Aplicação de Algoritmos de Aprendizado de Máquina à Previsão do Valor de Imóveis no Norte de Minas Gerais

Bruno Henrique de Pádua Silva
Instituto de Engenharia, Ciência e Tecnologia
Universidade Federal dos Vales do Jequitinhonha e Mucuri
Janaúba, Brasil
henrique.bruno@ufvjm.edu.br

Rogério Alves Santana
Faculdade de Ciências Sociais Aplicadas e Exatas
Teófilo Otoni, Brasil
rogerio.santana@ufvjm.edu.br

Honovan Paz Rocha
Instituto de Engenharia, Ciência e Tecnologia
Universidade Federal dos Vales do Jequitinhonha e Mucuri
Janaúba, Brasil
honovan.rocha@ufvjm.edu.br

***Resumo*—The Brazilian real estate market has grown much more than the country's economy in recent years, and the forecast is that it will still last, as the estimated growth for the construction industry is 2.5% for the year 2023 It is expected that purchases and sales will continue to grow significantly, because the real estate market is a very solid segment in Brazil, where properties are seen as a store of value, even more so when we talk about residential properties. The forecast for the real estate market in 2023 is positive and promising, as projections indicate that growth in the sector should continue, with a good return on investments. Considering this new reality, planning the property acquisition process has become increasingly important, bringing attention to prediction techniques as tools to support the choice of a new asset. Real estate price prediction is a complex subject and depends on several factors that can influence the value of properties, such as the Selic rate, demand for real estate credit, the supply, and location of properties and market conditions. In this work, we built two property databases in the macro-region of Northern Minas Gerais, and carried out processes to obtain structural bases. We use classic statistical methods, such as Linear and Polynomial Regression, a Perceptron and an Artificial Neural Network (ANN) of the Multilayer Perceptron type to predict property values on the constructed bases. The proposed methods showed promising results with errors close to zero in the bases used.**

***Keywords*—Real Estate Price Forecasting, Machine Learning, Linear Regression, Polynomial Regression, Perceptron, Multilayer Perceptron, North of Minas Gerais.**

***Resumo*—O mercado imobiliário brasileiro tem crescido muito mais que a economia do país nos últimos anos, e a previsão é que ele ainda perdure, pois, a estimativa de crescimento para indústria do setor de construção é de 2,5% para o ano de 2023. É esperado que as compras e vendas continuem crescendo bastante, isso porque o mercado imobiliário é um segmento bastante sólido no Brasil, onde imóveis são vistos como uma reserva de valor, ainda mais quando falamos de imóveis residenciais. A previsão para o mercado imobiliário em 2023 é positiva e promissora, como projeções indicam que o crescimento no setor deve permanecer, com bom retorno para investimentos. Considerando-se esta nova realidade, o planejamento do processo de aquisição de um imóvel tem ganhado cada vez mais importância, trazendo atenção para as técnicas de predição como ferramentas para suporte à escolha de um novo bem. A predição do preço de imóveis é um assunto complexo e depende de vários fatores que podem influenciar o valor dos imóveis, como a taxa Selic, a demanda por crédito imobiliário, a oferta e a localização das propriedades e as condições do mercado. Nesse trabalho, construímos duas bases de dados de imóveis da macrorregião do Norte de Minas Gerais, e realizamos processos para obtenção de bases estruturais. Utilizamos métodos clássicos da estatística, como a Regressão Linear e Polinomial, um Perceptron e uma Rede Neural Artificial (RNA) do tipo Perceptron Multicamadas para prever os valores de imóveis nas bases construídas. Os métodos propostos apresentaram resultados promissores com erros próximos de zero nas bases utilizadas.**

***Palavras-chave*—Previsão do preço de imobiliário, Aprendizado de máquina, Regressão Linear, Regressão Polinomial, Perceptron, Perceptron Multicamadas, Norte de Minas Gerais.**

I. INTRODUÇÃO

O mercado imobiliário consiste em um setor da economia onde ocorrem negociações de imóveis, de terrenos ou qualquer construção sobre esses terrenos. No momento em que participamos dessas negociações, a fim de comprar, vender ou alugar um desses bens, estamos participando de um negócio no mercado imobiliário. Essas negociações podem ser feitas tanto diretamente com o proprietário ou com uma imobiliária, e ainda se utiliza de várias formas de pagamento nesse seguimento, sendo o crédito via instituição financeira, a forma mais comum.

