50 ANOS DE HISTÓRIA E EXCELÊNCIA
XXIII COBREAP
CONGRESSO BRASILEIRO DE ENGENHARIA DE AVALIAÇÕES E PERÍCIAS
17 E 18 PRÉ-CONGRESSO | 18 A 21 CONGRESSO | NOVEMBRO | 2025 | JOÃO PESSOA - PARAÍBA

APLICANDO ALGORITMOS DE MINERAÇÃO DE DADOS EM AVALIAÇÕES DE IMÓVEIS: UM ESTUDO COMPARATIVO

Trabalho de Avaliação

THIAGO CESAR DE OLIVEIRA
LUCIO DE MEDEIROS
DANIEL HENRIQUE MARCO DETZEL

João Pessoa - PB
2025

REALIZAÇÃO:
IBAPE INSTITUTO BRASILEIRO DE AVALIAÇÕES E PERÍCIAS DE ENGENHARIA
IBAPE-PB INSTITUTO BRASILEIRO DE AVALIAÇÕES E PERÍCIAS DE ENGENHARIA

PATROCÍNIO:
CONFEA Conselho Federal de Engenharia e Agronomia
CREA-PB Conselho Regional de Engenharia e Agronomia da Paraíba
mútua Caixa de Assistência dos Profissionais do Crea

APLICANDO ALGORITMOS DE MINERAÇÃO DE DADOS EM AVALIAÇÕES DE IMÓVEIS: UM ESTUDO COMPARATIVO

RESUMO

A avaliação de imóveis tem ganhado relevância como suporte para operações financeiras baseadas no valor desses bens. Entretanto, em bases com grande volume de dados, tem-se observado uma diminuição na capacidade de predição quando são utilizados métodos tradicionais, como a Regressão Linear Múltipla – RLM. Com objetivo de verificar se a aplicação de algoritmos de mineração de dados poderia alcançar resultados estatísticos superiores, foram coletadas bases de dados de avaliação de imóveis de cinco cidades do Estado do Paraná junto à Caixa Econômica Federal. Após validações iniciais, foram geradas bases adicionais com valores reais, transformados e nominais, em versões originais e brutas. Cada uma foi submetida à aplicação de variados algoritmos de mineração de dados (Multilayer Perceptron, Support Vector Regression, K-star, M5Rules, e Random Forest), de forma isolada ou combinada (Regression by Discretization, Bagging e Stacking), com uso de *10-fold cross validation* no software Weka. Foram observados incrementos variados nos resultados estatísticos com o uso dos algoritmos em comparação com os obtidos pela RLM, especialmente quando utilizadas combinações de algoritmos. Os maiores incrementos foram obtidos em bases com maiores volumes de dados e naquelas em que foram realizadas limpezas iniciais mínimas. Os algoritmos foram ranqueados com base no número de resultados superiores obtidos.

Palavras-chave: ***Avaliação de imóveis; Caixa Econômica Federal; Mineração de dados; Algoritmos; Weka***

Observação: artigo publicado em periódico internacional. Em caso de seleção, deverá constar a referência para a publicação original.

SUMÁRIO

1. EXPOSIÇÃO

INTRODUÇÃO

Diversos negócios são realizados diariamente tendo o mercado imobiliário como finalidade principal ou acessória. Por exemplo: compra e venda de imóveis, garantias de empréstimos, financiamentos de projetos, análises de viabilidade econômico-financeira, análises patrimoniais, entre outros. Para que esses negócios aconteçam, é necessário, antes de tudo, saber quanto valem esses imóveis. Determinar os valores de mercado dos imóveis é o principal objetivo das avaliações.

O valor estimado por uma avaliação bem-feita deve estar muito próximo do valor real do imóvel. Isso é essencial para garantir a equidade nas transações imobiliárias, evitando prejuízos financeiros para qualquer das partes. Por exemplo, alguém pode comprar uma casa por um valor superior ao seu real, ou um banco pode ser prejudicado ao aceitar um apartamento como garantia de empréstimo e, ao tentar vendê-lo, descobrir que não recuperará o valor investido. Por isso, é fundamental que as avaliações estimem os valores com precisão e exatidão — o mais próximo possível do valor real.

Infelizmente, não é possível obter todas as informações sobre os imóveis e os fatores que influenciam os valores deles. Também não é viável conhecer toda a população de imóveis existente. Assim, como recomenda a Estatística, os avaliadores trabalham com amostras.

A etapa inicial de uma avaliação é a pesquisa de mercado imobiliário. Com as informações do imóvel a ser avaliado, o avaliador busca dados de ofertas ou vendas de imóveis com características semelhantes. Diversas informações podem ser coletadas a critério do avaliador, como área construída, número de cômodos, vagas de garagem, valores negociados, entre outras. Esses dados formam a base de dados de imóveis similares — a amostra.

A etapa seguinte é a análise dessa base. O principal objetivo da análise amostral é identificar quais características dos imóveis são relevantes para determinar seus valores e quais não são. Com técnicas estatísticas ou de análise de dados, é possível definir pesos para cada característica dos imóveis pesquisados. Esses pesos servirão como referência para estimar o valor dos imóveis. Além disso, os avaliadores costumam buscar dados “ruidosos” — ou seja, dados de imóveis significativamente diferentes da maioria da base. Esses dados são chamados de “outliers”. Pode ser que um imóvel não seja tão semelhante quanto o avaliador imaginou, ou pode ter ocorrido um erro na transcrição dos dados. Um “zero” a mais pode causar grande distorção nos cálculos. Por isso, durante a análise, os avaliadores costumam “limpar” a base, excluindo os dados ruidosos.

A análise resulta em modelos, que são usados para estimar os valores dos imóveis.

No Brasil, a Regressão Linear Múltipla – RLM é uma das técnicas mais aplicadas para análise de bases de dados de imóveis. Apesar de amplamente utilizada, a RLM exige conhecimento e habilidade do avaliador. Diversos pressupostos

básicos devem ser observados para que se possa afirmar que os modelos representam adequadamente a realidade do mercado. Isso significa que os modelos devem obedecer a regras específicas. A Parte 2 da Norma NBR 14653, da ABNT (2011), determina os procedimentos necessários a modelos de avaliações de imóveis com RLM. Por exemplo, exige-se que os erros sejam aleatórios, homocedásticos, não autocorrelacionados e com distribuição normal. Também não pode haver forte correlação entre as variáveis independentes. Outliers devem ser investigados e sua exclusão justificada — um a um. Há ainda outros pressupostos. Buscar transformações numéricas que melhorem os resultados estatísticos também consome tempo. Assim, observa-se empiricamente que análises com RLM no Brasil costumam levar um ou dois dias de trabalho do avaliador, com auxílio de softwares estatísticos.

Acredita-se que há uma relação direta entre o tempo gasto na análise e a complexidade das bases. Grandes bases de dados imobiliários são geralmente heterogêneas — e, portanto, trabalhosas para modelar com técnicas lineares.

Nas últimas décadas, houve um aumento significativo na quantidade de informações disponíveis. Corretores passaram a divulgar imóveis na internet, facilitando a coleta de dados. Empresas com grande presença no mercado, como a Caixa Econômica Federal – CEF, desenvolveram grandes bases de dados. Esperava-se que isso agilizasse as avaliações. No entanto, na maioria dos casos, o tempo para gerar modelos aumentou significativamente. Além disso, observou-se uma aparente redução na capacidade preditiva dos modelos, especialmente os com grandes volumes de dados.

