imóvel e o tipo de comercialização (venda ou locação). Os parâmetros utilizados nesse estudo foram:
    a. Bairro: Botafogo;
    b. Cidade: Rio de Janeiro;
    c. Tipo de imóvel: Apartamentos;
    d. Tipo de comercialização: Compra;
2. No código-fonte do webscraping identifica-se as tags HTML que contenham as informações a serem coletadas;
3. O framework identifica o número de páginas indexadas e seus respectivos links;
4. Através do *downloader* executa o download das informações necessárias de cada página visitada;
5. O framework interage com os links de cada página indexada, percorrendo todas as páginas que contenham as informações desejada;
6. Repete os itens de 2 a 5 até que todas as páginas sejam lidas e seus dados coletados;

7. Através do pipeline, estrutura os dados da forma desejada, nesse processo foi utilizado o modelo de dados JSON;

8. Por fim, o script armazena os dados em um arquivo JSON.

3.5.4. FASE 2: LIMPEZA, ORGANIZAÇÃO E TRANSFORMAÇÃO

Na primeira etapa da limpeza de dados foram removidas as observações que continham dados nulos para qualquer atributo.

Posteriormente, foram analisados os extremos de cada variáveis, ou seja, valores máximos e mínimos, pois são locais onde encontram-se mais erros.

Com as maiores inconsistências removidas, passou-se a etapa de enriquecimento dos dados. Utilizando a API do Google Maps, Geocoding API, obteve-se as coordenadas geográficas Latitude e Longitude de cada amostra do *dataset*.

Na sequência, após a obtenção dos dados geográficos, utilizou-se a fórmula de Haversine para calcular as distancias dos locais das amostras para pontos valorizantes e desvalorizantes na região avaliada.

A fórmula Haversine calcula a distância entre o ponto de localização principal e o ponto de destino com base no comprimento da linha reta tomando o valor da longitude e latitude de entrada (WINARNO; HADIKURNIAWATI; ROSSO, 2017).

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Equação 3.2: Equação da distância baseada na fórmula de Haversine.

onde:

φ1, φ2 são a latitude do ponto 1 e a latitude do ponto 2 (em radianos),

λ1, λ2 são a longitude do ponto 1 e longitude do ponto 2 (em radianos).

Os pontos valorizantes definidos foram os seguintes locais: (i) Estação do metrô; (ii) Praia de Botafogo; (iii) Shopping Rio Sul; (iv) Praia do Leme; (v) Praia de Copacabana.

Como ponto desvalorizante foi indicada a posição do Morro Santa Marta, devido a periculosidade da região.

Por fim, após a etapa de enriquecimento de dados, foram removidos os outliers com base na metodologia de valores limites baseados em amplitude interquartil:

$$LI = Q_1 - c * IIQ$$

$$LS = Q_3 + c * IIQ$$

Equação 3.3: Limite inferior e superior dos quartis (FARIAS; LAURENCEL, 2006).

Para a constante c foi utilizado o valor de 1.5, pois este valor é capaz de captar aproximadamente 99% dos dados sob uma curva normal, entretanto não capta 100% dos dados, eliminando assim, dados outliers (BUSSAB; MORETTIN, 2017).

3.5.5. FASE 3: TREINAMENTO E AVALIAÇÃO DA PERFORMANCE DOS ALGORITMOS DE MACHINE LEARNING

Após o saneamento das amostras, os dos dados restantes foram utilizados para treinamento dos modelos de *Machine Learning*: *CART*, *Random Forest*, *XGBoost* e *AdaBoost*. Além dos algoritmos anteriores, o modelo hedônico baseado na regressão linear múltipla também foi construído a partir dos dados do *dataset*.

Foram desenvolvidos códigos utilizando a linguagem de programação Python, as bibliotecas Pandas, Scikit-learn e Numpy para realização das atividades de treinamento e verificação da acurácia dos algoritmos.

Todos os algoritmos foram utilizados sem a aplicação de técnicas de *tuning* para melhor análise da eficiência de cada modelo. Para verificação da acurácia foi utilizado o processo de *cross-validation* (validação cruzada).

A validação cruzada é uma técnica de validação de modelo usado principalmente em configurações onde o objetivo é a previsão e se deseja estimar a precisão com que um modelo preditivo será executado na prática.

Em um problema de predição, um modelo geralmente recebe um conjunto de dados conhecidos no qual o treinamento é executado (conjunto de dados de treinamento), e um conjunto de dados

de dados desconhecidos (ou dados vistos pela primeira vez) contra os quais o modelo é testado (chamado conjunto de dados de validação ou conjunto de teste).

O objetivo da validação cruzada é testar a capacidade do modelo de prever novos dados que não foram usados em sua estimativa, a fim de sinalizar problemas como *overfitting* ou viés de seleção e dar uma visão sobre como o modelo irá generalizar para um conjunto de dados independente (SENI; ELDER, 2010).