A previsão e avaliação do preço dos imóveis sempre foi motivo de estudos, pois esse setor é um reflexo da economia do país. Logo, a previsão dos preços dos imóveis é importante

tanto para as famílias brasileiras na hora de tomar a decisão de investir na aquisição de um imóvel, quanto para economistas analisarem a situação econômica de uma região, cidade ou país. Ele visa auxiliar as pessoas na tomada de decisão em relação ao preço justo e localização de imóveis, que resultará em um preço desejado no futuro. Muitas vezes essas transações financeiras de imóveis são marcadas por assimetria de informação entre compradores, intermediários e vendedores. Por isso, existe a necessidade de coleta e processamento de informações para atenuar essa assimetria, muitas vezes a coleta dessas informações pode ser dispendiosa e ineficiente.

O objetivo principal deste trabalho é construir uma proposta de uso de aprendizado de máquina para prever preços de imóveis na região Norte de Minas Gerais, porém não existe um banco de dados disponível que possa ser utilizado para estimar preços de imóveis para esta região. No entanto, podemos usar a enorme quantidade de informações públicas e gratuitas na Internet, como dados de sites imobiliários, para coletá-los por meio de *Web scraping*. De forma geral, é possível gerar um banco de dados de imóveis estruturado que represente bem a região do Norte de Minas Gerais, de maneira que seja possível a previsão valores para novos imóveis pesquisados. Uma das vantagens dessa proposta é encurtar o tempo de previsão de preço, maximizar o retorno do investimento em imóveis, evitar prejuízos e alocação indevida dos recursos disponíveis.

Nesse trabalho, construímos duas bases de dados de imóveis da macro região norte de Minas Gerais, e realizamos processos para obtenção de bases estruturais. Em seguida, empregamos 4 métodos clássicos de aprendizado de máquina, Regressão Linear, Regressão Polinomial, Perceptron e Rede Neural Artificial (RNA) do tipo Perceptron Multicamadas, para prever com precisão valores de imóveis na região do estado do Norte de Minas Gerais (MG). Os métodos propostos apresentaram resultados promissores com baixos valores para o erro médio quadrático nas bases utilizadas.

A organização do trabalho é descrita a seguir. A Seção II apresenta alguns trabalhos relacionados relevantes ao problema de predição de preços de imóveis. Na Seção III abordamos assuntos sobre os principais conceitos de aprendizagem aplicado à predição do valor de imóveis, trabalhos relacionados, coleta de dados, pré-processamento de dados e avaliações de modelo. Na Seção IV apresentamos a metodologia utilizada para montagem das bases de dados e aplicação dos modelos de aprendizado de máquina. A Seção V apresenta os resultados, análises descritivas e discussões dos experimentos realizados. Por fim, a Seção VI apresenta a conclusão deste trabalho.

II. TRABALHOS RELACIONADOS

A. *Regressão Linear Múltipla*

O método de avaliação e previsão de preços de imóveis por meio de regressão linear múltipla envolve a estimação de parâmetros que representam a conduta do mercado imobiliário. Esses parâmetros são as variáveis independentes que se associam com a variável dependente, o preço do imóvel.

O trabalho de Dantas (1988) é conhecido como uns dos primeiros trabalhos com aplicação de métodos de regressão linear para estudar os preços dos imóveis. Utilizando como variável dependente, o valor unitário à vista do imóvel em metro quadrado, e como variáveis independentes, as características desse imóvel como a localização, o número de quartos e outros atributos. Porém, esse protótipo não fornece um ponto de vista do mercado e de outros fatores sobre o valor de um imóvel.

González (1997) analisou a evolução dos preços de aluguéis na cidade de Porto Alegre, de antes do Plano Real até 1996. Por fim, foram coletados dados de 980 apartamentos residenciais diretamente do mercado e dados de pesquisas mensais do Sindicato de imobiliárias de Porto Alegre. Não se tem confirmação de que a transação ocorreu de fato. Os dados foram analisados por meio de modelos hedônicos de preços, onde foi possível realizar algumas previsões.

No segmento da Regressão Linear, Selim (2008) apurou as principais propriedades que influenciam os preços de imóveis e de aluguéis na Turquia no modelo de preços hedônicos. Os principais atributos que influenciam os preços dos aluguéis são o tipo de casa, tipo de prédio, número de quartos, tamanho, e outras características estruturais como sistema de água, piscina e sistema de gás, em uma amostra com 5.741 observações e 46 atributos, além disso, foi empregado a forma funcional semi-logarítmica.