Paralelamente, temas como inteligência artificial, mineração de dados, algoritmos e big data passaram a permitir a extração de informações úteis de grandes bases. Essas técnicas são hoje usadas em buscas de texto, perfis de consumo, reconhecimento de imagens e outras aplicações complexas. No entanto, no Brasil, a RLM ainda parece ser a técnica preferida dos avaliadores.

Dessa forma, o objetivo do presente trabalho é de aplicar algoritmos de mineração de dados em bases com grande volume de dados de avaliações de imóveis e verificar se os resultados são superiores aos obtidos com RLM. Resultados estatísticos superiores levam a avaliações mais precisas e confiáveis — o que é essencial para evitar riscos financeiros. Além disso, o uso de algoritmos pode ajudar os avaliadores a acelerarem o processo de análise em casos de grandes bases de dados.

No primeiro trimestre de 2018, iniciou-se a busca por estudos semelhantes, cruzando palavras-chave de técnicas de análise de dados com termos relacionados à avaliação de imóveis. A literatura revelou ampla cobertura de temas. Big Data foi abordado por HROMADA (2015), RAE e SENER (2016), WU et al. (2016) e CHEN et al. (2016) em análises geográficas; ZHANG et al. (2016) e ÇATAK (2017) utilizaram Extreme Learning Machine (ELM); KOK et al. (2017) utilizaram Automated Valuation Machine (AVM); HAN et al. (2017) e HROMADA (2016) criaram índices e fatores para análise de Big Data.

Outros autores abordaram previsão de preços de imóveis: Shi et al. (2015) com técnicas de clusterização, SHEKARIAN e FALLAHPOUR (2013) e GIUDICE et al.

(2017) com algoritmos genéticos, e YALPIR (2018) com redes neurais artificiais. YEH e HSU (2018) abordaram a criação de fatores de estimativa com raciocínio baseado em casos (CBR). ĆETKOVIĆ et al. (2018) usaram redes neurais para criar índices de avaliação.

Alguns artigos compararam métodos, como MCCLUSKEY et al. (2014), JACKOWSKI (2015), JALALI e LEAKE (2016) e DEL GIUDICE et al. (2017), com novos ou aprimorados algoritmos. Esses estudos foram importantes para entender como comparar resultados entre técnicas. DELL (2013) e WYMAN et al. (2013) apresentaram abordagens focadas em estatística e na complexidade do mercado. MORANO e TAJANI (2014) focaram na identificação precisa de outliers.

Alguns estudos compararam resultados estatísticos de avaliações usando algoritmos de mineração de dados e outras técnicas matemáticas, como KONTRIMAS e VERIKAS (2011), SHI et al. (2015), YEH e HSU (2018), MCCLUSKEY et al. (2014), SHEKARIAN e FALLAHPOUR (2013), YALPIR (2018) e DEL GIUDICE et al. (2017). Todos mostraram superioridade das técnicas não lineares sobre a RLM. Não foram encontrados estudos semelhantes realizados no Brasil à época.

Diante disso, decidiu-se aplicar cinco algoritmos distintos de mineração de dados (Multilayer Perceptron, Support Vector Regression, K-star, M5Rules e Random Forest) e três técnicas de combinação de algoritmos (Regression by Discretization, Bagging e Stacking) em bases de avaliação de imóveis de cinco cidades do estado do Paraná, Brasil. A escolha dos algoritmos e das cidades buscou simular a maior variedade possível de resultados, tornando o método potencialmente replicável em outras regiões. Os resultados estatísticos foram comparados com os da abordagem tradicional de RLM. Todos os algoritmos foram implementados no software Weka.

O restante do artigo está organizado da seguinte forma: a Seção 2 descreve o software utilizado e os modelos coletados, além de apresentar esquemas e detalhes dos experimentos realizados. A Seção 3 apresenta os resultados e discussões. A Seção 4 conclui o artigo.

MATERIAIS E MÉTODOS

Software utilizado e modelos coletados

Após a pesquisa inicial, escolhemos o programa Waikato Environment Knowledge Analysis – Weka (WITTEN et al., 2016) para realizar as atividades de análise de dados deste estudo, pois: 1) possui uma interface gráfica amigável; 2) não requer a instalação de pacotes adicionais; 3) apresenta uma curva de aprendizado curta; e 4) é uma plataforma de código aberto. Assim, trabalhos similares podem ser realizados mesmo por avaliadores com pouco conhecimento em técnicas de mineração de dados. O Weka foi desenvolvido por membros do Departamento de Ciência da Computação da Universidade de Waikato, na Nova Zelândia. É um programa registrado sob a Licença Pública Geral (GNU, em inglês), desenvolvido na linguagem Java. O Weka pode ser descrito como um compilador de algoritmos de mineração de dados, como pré-processadores, classificadores, regressores, de análise de agrupamentos (clustering), regras de associação e seleção de atributos,

entre outros. O programa tem a vantagem de manter a interface gráfica amigável e o formato de apresentação dos resultados, mesmo com a utilização de diferentes algoritmos. Utilizamos a versão 3.8 do software.

Após a seleção da ferramenta, em meados de 2018, coletamos bases de dados de avaliação de imóveis da Caixa Econômica Federal – CEF, um dos cinco maiores bancos do Brasil. À época, a CEF detinha cerca de 70% dos contratos de financiamento habitacional do país. Para cada contrato de financiamento imobiliário é elaborado um laudo de avaliação por um profissional de engenharia ou arquitetura. Esses laudos permitem ao banco conhecer os valores adequados dos imóveis a serem financiados.

Para organizar e controlar os procedimentos técnicos necessários, a empresa contava com uma rede de 72 filiais no país. Cada filial era responsável por desenvolver modelos de Regressão Linear Múltipla (doravante chamados de macromodelos) para avaliação de imóveis das tipologias mais comuns (casas e apartamentos, em geral) em cidades com maior número de contratos — geralmente aquelas com maior população. No Estado do Paraná, existiam, à época, cinco dessas filiais, sediadas nas cidades de Curitiba, Ponta Grossa, Londrina, Maringá e Cascavel.

Com autorização da sede da empresa, iniciamos a coleta dos macromodelos — um de cada filial no Estado do Paraná à época. Solicitou-se às unidades que enviassem macromodelos residenciais de casas (ocupadas ou não), abrangendo cidades com o maior volume possível de dados. Recebemos macromodelos de cinco cidades diferentes, conforme apresentado na Tabela 1.

Tabela 1 – Informações de cidades (IBGE, 2010) e IBGE (2018)

Cidade	**População estimada em 2018**	**Área (km²)**	**Grau de urbanização[1] in 2010**	**Renda média formal em 2016 (salários-mínimos)**	**IDH-M[2] em 2010**
Curitiba	1.917.185	435,036	100,00%	3,9	0,823
Ponta Grossa	348.043	2.054,732	97,79%	2,7	0,763
Londrina	563.943	1.652,569	97,40%	2,8	0,778
Sarandi	95.543	103,463	99,15%	2,3	0,695
Cascavel	324.476	2.100,831	94,36%	2,5	0,782

Como pode ser observado, há diferenças consideráveis entre essas cidades. Há uma capital estadual (Curitiba), três capitais regionais (Ponta Grossa, Londrina e

[1] Grau de urbanização é a relação entre a população urbana e a população total do município.