O seguinte procedimento é seguido para cada um dos subconjuntos:

1. Um modelo é treinado usando parte dos dados como treinamento;

2. O modelo resultante é validado utilizando o restante dos dados, ou seja, é usado como um conjunto de teste para calcular uma medida de desempenho, como precisão.

A medida de desempenho relatada pela validação cruzada *k-fold* é então a média dos valores calculados no loop Figura 3.6.

Em resumo, a validação cruzada combina medidas, geralmente a média, de adequação na predição para derivar uma estimativa mais precisa do desempenho de predição do modelo.

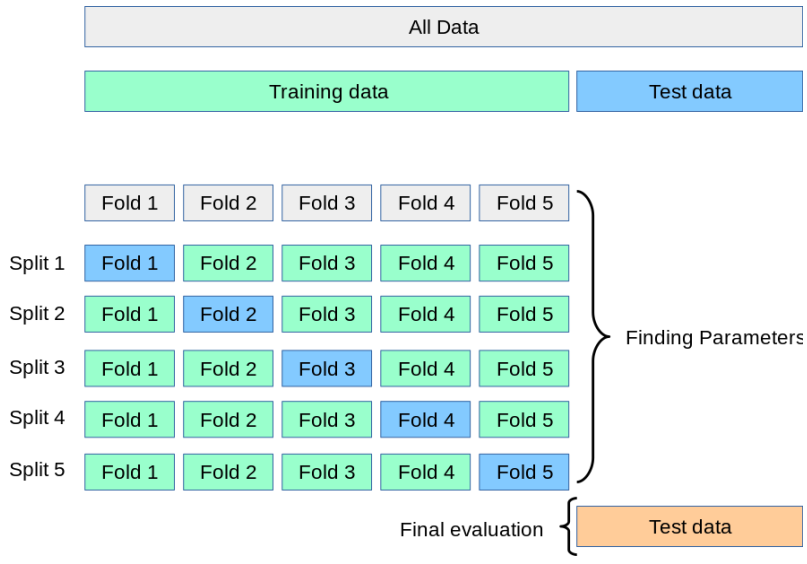

Figura 3.6: Processo de análise de eficiência de modelos através da validação cruzada.

Foram utilizadas três importantes métricas de avaliação de modelos de regressão: RMSE, MAPE e R2.

A métrica Raiz do Erro Quadrático Médio (HALLAK; PEREIRA FILHO, 2011), em inglês Root Mean Squared Error (RMSE), é baseada no quadrado da média das diferenças entre o valor predito e o real, conforme apresentado na fórmula:

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

Equação 3.4: Equação da métrica RMSE.

Onde:

y representa o valor real;

$\hat{y}$ representa o valor predito pelo modelo;

n refere-se ao tamanho amostral.

O erro percentual absoluto médio (Mean Absolute Percentage Error – MAPE), é uma métrica muito utilizada, pois seu valor é mais fácil de interpretar, pois, semelhantemente ao Coeficiente de Determinação, o valor pode ser expressado em um formato de porcentagem. O MAPE (Equação 3.5) representa a proporção do erro em relação ao valor real (NIU; NIU, 2019b).

$$\mathrm{MAPE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n} \frac{|y_i - \hat{y}_i|}{y_i} * 100$$

Equação 3.5: Equação da métrica MAPE.

Por fim, o Coeficiente de Determinação R^2 (Equação 3.6) tem a propriedade extremamente útil de fornecer um valor absoluto para a adequação de um modelo, como uma estatística resumida que descreve a quantidade de variância explicada (variando de 0 a 1) é bastante intuitivo. Como o R^2 não tem unidade, ele é extremamente útil como um índice de resumo para modelos estatísticos porque pode-se avaliar objetivamente o ajuste dos modelos e comparar os valores de R^2 entre os estudos de maneira semelhante às estatísticas de tamanho de efeito padronizado em algumas circunstâncias (NAKAGAWA; SCHIELZETH, 2013).

Assim, quanto maior o R^2, mais explicativo é o modelo linear, ou seja, melhor ele se ajusta à amostra. O valor do coeficiente de determinação indica qual a porcentagem dos valores de y pode ser explicado pelo modelo.

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

Equação 3.6: Equação do Coeficiente de Determinação (R^2).

3.6. RESULTADOS

Para coleta de dados automatizada, utilizando webscraping, foram utilizados os parâmetros abaixo para criação do estudo de caso:

i. Bairro: Botafogo;

ii. Cidade: Rio de Janeiro;

iii. Tipo de imóvel: Apartamentos;

iv. Tipo de comercialização: Compra;

A coleta de dados resultou em 2916 dados de anúncios de imóveis do tipo apartamento. O dataset inicial consiste em 7 atributos, conforme apresentado na Tabela 3.1.