Araújo et al. (2012) utilizaram regressão múltipla na avaliação de imóveis residenciais urbanos na cidade de Bonito, no estado do Mato Grosso do Sul. Eles utilizaram 27 observações de imóveis residenciais obtidas em sítios de imobiliárias da cidade de Bonito. A variável de localização assumiu o valor 0 para imóveis fora do centro e 1 para imóveis situados no centro. Os critérios utilizados para seleção das variáveis independentes foram Seleção Progressiva e Seleção Regressiva.

B. *Redes Neurais Artificiais*

Com o avanço dos trabalhos em avaliação e previsão de preços no mercado imobiliário, a utilização de técnicas de redes neurais artificiais começa a se tornar habitual.

Baptistella, Steiner e Chaves Neto (2005) estimaram os preços de imóveis urbanos na cidade de Guarapuava, no estado do Paraná, empregando redes neurais artificiais. Na RNA de múltiplas camadas eles utilizaram o algoritmo *Back-Propagation*

com o método de gradiente descendente. As Componentes Principais foram utilizadas para moderar os números de variáveis visando favorecer a execução, com o uso dessa técnica, o número de variáveis passou de 13 para 9 atributos. Além disso, a desempenho foi avaliada utilizando o erro quadrático médio[1] e a raiz quadrada do erro quadrático médio[2] no cálculo do erro da rede.

Moreira, Silva e Fernandes (2010) apresentaram uma nova conduta fundamentada na utilização de Redes Neurais Artificiais e do método de Tomada de Decisão Interativa e Multicritério. O trabalho desenvolveu uma avaliação na qual, inicialmente, é realizada uma carga de dados no sistema, depois é aplicado o método TODIM na base de dados nas alternativas, e, em seguida, as alternativas são ordenadas e um valor médio de aluguel para cada alternativa é calculado. A camada de saída possuía 4 neurônios capazes de representar a quantidade de faixas de valores de aluguéis que o resultado disponibiliza.

H. Lin e Chen (2011) apresentam um estudo de previsão de preço de imóveis em Taiwan, operando redes neurais artificiais e regressão de suporte vetorial[3]. Depois da coleta dos dados, calcularam-se as médias mensais dos preços dos imóveis. A Rede neural usou a técnica de Feedforward com *Backpropagation*. Muitas previsões dos preços dos imóveis realizadas pelo modelo SVR foram melhores do que a das redes neurais. No estudo, eles utilizaram as variáveis: taxa de redesconto, fornecimento de dinheiro e preço do mês anterior, tanto na RNA quanto SVR.

No artigo de H.-Y. Lin e KUENTAI (2015) eles tratam sobre a predisposição do preço unitário médio de imóveis na cidade de Taipei, capital de Taiwan, visando estabelecer um modelo de previsão de preços e os fatores-chave que definem esse preço. Em seu trabalho eles utilizaram dados do portal imobiliário de Taiwan. E para melhorar a eficiência do modelo eles utilizaram 12 variáveis de indicadores imobiliários e econômicos são utilizadas. Para o treinamento das redes neurais foi utilizado a biblioteca NN no Matlab 2008.

III. REFERENCIAL TEÓRICO

A. Aprendizado Supervisionado

Considere um conjunto de treinamento de N pares de exemplos de entrada e saída do tipo $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$, onde cada y_j foi gerado por uma função desconhecida $y_i = f$, O objetivo do aprendizado supervisionado é encontrar uma função h que se aproxime da função verdadeira f. Para medir a precisão de uma hipótese, fornecemos um *conjunto de testes* de exemplos que são distintos do conjunto de treinamento. Tecnicamente, a solução de um problema de *regressão* é encontrar uma expectativa condicional ou valor médio de y porque a probabilidade de acharmos exatamente o número de valor real certo para y tender a zero Russell (2010).

B. Regressão Linear

A regressão é uma técnica estatística que pode ser utilizada para prever valores de imóveis. O aprendizado de máquina pode ser aplicado em vários trabalhos para prever o valor aproximado de um imóvel com base em suas características. O processo envolve a realização de prognósticos por meio de técnicas de aprendizado de máquina, partindo-se de uma série de valores existentes obtidos de dados históricos, bem como de suposições controladas a respeito das condições futuras, para prever outros valores. Para se estimar o valor esperado, usa-se da Equação 1, que determina a relação entra (ambas) as variáveis:

$$y = \beta_0 + x\beta_1 + \epsilon \tag{1}$$

A Equação 1, pode ser reescrita em forma de matriz (Equação 2), onde $\mathbf{y}$ é um vetor de l observações, $\mathbf{X}$ é uma matriz de tamanho $l \times (c+1)$ (a primeira coluna com valores sendo $= 1$, representado a contante α, e c é a quantidade de variáveis explicativas), $\boldsymbol{\beta}$ é um vetor de $(c+1)$ variáveis explicativas e $\boldsymbol{\epsilon}$ é uma matriz de l de resíduos.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

O vetor $\hat{\boldsymbol{\beta}}$ de coeficientes da regressão Linear estimados (usando estimativa de método dos mínimos quadrados) $\boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-\mathbf{1}} \mathbf{X}^\top \mathbf{y}$, assumindo $m < n$ ser necessário para a matriz ser invertível; então, como é uma matriz de *Vandermonde*, a condição de invertibilidade é garantida para se manter se todos os valores forem distintos Russell (2010).