[2] Índice de Desenvolvimento Humano Municipal – IDH-M é uma medida composta de indicadores de três dimensões de desenvolvimento humano: longevidade, educação e renda. O indicador varia de 0 a 1 – sendo que quanto mais próximo de 1, maior é o nível de desenvolvimento humano.

Cascavel) e uma cidade de porte menor (Sarandi). Como mencionado, buscamos cidades com características diferentes — porém dentro de uma região conhecida pelos autores, no Estado do Paraná.

A pedido da CEF, os nomes das cidades não serão identificados na continuidade do artigo. Elas são apresentadas com números definidos por sorteio. Os macromodelos coletados foram elaborados utilizando programas específicos de avaliação imobiliária — desenvolvidos por funcionários ou ex-funcionários da empresa. Todos esses programas utilizam técnicas de RLM Os modelos coletados foram desenvolvidos com base em RLM. Algumas estatísticas descritivas dos macromodelos recebidos são apresentadas na Tabela 2.

Tabela 2 – Estatísticas descritivas dos macromodelos

Modelo	Total de instâncias	Instâncias utilizadas	Total de atributos	Atributos utilizados
Cidade 1	614	392	12	12
Cidade 2	687	542	39	12
Cidade 3	448	378	15	9
Cidade 4	134	84	18	10
Cidade 5	298	267	13	8

Cada instância refere-se a uma casa específica na cidade. Os atributos utilizados dizem respeito às características dessas casas que foram consideradas em cada modelo. Os atributos mais comuns utilizados nos modelos são: dados descritivos das casas (área construída, área do terreno, padrão construtivo, estado de conservação), localização da casa na cidade (geralmente a distância em relação ao centro), renda média da região e um atributo relacionado à data. Outros atributos utilizados em alguns modelos referem-se às divisões internas das casas (número de quartos, número de banheiros, número de vagas de garagem), infraestrutura do entorno (existência de pavimentação asfáltica, rede de esgoto, rede de drenagem), entre outros.

Na Tabela 2, ao observar a diferença entre o total de instâncias e as instâncias utilizadas, é possível perceber como os desenvolvedores dos modelos “limparam” as bases brutas, removendo os dados ruidosos.

Dando continuidade às estatísticas descritivas, podemos verificar o coeficiente de correlação — um dos principais resultados estatísticos — dos modelos coletados, conforme apresentado na Tabela 3.

Tabela 3 – Resultados estatísticos – coeficiente de correlação – R

Modelo	Equação de regressão	Função estimativa
Cidade 1	0,8576	0,8609
Cidade 2	0,7148	0,7124
Cidade 3	0,7876	0,7890
Cidade 4	0,9490	0,9341
Cidade 5	0,9429	0,9473

É importante destacar que esses resultados foram obtidos a partir dos modelos desenvolvidos após apenas as limpezas iniciais (com base nas colunas “instâncias utilizadas” e “atributos utilizados” da Tabela 2). Não houve aplicação adicional de nenhum método sistemático de separação aleatória de dados — como divisão percentual ou validação cruzada. Esses métodos são conhecidos na literatura de mineração de dados por garantirem que um modelo possa representar a realidade mesmo sem testar seus resultados com dados externos — ou seja, dados que não estão presentes na base original.

Método

O método pode ser descrito em quatro etapas, conforme apresentado de forma ilustrativa no fluxograma da Figura 1:

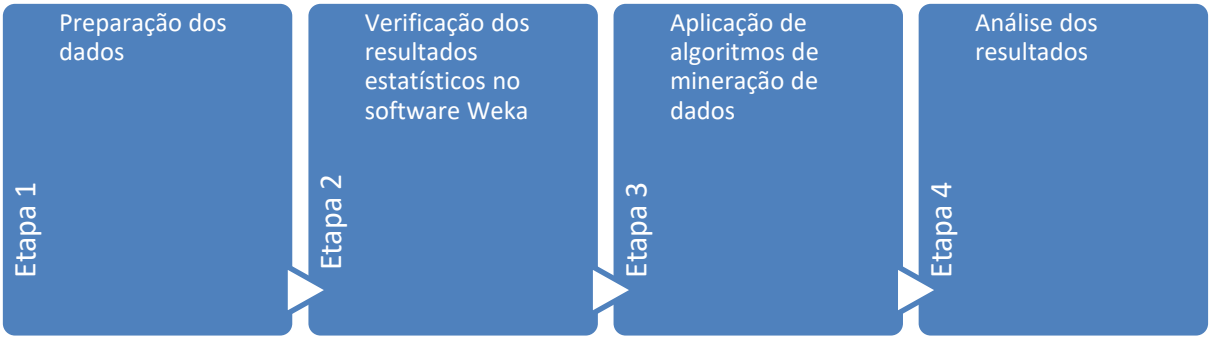


Figura 1 – Fluxograma da metodologia

A primeira etapa consistiu na preparação dos dados dos macromodelos coletados, para possibilitar as verificações e comparações posteriores.

Separamos as bases de dados utilizadas no desenvolvimento dos macromodelos em dois grupos principais: 1) Bases Limpas; e 2) Bases Brutas. As bases limpas possuem a mesma estrutura das bases utilizadas nos macromodelos enviados, com os mesmos dados (ver colunas “instâncias utilizadas” e “atributos utilizados” na Tabela 2). Já as bases brutas são compostas pelas bases originais dos macromodelos, com a reincorporação dos dados das instâncias que foram desconsideradas nas limpezas iniciais (ver “instâncias totais” na Tabela 2). Nas bases

brutas, mantivemos a seleção de atributos dos macromodelos, para que os resultados pudessem ser comparáveis (ver “atributos utilizados” na Tabela 2).

Além disso, optamos por testar algumas situações adicionais nas bases de dados: 1) dados reais e dados transformados; 2) dados numéricos e dados nominais; e 3) dados com transformação logarítmica no atributo alvo (Y). A intenção de criar bases adicionais com valores transformados foi verificar se essas transformações poderiam também proporcionar melhorias nos resultados estatísticos.

Como foram escolhidas cinco cidades, com duas bases e seis tipos de transformações para cada base, um total de 60 bases de dados foi utilizado na análise. Justificamos o uso extensivo de bases para tentar verificar o potencial e as limitações dos algoritmos de mineração de dados na avaliação imobiliária.

Todas as bases de dados foram então convertidas para o formato ARFF (Attribute-Relation File Format) — o formato padrão de arquivos de dados do Weka.

A segunda etapa teve como objetivo verificar se o software Weka era capaz de obter resultados estatísticos consistentes utilizando RLM. Essa etapa foi relevante porque os resultados serviriam como linha de base para todos os testes comparativos posteriores.

Para essa verificação, utilizamos os mesmos dados dos macromodelos (bases limpas, dados transformados, dados numéricos e dados com transformação logarítmica no atributo dependente), pois essas configurações foram utilizadas para obter os resultados estatísticos apresentados na Tabela 3, permitindo a comparação.

Para tentar simular a forma como os resultados da Tabela 3 foram obtidos, configuramos o algoritmo de RLM do Weka para desativar os recursos de robustez (método de seleção de atributos e eliminação de atributos colineares) e desativar as técnicas de separação de dados.