Tabela 3.1: Definição dos atributos do *dataset*

Atributo	**Definição**
Preço de venda	Valor de venda da propriedade (ponto flutuante)
Área	Área construída do imóvel (ponto flutuante)
Nº de quartos	Quantidade de quartos do imóvel (inteiro)
Nº de banheiros	Quantidade de banheiros do imóvel (inteiro)
Nº de garagens	Quantidade de vagas de garagem disponíveis para o imóvel (inteiro)
Taxa	Taxa condominial do apartamento (ponto flutuante)
CEP	Código postal que identifica o local do imóvel

Esses dados brutos então foram utilizados na próxima fase, no processo de Limpeza, organização e transformação.

Utilizando os parâmetros de entrada para o O *dataset* limpo resultou em 2026 observações do total coletado 2916. As amostras saneadas foram utilizadas na etapa seguinte, treinamento de modelos de *machine learning* e do modelo hedônico, regressão linear múltipla.

O resumo das informações do dataset estão na tabela abaixo:

Tabela 3.2: Resumo de informações contidas no dataset

	Preço	Área	Quartos	Banheiros	Garagem	Taxa	Metrô	Praia de Botafogo	Shopping Rio Sul	Praia do Leme	Praia de Copacabana	Morro Santa Marta
# Linhas	2.026	2.026	2.026	2.026	2.026	2.026	2.026	2.026	2.026	2.026	2.026	2.026
Média	1.046.389,44	87,16	2,34	1,73	0,91	784,01	602,57	867,30	1.204,92	2.146,92	2.369,61	970,82
Desvio Padrão	479.282,53	31,38	0,75	0,73	0,66	543,19	304,29	505,22	462,25	433,01	335,29	393,13
Valor Mínimo	76.000,00	9,00	1	1	0	0,00	16,96	30,38	147,36	1.163,88	1.749,66	36,92
1º Quartil	700.000,00	69,00	2	1	0	401,75	404,94	379,80	845,63	1.778,17	2.166,13	689,84
Mediana	989.000,00	85,00	2	2	1	800,00	625,86	884,97	1.224,88	2.166,58	2.314,72	977,03
3º Quartil	1.349.350,00	105,00	3	2	1	1.149,25	800,11	1.326,67	1.624,64	2.520,17	2.615,11	1.205,46
Valor Máximo	2.807.300,00	182,00	4	3	2	2.200,00	1.308,96	1.888,08	2.028,48	3.004,19	3.265,36	1.898,09

A Figura 3.7 apresenta o coeficiente de correlação entre as variáveis do *dataset* limpo.

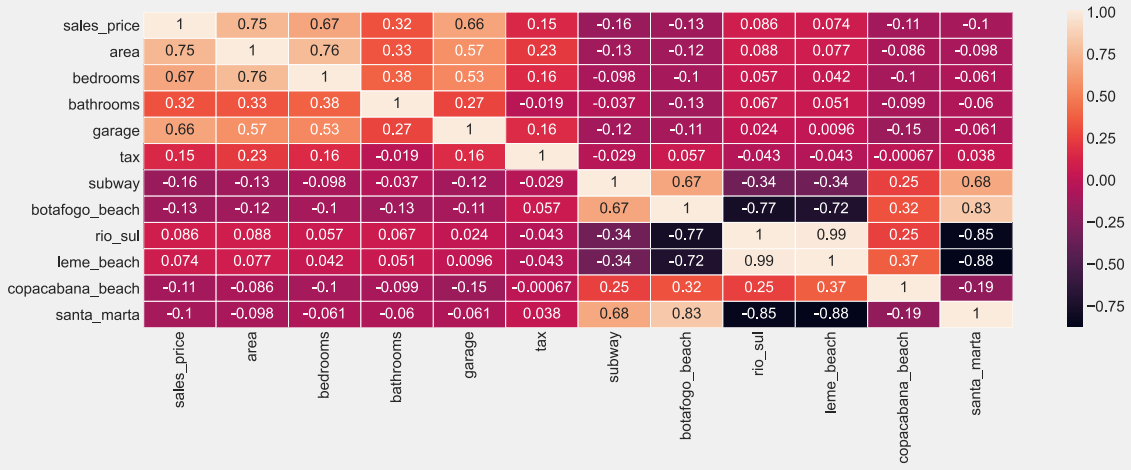

Figura 3.7: Gráfico de correlação entre as variáveis.

Como é possível observar no *heatmap*, os atributos área, número de quartos e quantidade de vagas de garagem apresentam as maiores correlações com o valor do imóvel. A variável número de banheiros vem na sequência, seguida pela taxa e distância até o metrô.

Os conjuntos de dados foram usados para treinar e avaliar os modelos de *Machine Learning* mencionados, assim como a RLM. Antes de treinar qualquer um dos modelos, as diferentes técnicas de pré-processamento descritas antes foram aplicadas aos dados. Os modelos foram avaliados com base em sua precisão na estimativa dos preços das observações utilizando a validação cruzada.

A Figura 3.8 apresenta os resultados obtidos pelos modelos utilizados no estudo. É perceptível que os modelos ensemble Random Forest e XGBoost possuem os menores valores de RMSE e MAPE, bem como os maiores valores para o coeficiente de determinação (R^2), indicando que são modelos que entregam maiores assertividades em relação as avaliações patrimoniais.