Após aplicar o algoritmo de RLM, observamos que os resultados estatísticos obtidos pelo Weka foram iguais aos descritos na Tabela 3 — com exceção do modelo da Cidade 4. Para esse modelo, o algoritmo do Weka gerou uma equação de regressão diferente da do modelo original, resultando em resultados estatísticos distintos (R original: 0,949 / R do Weka: 0,942). Embora a variação tenha sido pequena, foi realizado um teste t com 95% de confiança. O resultado confirmou que as médias preditivas são iguais com 95% de confiança.

Assim, todos os resultados estatísticos dos macromodelos foram verificados no Weka, o que nos permitiu avançar para a próxima etapa do experimento.

Na terceira etapa, escolhemos os seguintes algoritmos para aplicar às bases de dados coletadas:

1. Multilayer Perceptron – MLP: (WITTEN et al., 2016);
2. Support Vector Regression – SMO: (SHEVADE et al., 1999), (PLATT, 1998) and (SMOLA e SCHSÖLKOPF, 2004);
3. K-star – K*: (CLEARY AND TRIGG, 1995);

4. M5Rules – M5R: (HOLMES et al., 1999), (QUINLAN, 1992) e (WANG e WITTEN, 1997); e
5. Random Forest – RF: (BREIMAN, 2001).

Também escolhemos os seguintes algoritmos que utilizam agrupamento ou conjunto de técnicas (listados no Weka como “meta learners”):

1. Regression by Discretization – Logistic – RBD-L: (FRANK e BOUCKAERT, 2009) and (LE CESSIE e VAN HOUWELINGEN, 1992);
2. Bagging – B-: (BREIMAN, 1994); and
3. Stacking – S-: (WOLPERT, 1992).

Essa lista apresenta: 1) o nome completo do algoritmo; 2) a abreviação utilizada neste trabalho; e 3) referências com detalhes sobre cada um dos métodos. Como decidimos utilizar diversos algoritmos, optamos por não detalhar o funcionamento de cada um ou explicar como realizam seus cálculos, pois isso poderia tornar o texto muito extenso e desviar o foco do leitor do objetivo principal do artigo. No entanto, como demonstram as referências, a maioria desses algoritmos foi desenvolvida na década de 1990. O Multilayer Perceptron é um dos primeiros tipos de redes neurais artificiais desenvolvidos. Portanto, não são novidades na literatura de mineração de dados. Mas acreditamos que a maioria dos avaliadores imobiliários ainda não os utiliza em aplicações práticas.

Os oito algoritmos citados foram escolhidos com base nos seguintes critérios: 1) capacidade de realizar atividades de regressão; 2) presença em artigos analisados na pesquisa inicial; e 3) resultados superiores em análises preliminares. Ao montar a lista, selecionamos pelo menos um algoritmo de cada “classe” oferecida pelo pacote Weka. As principais classes são: 1) Funções (MLP e SMO; 2) Similaridade (K*); 3) Regras (M5R); e 4) Árvores de Decisão (RF). Os demais algoritmos da segunda lista são classificados no Weka como “meta”, ou seja, algoritmos que utilizam outros algoritmos para tentar melhorar a capacidade preditiva.

Os algoritmos foram aplicados no Weka utilizando o módulo Experimenter — disponível no pacote de instalação. Esse módulo permite aplicar diversos algoritmos ou combinações de algoritmos a uma ou mais bases de dados, possibilitando a comparação dos resultados com os de uma referência. Todas as iterações utilizaram *10-fold cross validation*, com 1 repetição (já que a *10-fold cross validation* já realiza 10 iterações), e nível de confiança de 95% para comparação dos resultados. Foram utilizados três resultados estatísticos para as comparações: Coeficiente de Correlação – R, Erro Absoluto Relativo (em inglês: Relative Absolute Error – RAE), e Raiz do Erro Quadrático Relativo (em inglês, Root Relative Squared Error – RRSE), com as equações apresentadas a seguir:

$$R = \sqrt{1 - \frac{(\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2}{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}} \quad (1)$$

$$RAE = \frac{|\hat{y}_1 - y_1| + \cdots + |\hat{y}_n - y_n|}{|y_1 - \bar{y}| + \cdots + |y_n - \bar{y}|} \quad (2)$$

$$RRSE = \sqrt{\frac{(\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2}{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}} \quad (3)$$

onde:

$\hat{y}$ representa o valor estimado para cada instância;

y representa o valor real de cada instância;

$\bar{y}$ representa a média entre os valores reais de todas as instâncias; e

n representa o número total de instâncias.

Os resultados estatísticos obtidos com a aplicação dos algoritmos de mineração de dados foram considerados superiores quando apresentaram desempenho estatisticamente melhor do que os resultados obtidos com o método de referência — a RLM — com nível de confiança de 95%. Para o coeficiente de correlação (R), a relação é positiva (quanto maior, melhor), enquanto para o Erro Absoluto Relativo (RAE) e a Raiz do Erro Quadrático Relativo (RRSE), a relação é negativa (quanto menor, melhor).

Para obter os resultados estatísticos, decidimos utilizar o método *10-fold cross validation*, amplamente recomendado e utilizado na literatura de mineração de dados. Esse método consiste em separar aleatoriamente a base de dados em 10 subconjuntos (folds em inglês). Em seguida, os dados de 9 desses subconjuntos são usados para treinar o algoritmo, e os dados do subconjunto restante são usados para testar. Os resultados estatísticos são obtidos com base nesse subconjunto de teste. O procedimento é repetido, testando com um subconjunto diferente a cada vez, até que todos os 10 subconjuntos tenham sido utilizados como teste. A média entre os 10 resultados obtidos representa o resultado.

Realizamos 10 iterações utilizando *10-fold cross validation*, considerando os dados das bases limpas e brutas das cinco cidades. Para cada iteração, utilizamos 17 algoritmos: seis individuais (RLM, MLP, SMOReg, K*, M5R e RF), um com agrupamento simples (RbD-Logistic), e dez com agrupamentos combinados (cinco com Bagging – B- e cinco com Stacking – S-). Utilizamos abreviações dos algoritmos para facilitar a leitura das informações nas tabelas.

Cada grupo de Bagging utilizou um método individual (exceto RLM) para as combinações. Por exemplo, “B-MLP” refere-se ao uso de Bagging com o algoritmo Multilayer Perceptron. Cada grupo de Stacking utilizou um método individual (também diferente de RLM) como *metalearner*, responsável por selecionar as previsões dos outros métodos combinados. Por exemplo, “S-RF” refere-se ao uso de Stacking com o Random Forest como *metalearner*.

Como foram utilizadas 60 bases de dados, multiplicadas por 17 variações de algoritmos, foram gerados 1.020 resultados. Como foram considerados três indicadores estatísticos (R, RAE e RRSE), o total de resultados obtidos foi de 3.060.

Observamos que o tempo necessário para cada iteração das 17 variações de algoritmos é aparentemente proporcional à quantidade de dados nas bases de cada cidade. As iterações com as bases limpas levaram cerca de uma hora e meia para serem concluídas em um notebook padrão (i5, 8 GB de RAM, SSD – utilizado à época). As iterações com as bases brutas levaram cerca de duas horas e meia. Os algoritmos individuais são razoavelmente rápidos, sendo o Random Forest (RF) o mais demorado. As análises com Stacking (S-) são consideravelmente mais lentas do que as com Bagging (B-) e Regression by Discretization – Logistic (RBD-L).

A quarta etapa do método teve como objetivo analisar todos os resultados estatísticos obtidos, comparar e classificar os algoritmos utilizados. Essa etapa é descrita na próxima seção.

Resultados

Após executar todas as iterações descritas anteriormente, sistematizamos os resultados estatísticos em grandes tabelas, para que pudessem ser analisados posteriormente. O módulo Experimenter do software Weka marca quais resultados são significativamente melhores ou piores do que o método de referência, com base no nível de confiança definido. Os resultados estatísticos da RLM no Weka — também obtidos com *10-fold cross validation* — foram considerados como referência para essas marcações.

Em seguida, para cada cidade e tipo de base (limpa ou bruta), identificamos os algoritmos que apresentaram resultados estatísticos significativamente superiores aos da RLM. Esses resultados foram destacados em amarelo. O melhor resultado estatístico de cada tipo (R, RAE e RRSE) foi destacado em verde. Os resultados de referência da RLM foram destacados em laranja. Esse procedimento está exemplificado na Tabela 4. Foram elaboradas seis dessas tabelas (para bases limpas e brutas – R, RAE e RRSE).

Tabela 4 – Comparação de R – Bases Limpas (exemplo)

CITY 1 DATABASE	MLR	MLP	SMO	K*	M5R	RF	RBD-L	B-MLP	B-SMO	B-K*	B-M5R	B-RF	S-MLP	S-SMO	S-K*	S-M5R	S-RF
real_numerical	0.56	0.74v	0.57	0.67v	0.76v	0.81v	0.70v	0.77v	0.56	0.70v	0.81v	0.81v	0.81v	0.82v	0.76v	0.81v	0.81v
real_nominal	0.56	0.51*	0.55*	0.69v	0.73v	0.79v	0.64v	0.68v	0.57v	0.71v	0.81v	0.77v	0.78v	0.79v	0.71v	0.78v	0.78v
transformed_numerical	0.84	0.78*	0.84*	0.70*	0.84*	0.82*	0.82*	0.83*	0.84*	0.71*	0.85v	0.81*	0.85v	0.85v	0.81*	0.85v	0.84*
transformed_nominal	0.83	0.62*	0.82*	0.70*	0.82*	0.79*	0.75*	0.77*	0.82*	0.71*	0.83*	0.78*	0.81*	0.82*	0.76*	0.83*	0.81*
log Y_numerical	0.57	0.73v	0.57v	0.69v	0.78v	0.81v	0.71v	0.79v	0.57v	0.71v	0.82v	0.81v	0.82v	0.83v	0.77v	0.83v	0.81v
log Y_nominal	0.57	0.54*	0.54*	0.70v	0.76v	0.78v	0.64v	0.71v	0.57v	0.72v	0.80v	0.78v	0.80v	0.81v	0.75v	0.81v	0.81v

CITY 2 DATABASE	RLM	MLP	SMO	K*	M5R	RF	RBD-L	B-MLP	B-SMO	B-K*	B-M5R	B-RF	S-MLP	S-SMO	S-K*	S-M5R	S-RF
real_numerical	0.57	0.46*	0.56*	0.63v	0.63v	0.74v	0.61v	0.61v	0.56*	0.65v	0.69v	0.74v	0.73v	0.75v	0.69v	0.75v	0.73v
real_nominal	0.57	0.43*	0.55*	0.63v	0.63v	0.74v	0.60v	0.58v	0.55*	0.64v	0.70v	0.74v	0.72v	0.75v	0.67v	0.75v	0.72v
transformed_numerical	0.68	0.57*	0.68	0.65*	0.68*	0.74v	0.66*	0.65*	0.68*	0.67*	0.70v	0.74v	0.75v	0.75v	0.69	0.76v	0.72v
transformed_nominal	0.68	0.51*	0.68*	0.64*	0.66*	0.74v	0.64*	0.63*	0.67*	0.66*	0.70v	0.74v	0.75v	0.75v	0.67*	0.75v	0.73v
log Y_numerical	0.57	0.50*	0.57*	0.64v	0.67v	0.75v	0.61v	0.63v	0.57*	0.65v	0.71v	0.74v	0.74v	0.75v	0.68v	0.76v	0.75v
log Y_nominal	0.58	0.44*	0.56*	0.63v	0.66v	0.74v	0.61v	0.61v	0.56*	0.65v	0.70v	0.74v	0.74v	0.75v	0.66v	0.76v	0.73v

CITY 3 DATABASE	RLM	MLP	SMO	K*	M5R	RF	RBD-L	B-MLP	B-SMO	B-K*	B-M5R	B-RF	S-MLP	S-SMO	S-K*	S-M5R	S-RF
real_numerical	0.73	0.69*	0.72*	0.63*	0.76v	0.73	0.75v	0.75v	0.72*	0.65*	0.79v	0.74v	0.78v	0.79v	0.74v	0.78v	0.79v
real_nominal	0.73	0.67*	0.72*	0.62*	0.77v	0.72*	0.74v	0.75v	0.72*	0.63*	0.78v	0.73*	0.77v	0.79v	0.71*	0.78v	0.78v
transformed_numerical	0.78	0.68*	0.76*	0.70*	0.75*	0.74*	0.78*	0.77*	0.77*	0.69*	0.76*	0.74*	0.80v	0.79v	0.73*	0.80v	0.79v
transformed_nominal	0.77	0.66*	0.76*	0.70*	0.74*	0.73*	0.73*	0.74*	0.76*	0.69*	0.77*	0.73*	0.78v	0.79v	0.73*	0.79v	0.78v
log Y_numerical	0.72	0.67*	0.71*	0.64*	0.73v	0.74v	0.72v	0.75v	0.70*	0.65*	0.76v	0.74v	0.76v	0.78v	0.73v	0.77v	0.79v
log Y_nominal	0.71	0.67*	0.70*	0.63*	0.73v	0.72v	0.68*	0.74v	0.70*	0.64*	0.77v	0.72v	0.77v	0.77v	0.67*	0.77v	0.76v

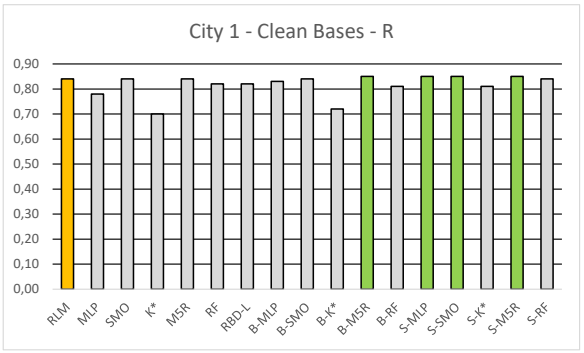
CITY 4 DATABASE	RLM	MLP	SMO	K*	M5R	RF	RBD-L	B-MLP	B-SMO	B-K*	B-M5R	B-RF	S-MLP	S-SMO	S-K*	S-M5R	S-RF
real_numerical	0.88	0.85*	0.87*	0.67*	0.82*	0.87*	0.72*	0.85*	0.88*	0.70*	0.85*	0.87*	0.87*	0.88	0.81*	0.89v	0.89v
real_nominal	0.88	0.85*	0.87*	0.67*	0.82*	0.87*	0.72*	0.85*	0.88*	0.70*	0.85*	0.87*	0.86*	0.88	0.81*	0.88	0.89v
transformed_numerical	0.91	0.87*	0.91	0.12*	0.85*	0.88*	0.74*	0.89*	0.92v	-0.09*	0.85*	0.88*	0.92v	0.91*	0.84*	0.91*	0.88*
transformed_nominal	0.91	0.87*	0.91	0.12*	0.85*	0.88*	0.74*	0.89*	0.92v	-0.09*	0.85*	0.88*	0.91*	0.91*	0.84*	0.91*	0.89*
log Y_numerical	0.88	0.82*	0.88*	0.68*	0.77*	0.86*	0.72*	0.86*	0.89v	0.70*	0.85*	0.86*	0.85*	0.85*	0.81*	0.82*	0.84*
log Y_nominal	0.88	0.82*	0.88*	0.68*	0.77*	0.86*	0.72*	0.86*	0.89v	0.70*	0.85*	0.86*	0.85*	0.85*	0.81*	0.82*	0.85*