A RLM surpreendente ficou à frente dos modelos CART e AdaBoost, tendo uma performance intermediária. Vale ressaltar que nenhum modelo passou por processo de *tuning*.

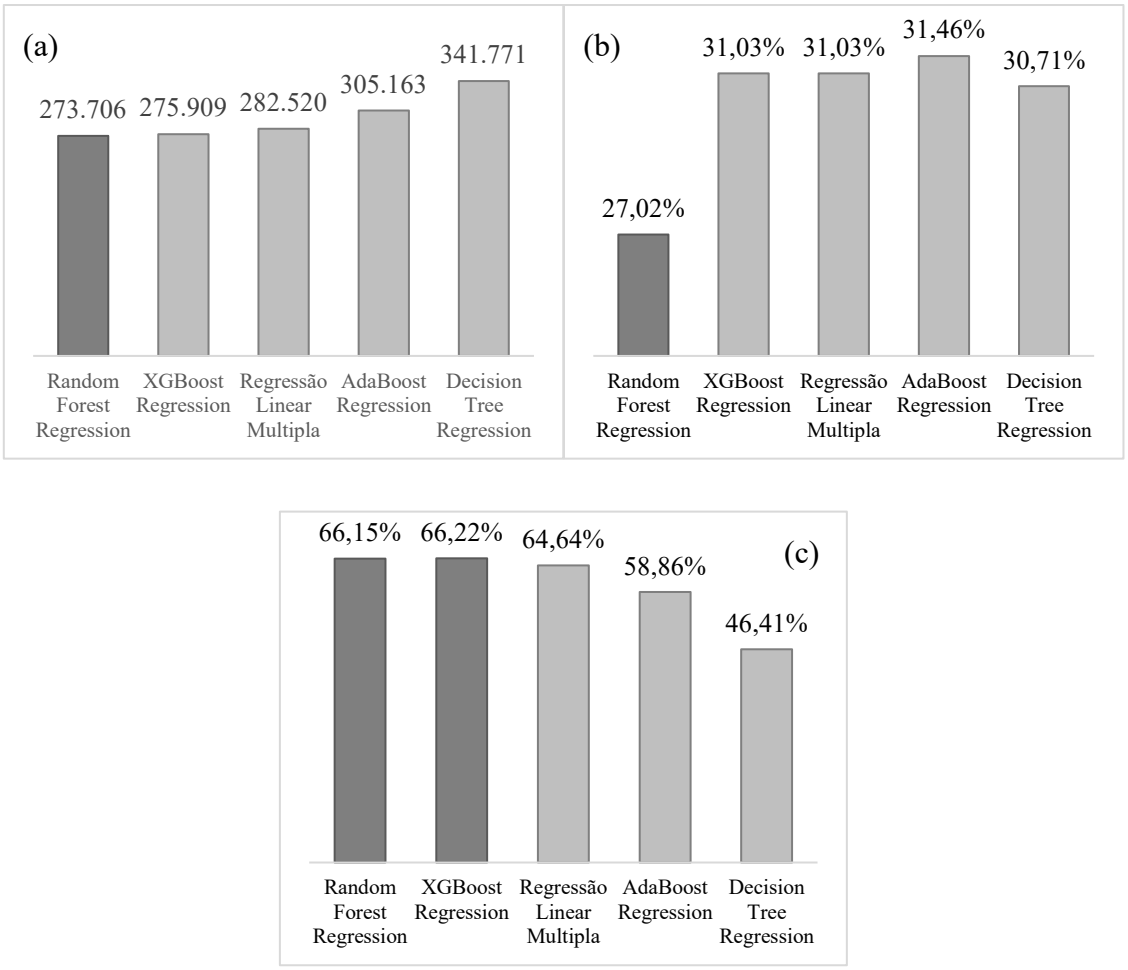

Figura 3.8: Comparação de performance dos modelos: (a) RMSE, (b) MAPE e (c) R²

Comparando os resultados obtidos com a literatura, existem alguns pontos importantes a ressaltar.

Entre os resultados obtidos destaca-se a liderança de algoritmo do tipo *Bagging* seguido logo por algoritmos *Boosting*. No estudo apresentado por (KUMKAR et al., 2018b), os algoritmos mais precisos foram *XGBoost*, *Gradient Boosting* e o *Random Forest*, sequencialmente em 1º, 2º e 3º lugares. Portanto os algoritmos baseados em técnicas *Boosting* ficando à frente de algoritmos *Bagging*.

Entretanto, o que mais destaca-se neste comparativo são os resultados obtidos pelo critério MAPE. O melhor resultado no presente trabalho foi com o algoritmo Random Forest, com 27,02%, comparando com o trabalho do colega que obteve 15%, quase a metade do valor encontrado. Tratando-se de *XGBoost* a diferença foi de 31,03% contra 13,7%, uma precisão 2 vezes melhor.