CITY 5 DATABASE	RLM	MLP	SMO	K*	M5R	RF	RBD-L	B-MLP	B-SMO	B-K*	B-M5R	B-RF	S-MLP	S-SMO	S-K*	S-M5R	S-RF
real_numerical	0.91	0.92v	0.91v	0.88*	0.92v	0.93v	0.92v	0.94v	0.91*	0.89*	0.94v	0.93v	0.93v	0.93v	0.88*	0.93v	0.92v
real_nominal	0.92	0.89*	0.91*	0.88*	0.92v	0.93v	0.92v	0.92*	0.92*	0.89*	0.93v	0.92v	0.92v	0.93v	0.90*	0.93v	0.92v
transformed_numerical	0.93	0.90*	0.93*	0.86*	0.93*	0.91*	0.92*	0.93*	0.93*	0.87*	0.94v	0.92*	0.93v	0.93*	0.92*	0.93*	0.92*
transformed_nominal	0.93	0.82*	0.93*	0.87*	0.93*	0.91*	0.84*	0.91*	0.93*	0.88*	0.93*	0.91*	0.93*	0.93*	0.91*	0.93*	0.92*
log Y_numerical	0.89	0.90v	0.89*	0.86*	0.91v	0.92v	0.88*	0.92v	0.89*	0.87*	0.93v	0.92v	0.91v	0.92v	0.91v	0.92v	0.91v
log Y_nominal	0.89	0.86*	0.88*	0.87*	0.91v	0.90v	0.80*	0.89*	0.89*	0.88*	0.93v	0.91v	0.91v	0.92v	0.89*	0.91v	0.91v

v - significaly superior result
* - significaly inferior result
considered confidence level - 95%

Em seguida, distribuímos os resultados estatísticos superiores e os melhores em gráficos (à esquerda), separados por cidade. Também calculamos quantas vezes cada algoritmo obteve resultados estatísticos superiores aos da RLM (à direita). Esses resultados estão exemplificados na Tabela 5. Foram elaboradas dez dessas tabelas (para bases limpas e brutas – separadas por cidade).

Tabela 5 – Cidade 1 – Bases Limpas (exemplo)



CLASS.	ALGOR.	QTD.
1	B-M5R	3,00
	S-SMO	
	S-M5R	
2	S-MLP	1,00

Em seguida, calculamos o incremento estatístico obtido pelo algoritmo que apresentou o melhor desempenho (marcado em verde nas tabelas e gráficos anteriores) em cada caso, em comparação com os resultados da RLM (marcados em laranja). Em seguida, compilamos os resultados na Tabela 6. Essa tabela também apresenta os algoritmos recomendados pela análise quantitativa, ou seja, o número de vezes que cada algoritmo foi significativamente superior aos resultados da RLM.

Tabela 6 – Resultados compilados – Bases Limpas

Resultado	Cidade 1		Cidade 2		Cidade 3		Cidade 4		Cidade 5	
	Incr.	Algor.	Incr.	Algor.	Incr.	Algor.	Incr.	Algor.	Incr.	Algor.
R	1,19%	B-M5R S-MLP S-SMO S-M5R	11,76%	S-M5R	2,56%	S-MLP S-M5R	1,10%	B-SMO S-MLP	1,08%	B-MLP B-M5R
RAE	4,46%	S-SMO	14,46%	S-SMO	8,59%	S-SMO	-	-	6,51%	B-M5R
RRSE	2,57%	S-M5R	10,02%	S-M5R	4,78%	S-M5R	-	-	1,83%	B-M5R
RECOM.	S-SMO		S-M5R		S-M5R		RLM		B-M5R	

A Tabela 7 apresenta a compilação dos incrementos nos resultados estatísticos obtidos na análise das bases brutas, bem como os algoritmos recomendados pela análise quantitativa.

Tabela 7 – Resultados compilados – Bases Brutas

Resultado	Cidade 1		Cidade 2		Cidade 3		Cidade 4		Cidade 5	
	Incr.	Algor.	Incr.	Algor.	Incr.	Algor.	Incr.	Algor.	Incr.	Algor.
R	30,19%	S-SMO	47,06%	S-RF	4,55%	S-SMO	-	-	2,20%	RF S-SMO S-RF
RAE	24,63%	S-SMO	51,31%	B-K*	7,34%	RF	7,72%	S-SMO	11,69%	B-RF
RRSE	19,04%	S-SMO	38,17%	RBD-L	4,04%	S-SMO	-	-	8,38%	S-SMO
RECOM.	S-SMO		B-SMO		S-SMO		RML		S-SMO	

Como observado, apenas nas bases de dados da Cidade 4 a aplicação dos algoritmos de mineração de dados não obteve resultados superiores aos obtidos com a RLM. Em todas as demais bases, foram observados incrementos em relação aos valores de referência. Os maiores incrementos ocorreram quando os algoritmos foram aplicados às bases brutas, em comparação com as bases limpas.

A aplicação dos algoritmos resultou em pequenos incrementos nas Cidades 3 e 5, tanto nas bases limpas quanto nas brutas. Na Cidade 1, os incrementos foram pequenos nas bases limpas e maiores nas bases brutas. Já na Cidade 2, os incrementos foram mais expressivos em ambas as bases, sendo muito mais elevados nas bases brutas.

Para identificar a origem dessas diferenças, observamos que a quantidade de dados desconsiderados ou excluídos das bases brutas (para formar as bases limpas) pode estar relacionada aos incrementos observados. A Tabela 8 apresenta essa relação por cidade.

Tabela 8 – Quantidades de dados nas bases

Informação	Cidade 1	Cidade 2	Cidade 3	Cidade 4	Cidade 5
Instancias de Bases Brutas	614	687	448	134	298
Instâncias de Bases Limpas	392	542	378	84	267
“Limpeza” em quantidade	222	145	70	50	31
“Limpeza” em percentual	36,16%	21,11%	15,63%	37,31%	10,40%

Assim, observamos que os maiores incrementos nos resultados após a aplicação dos algoritmos de mineração de dados foram verificados em bases de dados de cidades com: 1) maior quantidade de dados nas bases brutas, e 2) menor nível de “limpeza” na quantidade de dados para formar as bases limpas (aquelas efetivamente modeladas).