Já em (NIU; NIU, 2019b), os autores desenvolveram um algoritmo próprio que obteve a liderança na precisão das avaliações patrimoniais, em segundo lugar um algoritmo de redes neurais e tendo logo na sequência o *Random Forest Regression*. Eliminando os dois primeiros algoritmos devido suas características distintas, um algoritmo do tipo *Bagging* ficou à frente de algoritmos do tipo *Boosting,* portanto, um resultado semelhante com o presente estudo. Tratando-se de performance o algoritmo de Random Forest obteve aproximadamente 19% na métrica MAPE, ficando à frente do resultado apresentado neste trabalho.

Por fim, vale ressaltar que os autores (KUMKAR et al., 2018b) utilizaram *tuning* dos hiper parâmetros para aprimorar o desempenho *d*os modelos, o que influencia diretamente na precisão do modelo. Já em (NIU; NIU, 2019b) os autores não informaram a utilização de *tuning*, entretanto os dados utilizados pelos autores foram cedidos por uma empresa imobiliária, sendo a forma de coleta de dados para treinamento dos modelos divergentes do presente estudo.

3.7. CONCLUSÕES

Neste trabalho foi apresentado um modelo de avaliação patrimonial totalmente automatizado, desde a coleta dos dados até a entrega de modelos de *ML* treinados e capazes de predizer valores de novos imóveis entrantes.

Foi desenvolvido uma aplicação capaz de rastrear dados de anúncios de vendas de imóveis postados no site olx.com.br através de técnicas de webscraping recuperando o preço de venda e atributos estruturais como: área, quantidade de quartos, quantidade de banheiros, quantidade de vagas de estacionamento e valor da taxa condominial; e atributos de localização, sendo estes relacionados as distâncias para pontos valorizantes e desvalorizantes.

Para a entrega dos resultados 4 algoritmos de *ML*, sendo 3 do tipo ensemble e 1 do tipo árvore, além do modelo hedônico baseado no método dos mínimos quadrados, regressão linear múltipla, foram analisados e comparados para verificação da eficiência e acurácia tratando-se de avaliações patrimoniais no bairro de Botafogo, Rio de Janeiro/BR.

Para examinar os determinantes do preço da habitação através dos modelos de *ML* e pelo modelo hedônico, foram selecionados todos os atributos. A área construída foi o fator

mais influente, apresentando correlação positiva com o preço do imóvel (r = 75%), ou seja, imóveis com maior área possuem a tendência de serem mais caros e mais luxuosos. Além disso, os outros atributos estruturais tiveram maiores correlações do que os atributos de localização que acabaram influenciando pouco no valor do imóvel.

Os resultados confirmam que os algoritmos ensemble, principalmente os modelos que utilizam a técnica de *bagging*, são melhores estimadores de valores imobiliários do que os modelos hedônicos. A norma brasileira NBR 14653 precisa acompanhar a evolução tecnológica e inserir as modelagens de *ML* em seu escopo, bem como as determinações necessárias para utilização desses modelos, dando embasamento técnico e jurídico aos engenheiros avaliadores que utilizarem tais técnicas para realização de suas atividades.

Comparado com o modelo de avaliação de imóveis tradicional, o modelo proposto neste estudo tem as seguintes vantagens. Primeiramente, a modelagem proposta pode produzir dados de alta qualidade sem a necessidade de supervisão de uma pessoa, eliminando, com isso, dois dos maiores problemas da engenharia de avaliação, a coleta de dados em quantidade e qualidade suficiente e a subjetividade, pois não há interferência humana no processo. Em segundo lugar, os modelos de treinamento podem ser utilizados concomitantemente, e escolher aquele que entrega o melhor resultado, ou seja, uma maior acurácia nas avaliações. Terceiro, o sistema proposto pode produzir o preço de previsão de vendas de imóveis com precisão razoável, graças ao aprendizado do conjunto que integra diferentes algoritmos de aprendizado de máquina.

Esse último ponto pode ser aprimorado utilizando técnicas de LNP e reconhecimento de imagens, identificando outros atributos importantes, como: características de acabamento do imóvel, estrutura condominial (piscinas, quadras, dentre outras amenidades que influenciam no valor imobiliário), vista para o mar, para citar alguns.

Os algoritmos utilizados também podem ter um aumento de performance utilizando técnicas de *tuning*, processo de otimização dos hiper parâmetros para melhor controle do aprendizado dos modelos de *ML*.

3.8. REFERÊNCIAS BIBLIOGRÁFICAS

ABIDOYE, R. B.; CHAN, A. P. C. Artificial neural network in property valuation: application framework and research trend. **Property Management**, 16 out. 2017.

ABIDOYE, R. B.; CHAN, A. P. C. Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network. **Pacific Rim Property Research Journal**, v. 24, n. 1, p. 71–83, 2 jan. 2018a.

ABIDOYE, R. B.; CHAN, A. P. C. Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network. **Pacific Rim Property Research Journal**, v. 24, n. 1, p. 71–83, 2 jan. 2018b.

ABIDOYE, R. B.; CHAN, A. P. C. Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network. **Pacific Rim Property Research Journal**, v. 24, n. 1, p. 71–83, 2 jan. 2018c.

ABNT. **ABNT NBR 14653-2 - Avaliação de bens Parte 2: Imóveis Urbanos**. [s.l.] ABNT, 2011.

ABUNAHMAN, S. A. **Curso básico de engenharia legal e de avaliações**. 4. ed. São Paulo: PINI, 2008.

ACEBIP. **Associação Brasileira das Entidades de Crédito Imobiliário e Poupança**. Disponível em: <https://www.abecip.org.br/>. Acesso em: 27 mar. 2020.

AHMED, H. et al. Producing Standard Rules for Smart Real Estate Property Buying Decisions based on Web Scraping Technology and Machine Learning Techniques. **International Journal of Advanced Computer Science and Applications (IJACSA)**, v. 11, n. 3, 40/30 2020a.

AHMED, H. et al. Producing standard rules for smart real estate property buying decisions based on web scraping technology and machine learning techniques. **International Journal of Advanced Computer Science and Applications**, v. 11, n. 3, p. 498–505, 2020b.

AHN, T.; CHARNES, A.; COOPER, W. W. Some statistical and DEA evaluations of relative efficiencies of public and private institutions of higher learning. **Socio-Economic Planning Sciences**, v. 22, n. 6, p. 259–269, jan. 1988.

ANNAMORADNEJAD, R. et al. **Using Web Mining in the Analysis of Housing Prices: A Case study of Tehran**. 2019 5th International Conference on Web Research,

ICWR 2019. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2019a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069907225&doi=10.1109%2fICWR.2019.8765250&partnerID=40&md5=6faec40df83f2429b4c6357594c31367>

ANNAMORADNEJAD, R. et al. **Using Web Mining in the Analysis of Housing Prices: A Case study of Tehran**. 2019 5th International Conference on Web Research (ICWR). **Anais**... In: 2019 5TH INTERNATIONAL CONFERENCE ON WEB RESEARCH (ICWR). abr. 2019b.

BACEN. **Banco Central do Brasil**. Disponível em: <https://www.bcb.gov.br/estatisticas/mercadoimobiliario>. Acesso em: 27 mar. 2020.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1 ago. 1996.

BREIMAN, L. et al. **Classification And Regression Trees**. Boca Raton: Routledge, 2017.

BUITRAGO-SUESCÚN, O. Y. et al. Data Envelopment Analysis for Efficiency Measurement on Higher Education Institutions: a State of the Art Review. **Revista Científica General José María Córdova**, v. 15, n. 19, p. 147–173, jun. 2017.

BUREAU, U. C. **Census.gov**. Disponível em: <https://www.census.gov/en.html>. Acesso em: 29 maio. 2021.

BUSSAB, W. DE O.; MORETTIN, P. A. **Estatística Básica**. [s.l.] SARAIVA EDITORA, 2017.

CASADO, F. L. ANÁLISE ENVOLTÓRIA DE DADOS: CONCEITOS, METODOLOGIA E ESTUDO DA ARTE NA EDUCAÇÃO SUPERIOR. v. 20, n. 01, p. 13, 2007.

CBIC. **Indicadores Imobiliários Nacionais**. Disponível em: <https://cbic.org.br/estudos/>. Acesso em: 13 dez. 2020.

ČEH, M. et al. Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. **ISPRS International Journal of Geo-Information**, v. 7, n. 5, p. 168, maio 2018.

CHARNES, A.; COOPER, W. W.; RHODES, E. Measuring the efficiency of decision making units. **European Journal of Operational Research**, v. 2, n. 6, p. 429–444, nov. 1978.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 785–794, 13 ago. 2016.

CREA. **CREA – Canadian Real Estate Association**, 2021. Disponível em: <https://www.crea.ca/>. Acesso em: 29 maio. 2021

CROSBY, N.; LAVERS, A.; MURDOCH, J. Property valuation variation and the "margin of error" in the UK. **Journal of Property Research**, v. 15, n. 4, p. 305–330, 1 jan. 1998.

DANTAS, R. A.; MAGALHÃES, A. M.; VERGOLINO, J. R. DE O. Avaliação de imóveis: a importância dos vizinhos no caso de Recife. **Economia Aplicada**, v. 11, n. 2, p. 231–251, jun. 2007.

DANTAS, RUBENS ALVES, R. **UMA NOVA METODOLOGIA PARA AVALIAÇÃO DE IMÓVEIS UTILIZANDO REGRESSÃO ESPACIAL**. . In: XXI COBREAP. ESPIRITO SANTO: 2001.

DE'ATH, G.; FABRICIUS, K. E. Classification and regression trees: a powerful yet simple technique for ecological data analysis. **Ecology**, v. 81, n. 11, p. 3178–3192, 1 nov. 2000.

DO, A. Q.; GRUDNITSKI, G. A neural network approach to residential property appraisal. **The Real Estate Appraiser**, v. 58, p. 38–45, 1 jan. 1992.