Na Cidade 4, mesmo com um alto percentual de limpeza, nenhum incremento foi obtido com a aplicação dos algoritmos de mineração de dados. Isso sugere que pode ser necessário atender aos dois critérios mencionados acima para possibilitar o aumento dos resultados estatísticos com a aplicação dos algoritmos.

Além disso, o desempenho inferior dos algoritmos em algumas bases de dados pode estar relacionado à existência de mercados imobiliários naturalmente “linearizados”. Ou seja, mercados em que as relações lineares entre os atributos dos imóveis explicam muito bem o comportamento do mercado. Esse comportamento peculiar pode ocorrer em mercados imobiliários menos complexos.

A Tabela 9 apresenta os algoritmos mais bem classificados, considerando a análise das bases limpas das cinco cidades, calculados com base na quantidade de resultados estatísticos superiores aos de referência (RLM).

Tabela 9 – Ranking de melhores algoritmos – Bases Limpas

RANK.	ALGOR.	QTY.
1	B-M5R	12
2	S-SMO	10
	S-M5R	
3	S-RF	7
4	RF	5
	B-RF	
	S-MLP	
5	B-MLP	3
6	S-K*	2
7	SMO	1
	K*	
	M5R	
	B-SMO	
	B-K*	
8	MLP	0
	RBD-L	

A Tabela 10 apresenta os algoritmos mais bem classificados, considerando a análise das bases brutas das cinco cidades, calculados com base na quantidade de resultados estatísticos superiores aos de referência (RLM).

Tabela 10 – Ranking dos melhores algoritmos – Bases Brutas

RANK.	ALGOR.	QTY.
1	S-SMO	13,00
2	RF	12,00
	S-M5R	
3	S-RF	11,00
4	B-RF	10,00
5	B-M5R	9,00
6	SMO	8,00
	M5R	
	B-SMO	
	S-K*	
7	S-MLP	6,00
8	K*	5,00
	B-K*	
9	RBD-L	2,00
	B-MLP	
10	MLP	0,00

É possível perceber, na análise dessas duas tabelas, que os algoritmos B-M5R e S-SMO obtiveram as maiores quantidades de resultados estatísticos superiores nas aplicações dos algoritmos de mineração de dados nas bases limpas e brutas, respectivamente. No entanto, ao somar as quantidades entre as duas bases, o S-SMO

foi o método que apresentou mais resultados estatísticos superiores em comparação com os da referência (RLM).

O melhor método isolado classificado foi o Random Forest (RF). Mais uma vez, observou-se que os métodos combinados obtiveram resultados estatísticos superiores com mais frequência do que os métodos isolados.

O algoritmo MLP não obteve nenhum resultado superior em relação aos da referência, nem nas bases limpas, nem nas brutas.

2. CONCLUSÕES

Demonstramos com o presente trabalho que a aplicação de algoritmos de mineração de dados permite obter resultados estatísticos superiores aos obtidos com o uso da RLM em bases de dados de avaliação de imóveis. Resultados estatísticos superiores refletem estimativas mais precisas e corretas dos valores dos imóveis, o que possibilita a redução dos riscos em transações financeiras imobiliárias.

Também verificamos que a possibilidade de obter resultados superiores aumenta, na maioria dos casos, com o uso de combinações de algoritmos — com uso de métodos como Bagging e Stacking.

Neste estudo, optamos por variar tanto o número de algoritmos aplicados quanto as próprias bases de dados. Mostramos que a variedade, inicialmente considerada o principal problema das grandes bases de dados imobiliárias, também se apresentou como a principal solução. Os resultados também variaram, o que indica que os mercados imobiliários das cidades analisadas possuem dinâmicas diferentes. De modo geral, observamos que os maiores incrementos com a aplicação dos algoritmos de mineração de dados ocorreram em bases com mais dados e com menos limpeza.

Na análise de dados tradicional, há uma prática estatística de tentar eliminar quaisquer inconsistências das bases. Algumas técnicas podem ser úteis para localizar possíveis erros — a fim de garantir a confiabilidade dos dados. No entanto, a busca por melhores resultados estatísticos pode muitas vezes levar a uma limpeza excessiva. Outliers podem ou não ser ruídos. Algumas informações, embora verdadeiras, podem simplesmente ser diferentes da “massa” de dados e, por isso, acabam sendo excluídas da modelagem. Especialmente em avaliações de imóveis, muitas vezes as diferenças observadas não são inconsistências — elas podem simplesmente refletir o nível de complexidade do mercado imobiliário analisado, o que pode exigir métodos não lineares para a análise dos dados. E é justamente aí que acreditamos que os algoritmos de mineração de dados superam os métodos tradicionais, como a RLM.

O tempo necessário para a modelagem também foi considerado neste trabalho. O treinamento e teste dos modelos levaram várias horas em um notebook padrão à época. Pode-se argumentar que um avaliador experiente poderia gastar o mesmo tempo para obter resultados semelhantes. No entanto, o tempo de processamento pode ser escalado em computadores mais potentes, em clusters ou em ambientes de computação distribuída (distribuição horizontal). Algumas tecnologias mais recentes podem realmente reduzir o tempo de processamento necessário. Grandes empresas de tecnologia utilizam processamento em nuvem em suas atividades de mineração de dados, o que lhes permite extrair informações úteis de grandes bases. O tempo de processamento pode ser escalado vertical ou horizontalmente para reduzir o tempo necessário para obter informações. Mas as habilidades, experiência e sensibilidade de um avaliador não podem ser escaladas.

É importante ressaltar que os resultados obtidos neste estudo se referem a situações específicas em uma determinada data — o que pode não refletir o mesmo

desempenho em outras situações. Portanto, análises experimentais devem sempre ser realizadas. O software Weka é uma ferramenta recomendada para esse fim, pois oferece diversas soluções computacionais para análise de bases de dados, testes e comparações de algoritmos, com o objetivo de extrair informações úteis.

A desvantagem da aplicação de algoritmos de mineração de dados na avaliação imobiliária é a possível dificuldade de interpretação. A RLM, embora trabalhosa, é de fácil explicação.

Conforme demonstrado neste trabalho, os resultados estatísticos das avaliações imobiliárias podem ser aprimorados com o uso de algoritmos de mineração de dados, especialmente em bases grandes e complexas — como as dos diversos mercados habitacionais. Isso os torna interessantes para atividades de avaliação de imóveis em um mundo com alta disponibilidade de informações. Com o uso de tecnologias mais recentes, como a distribuição horizontal, acreditamos que o tempo de processamento necessário para gerar modelos de avaliação de imóveis a partir de grandes bases de dados pode ser significativamente reduzido — mantendo, ainda assim, a precisão e exatidão necessárias.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS – ABNT. (2011), **NBR 14653-2 - Avaliação de Bens - Parte 2 : Imóveis Urbanos**, *Abnt*, Brasil.

BREIMAN, L. (1994), **Bagging Predictors**, *Machine Learning, Technical Report No. 421*, No. 2, p. 19.

BREIMAN, L. (2001), **Random Forests**, *Machine Learning*, Kluwer Academic Publishers Hingham, Vol. 45 No. 1, pp. 5–32.

ÇATAK, F.Ö. (2017), **Classification with boosting of extreme learning machine over arbitrarily partitioned data**, *Soft Computing*, Springer Berlin Heidelberg, Vol. 21 No. 9, pp. 2269–2281.