DRUCKER, H. Improving Regressors Using Boosting Techniques. **Proceedings of the 14th International Conference on Machine Learning**, 17 ago. 1997.

FARIAS, A. M. L. DE; LAURENCEL, L. DA C. **Estátistica Descritiva**. UFF: [s.n.].

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–37, 15 mar. 1996.

FEURER, M. et al. Auto-sklearn: Efficient and Robust Automated Machine Learning. In: HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. (Eds.). . **Automated Machine Learning**. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2019a. p. 113–134.

FEURER, M. et al. Efficient and Robust Automated Machine Learning. p. 9, 2019b.

GONZÁLEZ, M. A. S. **Aplicação de Técnicas de Descobrimento de Conhecimento em Bases de Dados e de Inteligência Artificial em Avaliação de Imóveis**. Rio Grande do Sul: UFRGS, dez. 2002.

GOVERNMENT OF CANADA, S. C. **The Daily — New Housing Price Index, April 2021**. Disponível em: <https://www150.statcan.gc.ca/n1/daily-quotidien/210520/dq210520d-eng.htm>. Acesso em: 29 maio. 2021.

GROVER, R. Mass valuations. **Journal of Property Investment & Finance**, 7 mar. 2016.

HALLAK, R.; PEREIRA FILHO, A. J. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS: sensibilidade a variações com os esquemas de advecção e assimilação de dados. **Revista Brasileira de Meteorologia**, v. 26, p. 591–608, dez. 2011.

HARTEN, J. G.; KIM, A. M.; BRAZIER, J. C. Real and fake data in Shanghai's informal rental housing market: Groundtruthing data scraped from the internet. **Urban Studies**, v. 58, n. 9, p. 1831–1845, 1 jul. 2021.

HO, T. K. The random subspace method for constructing decision forests. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 20, n. 8, p. 832–844, ago. 1998.

IBGE. **Pesquisa Anual da Indústria da Construção - PAIC | IBGE**. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/industria/9018-pesquisa-anual-da-industria-da-construcao.html?=&t=destaques>. Acesso em: 27 mar. 2020.

JIANG, H.; JIN, X.-H.; LIU, C. The effects of the late 2000s global financial crisis on Australia's construction demand. **Construction Economics and Building**, v. 13, n. 3, p. 65–79, 18 set. 2013.

JOVIC, A.; BRKIC, K.; BOGUNOVIC, N. **An overview of free software tools for general data mining**. 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). **Anais**... In: 2014 37TH INTERNATIONAL CONVENTION ON INFORMATION AND COMMUNICATION TECHNOLOGY, ELECTRONICS AND MICROELECTRONICS (MIPRO). maio 2014.

KLAMER, P.; BAKKER, C.; GRUIS, V. Research bias in judgement bias studies – a systematic review of valuation judgement literature. **Journal of Property Research**, v. 34, n. 4, p. 285–304, 2 out. 2017.

KUMKAR, P. et al. **Comparison of Ensemble Methods for Real Estate Appraisal**. Proceedings of the 3rd International Conference on Inventive Computation Technologies, ICICT 2018. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2018a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082649654&doi=10.1109%2fICICT43934.2018.9034449&partnerID=40&md5=ec62aa51060a948ee7a4f759733ffb59>

KUMKAR, P. et al. **Comparison of Ensemble Methods for Real Estate Appraisal**. 2018 3rd International Conference on Inventive Computation Technologies (ICICT). **Anais**... In: 2018 3RD INTERNATIONAL CONFERENCE ON INVENTIVE COMPUTATION TECHNOLOGIES (ICICT). nov. 2018b.

LANCASTER, K. J. A New Approach to Consumer Theory. **Journal of Political Economy**, v. 74, n. 2, p. 132–157, 1966.

LASOTA, T. et al. **Comparison of Ensemble Approaches: Mixture of Experts and AdaBoost for a Regression Problem**. (N. T. Nguyen et al., Eds.)Intelligent Information and Database Systems. **Anais**...: Lecture Notes in Computer Science.Cham: Springer International Publishing, 2014.

LINS, M. P. E.; NOVAES, L. F. DE L.; LEGEY, L. F. L. Real Estate Appraisal: A Double Perspective Data Envelopment Analysis Approach. **Annals of Operations Research**, v. 138, n. 1, p. 79–96, 1 set. 2005.

MCCLUSKEY, W. J. et al. Prediction accuracy in mass appraisal: a comparison of modern approaches. **Journal of Property Research**, v. 30, n. 4, p. 239–265, 1 dez. 2013.

MELANDA, E.; HUNTER, A.; BARRY, M. Identification of locational influence on real property values using data mining methods. **Cybergeo : European Journal of Geography**, 4 fev. 2016.

MING-SYAN CHEN; JIAWEI HAN; YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 866–883, dez. 1996.

MOHER, D. et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. **PLOS Medicine**, v. 6, n. 7, p. e1000097, 21 jul. 2009.

MTE. **CAGED**. Disponível em: <http://pdet.mte.gov.br/caged>. Acesso em: 27 mar. 2020.

NAKAGAWA, S.; SCHIELZETH, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. **Methods in Ecology and Evolution**, v. 4, n. 2, p. 133–142, 1 fev. 2013.

NEDER, H. D. et al. Índice de defasagem do imposto predial e territorial urbano (IPTU) dos Municípios de Minas Gerais: um estudo de caso para Uberlândia (MG). **Revista ESPACIOS**, v. 38, n. 46, 6 out. 2017.

NIU, J.; NIU, P. **An intelligent automatic valuation system for real estate based on machine learning**. (T. J.M.R.S, Ed.)ACM International Conference Proceeding Series. **Anais**...Association for Computing Machinery, 2019a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077820799&doi=10.1145%2f3371425.3371454&partnerID=40&md5=25936a4abf77f70deb07c0299ab0c874>

NIU, J.; NIU, P. **An intelligent automatic valuation system for real estate based on machine learning**. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing. **Anais**...: AIIPCC '19.New York, NY, USA: Association for Computing Machinery, 19 dez. 2019b. Disponível em: <http://doi.org/10.1145/3371425.3371454>. Acesso em: 14 dez. 2020

NUNES, D. B. et al. Modelo de regressão linear múltipla para avaliação do valor de mercado de apartamentos residenciais em Fortaleza, CE. **Ambiente Construído**, v. 19, n. 1, p. 89–104, mar. 2019.

PATEL, J. M. Web Scraping in Python Using Beautiful Soup Library. In: PATEL, J. M. (Ed.). . **Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale**. Berkeley, CA: Apress, 2020. p. 31–84.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825–2830, 2011.

PELLI, A. N. **Redes neurais artificiais aplicadas às avaliações em massa estudo de caso para a cidade de Belo Horizonte / MG**. [s.l.] Universidade Federal de Minas Gerais, 10 mar. 2006.

PETERSON, S.; FLANAGAN, A. Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. **Journal of Real Estate Research**, American Real Estate Society. v. 31(2), p. 147–164, 2009.

RICS. **The future of valuations**. Disponível em: <https://www.rics.org/globalassets/rics-website/media/knowledge/research/insights/future-of-valuations-insights-paper-rics.pdf>. Acesso em: 5 jun. 2021.

ROSEN, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. **Journal of Political Economy**, v. 82, n. 1, p. 34–55, jan. 1974.

ROYAL INSTITUTION OF CHARTERED SURVEYORS (ED.). **RICS valuation - Global standards: incorporating the IVSC International Valuation Standards**. London: RICS, 2017.

SALEM, H.; MAZZARA, M. ML-based Telegram bot for real estate price prediction. **Journal of Physics: Conference Series**, v. 1694, n. 1, p. 012010, 1 dez. 2020.

SENI, G.; ELDER, J. F. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. **Synthesis Lectures on Data Mining and Knowledge Discovery**, v. 2, n. 1, p. 1–126, 1 jan. 2010.

SHARMA, N. et al. Real Estate Price's Forecasting Through Predictive Modelling. In: JOSHI, A.; KHOSRAVY, M.; GUPTA, N. (Eds.). . **Machine Learning for Predictive Analysis**. Lecture Notes in Networks and Systems. Singapore: Springer Singapore, 2021. v. 141p. 589–597.

STEINER, M. T. A. et al. Métodos estatísticos multivariados aplicados à engenharia de avaliações. **Gestão & Produção**, v. 15, n. 1, p. 23–32, abr. 2008.

VALIER, A. Who performs better? AVMs vs hedonic models. **Journal of Property Investment & Finance**, 27 mar. 2020.

WANG, F. et al. **House Price Prediction Approach based on Deep Learning and ARIMA Model**. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). **Anais**... In: 2019 IEEE 7TH INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND NETWORK TECHNOLOGY (ICCSNT). out. 2019.

WEINSTOCK, L. R. Introduction to U.S. Economy: Housing Market. p. 3, 2021.

WINARNO, E.; HADIKURNIAWATI, W.; ROSSO, R. N. **Location based service for presence system using haversine method**. 2017 International Conference on Innovative and Creative Information Technology (ICITech). **Anais**... In: 2017 INTERNATIONAL CONFERENCE ON INNOVATIVE AND CREATIVE INFORMATION TECHNOLOGY (ICITECH). nov. 2017.

WU, X. et al. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p. 1–37, 1 jan. 2008.

ZHOU, G. et al. Artificial Neural Networks and the Mass Appraisal of Real Estate. **International Journal of Online and Biomedical Engineering (iJOE)**, v. 14, n. 03, p. 180–187, 30 mar. 2018.

ZURADA, J.; LEVITAN, A. S.; GUAN, J. A comparison of regression and artificial intelligence methods in a mass appraisal context. **Journal of Real Estate Research**, v. 33, n. 3, p. 349–387, 2011.