LE CESSIE, S. e VAN HOUWELINGEN, J.C. (1992), **Ridge Estimators in Logistic Regression**, *Applied Statistics*, Vol. 41 No. 1, pp. 191–201.

ĆETKOVIĆ, J., LAKIĆ, S., LAZAREVSKA, M., ŽARKOVIĆ, M., VUJOŠEVIĆ, S., CVIJOVIĆ, J. e GOGIĆ, M. (2018), **Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application**, *Complexity*, Vol. 2018, p. 10.

CHEN, Y., LIU, X., LI, X., LIU, Y. e XU, X. (2016), **Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning**, *Applied Geography*, Elsevier Ltd, Vol. 75, pp. 200–212.

CLEARY, J.G. e TRIGG, L.E. (1995), **K*: An Instance-based Learner Using an Entropic Distance Measure**, *12th International Conference on Machine Learning*, pp. 108–114.

DELL, G. (2013), **Common statistical errors and mistakes valuation and reliability**, *The Appraisal Journal*, pp. 332–347.

FRANK, E. e BOUCKAERT, R.R. (2009), **Conditional Density Estimation with Class Probability Estimators**, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Berlim, Vol. 5828 LNAI, pp. 65–81.

DEL GIUDICE, V., DE PAOLA, P. e CANTISANI, G.B. (2017), **Rough Set Theory for Real Estate Appraisals: An Application to Directional District of Naples**, *Buildings*, Vol. 7 No. 1, available at:https://doi.org/10.3390/buildings7010012.

GIUDICE, V. DEL, PAOLA, P. DE e FORTE, F. (2017), **Using Genetic Algorithms for Real Estate Appraisals**, pp. 1–12.

HAN, S., KO, Y., KIM, S. e SHIN, D.H. (2017), **Home sales index prediction model based on cluster and principal component statistical approaches in a big data analytic concept**, *KSCE Journal of Civil Engineering*, Vol. 21 No. 1, pp. 67–75.

HOLMES, G., HALL, M. e FRANK, E. (1999), **Genereting Rule Sets from Model Trees**, *Twelfth Australian Joint Conference on Artificial Intelligence*, pp. 1–12.

HROMADA, E. (2015), **Mapping of real estate prices using data mining techniques**, *Procedia Engineering*, Elsevier B.V., Vol. 123, pp. 233–240.

HROMADA, E. (2016), **Real estate valuation using data mining software**, *Procedia Engineering*, The Author(s), Vol. 164 No. June, pp. 284–291.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE (2018), **Estimativas Da População Residente Com Data de Referência 1º de Julho de 2018**, disponível em: https://cidades.ibge.gov.br.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE (2010), **Censo Demográfico 2010**, disponível em: https://cidades.ibge.gov.br.

JACKOWSKI, K. (2015), **Adaptive Splitting and Selection Algorithm for Regression**, Vol. 9, pp. 425–448.

JALALI, V. e LEAKE, D. (2016), **Enhancing case-based regression with automatically-generated ensembles of adaptations**, *Journal of Intelligent Information Systems*, Vol. 46 No. 2, pp. 237–258.

KOK, N., KOPONEN, E.-L. e MARTÍNEZ-BARBOSA, C.A. (2017), **Big data in real estate? from manual appraisal to automated valuation**, *Journal of Portfolio Management*, Vol. 43 No. 6, pp. 202–211.

KONTRIMAS, V. e VERIKAS, A. (2011), **The mass appraisal of the real estate by computational intelligence**, *Applied Soft Computing Journal*, Vol. 11 No. 1, pp. 443–448.

MAKRIDAKIS, S., WHEELWRIGHT, S.C. e HYNDMAN, R.J. (1998), **Forecasting: Methods and Applications**, 3rd edition, John Wiley & Sons.

MCCLUSKEY, W.J., DAUD, D.Z. e KAMARUDIN, N. (2014), **Boosted regression trees**, *Journal of Financial Management of Property and Construction*, Vol. 19 No. 2, pp. 152–167.

MORANO, P. e TAJANI, F. (2014), **Least median of squares regression and minimum volume ellipsoid estimator for outliers detection in housing appraisal**, *Int. J. Business Intelligence and Data Mining*, Vol. 9 No. 2, pp. 91–111.

PLATT, J.C. (1998), **Fast Training of Support Vector Machines using Sequential Minimal Optimization**, edited by Schoelkopf, B., Burges, C. and Smola, A.*Advances in Kernel Methods - Support Vector Learning*, MIT Press, p. 25.

QUINLAN, R.J. (1992), **Learning with Continuous Classes**, *5th Australian Joint Conference on Artificial Intelligence*, pp. 343–348.

RAE, A. e SENER, E. (2016), **How website users segment a city: The geography of housing search in London**, *Cities*, The Authors, Vol. 52, pp. 140–147.

SHEKARIAN, E. e FALLAHPOUR, A. (2013), **Predicting house price via gene expression programming**, *International Journal of Housing Markets and Analysis*, Vol. 6, pp. 250–268.

SHEVADE, S.K., KEERTHI, S.S., BHATTACHARYYA, C. e MURTHY, K.R.K. (1999), **Improvements to the SMO Algorithm for SVM Regression**, *IEEE Transactions on Neural Networks*.

SHI, D., GUAN, J., ZURADA, J. e LEVITAN, A.S. (2015), **An Innovative Clustering**

Approach to Market Segmentation for Improved Price Prediction, *Journal of International Technology & Information Management*, Vol. 24 No. 1, pp. 15–32.

SMOLA, A.J. e SCHSÖLKOPF, B. (2004), **A tutorial on support vector regression**, *Statistics and Computing*, No. 14, pp. 199–222.

WANG, Y. e WITTEN, I.H. (1997), **Induction of model trees for predicting continuous classes**, *Poster Papers of the 9th European Conference on Machine Learning*.

WITTEN, I.H., FRANK, E., HALL, M.A. e PAL, C.J. (2016), **Data Mining Practical Machine Learning Tools And Techniques**, *Morgan Kaufmann Series in Data Management Systems*, 4th ed., Morgan Kaufmann.

WOLPERT, D.H. (1992), **Stacked Generalization**, *Neural Networks*, Elsevier, Vol. 5 No. 2, pp. 241–259.

WU, C., YE, X., REN, F., WAN, Y., NING, P. e DU, Q. (2016), **Spatial and Social Media Data Analytics of Housing Prices in Shenzhen, China**, *PLoS ONE*, Vol. 11 No. 10, p. e0164553.

WYMAN, D., WORZALA, E. e SELDIN, M. (2013), **Hidden complexity in housing markets: A case for alternative models and techniques**, *International Journal of Housing Markets and Analysis*, Vol. 6 No. 4, pp. 383–404.

YALPIR, Ş. (2018), **Enhancement of parcel valuation with adaptive artificial neural network modeling**, *Artificial Intelligence Review*, Vol. 49 No. 3, pp. 393–405.

YEH, I.-C. e HSU, T.-K. (2018), **Building real estate valuation models with comparative approach through case-based reasoning**, *Applied Soft Computing Journal*, Elsevier B.V., Vol. 65, pp. 260–271.

ZHANG, N., CHEN, H., CHEN, X. e CHEN, J. (2016), **ELM Meets Urban Big Data Analysis: Case Studies**, *Computational Intelligence and Neuroscience*, Vol. 2016, p. 10.