C. Regressão Polinomial

A regressão polinomial é uma forma de regressão linear na qual a relação entre a variável independente x e a variável dependente y é modelada como um polinômio de n-ésimo grau e se ajusta a uma relação não linear entre o valor de x e a média condicional correspondente de y, denotada por $P(y|x)$. O modelo de regressão polinomial e dado por:

$$y = \beta_0 + x\beta_1 + x^2\beta_2 + \cdots + x^g\beta_g + \epsilon \tag{3}$$

A Equação 3 também pode ser escrita na forma matricial da equação 2.

[1]MSE — *Mean Squared Error*
[2]RMSE — *Root Mean Squared Error*
[3]SVR — *Support Vector Regression*

D. Perceptron

O Perceptron é um algoritmo para aprender um classificador binário chamada função de limiar: uma função que mapeia sua entrada $\mathbf{X}$ (um vetor de valor real) para um valor de saída y (um único valor entre 0 e 1). O qual é dado por:

$$\mathbf{y} = \text{ativação}(\boldsymbol{\theta} + \mathbf{XW}) \tag{4}$$

onde $\mathbf{W}$ é o vetor de pesos, $\mathbf{XW}$ é o produto de $\sum_{i=1}^{l} x_i w_i$, onde l é o número de entradas para o perceptron, e $\boldsymbol{\theta}$ é o viés. O viés afasta o limite de decisão da origem e não depende de nenhum valor de entrada. Caso não seja utilizado uma função de ativação limiar, o processo se torna um modelo regressivo, diferindo de regressão linear devido à forma de estimação de parâmetros.

E. Perceptron Multicamadas

Perceptron Multicamadas (MLP — *Multi Layer Perceptron*) é uma rede neural com uma ou mais camadas ocultas com um número indeterminado de neurônios. A camada oculta possui esse nome porque não é possível prever a saída desejada nas camadas intermediárias.

Entrada: dados de entrada, quantidade de neurônios
início
  # Etapa I: Inicialização.
  Atribuir valores aleatórios para os pesos e limite
  Escolha dos valores iniciais influencia o comportamento da redes
  Na ausência de conhecimento prévio os pesos e limites devem ter valores iniciais aleatórios e pequenos uniformemente distribuídos
  # Etapa II Iteração.
  repita
    # Etapa III: Ativação.
    Calcular os valores dos neurônios da camada oculta
    Calcular os valores dos neurônios da camada de saída
    # Etapa IV: Treinar os Pesos.
    Calcular os erros dos neurônios das camadas de saída e oculta
    Calcular a correção dos pesos
    Atualizar os pesos dos neurônios das camadas de saída e oculta
  até *que satisfaça o critério de erro*;
fim

Algoritmo 1: Perceptron Multicamadas.

Para treinar a rede MLP, o algoritmo comumente utilizado é a retro-propagação, diferentemente do Perceptron e Adaline, onde existe apenas um único neurônio de saída y, a MLP pode relacionar o conhecimento a vários neurônios de saída. Conforme é mostrado no Algoritmo 1, o algoritmo de aprendizado MLP é chamado *backpropagation* é composto de 4 etapas. **Observação**: aumentar o número de camadas e neurônios nem sempre é a melhor solução para uma melhoria de desempenho/acurácia.

F. Funções de Ativação

1) Identidade: A função Identidade ou linear apenas aplica um fator de multiplicação ao valor que recebe, apesar de cumprir todos os requisitos, essa função é limitada em sua capacidade de compreender relações mais complexas entre os dados, justamente por ser linear.

$$\begin{aligned} \text{Identidade}(x) &= y(x) = x \\ y'(x) &= 1 \end{aligned} \tag{5}$$

Além disso, sua derivada é constante, o que faz com que o gradiente a cada etapa de *backpropagation* seja constante, assim a etapa de descida do gradiente não tende a convergir para produzir um erro estável próximo de zero. Na camada de saída, a função de ativação linear pode ser utilizada em problemas de regressão, já que produz resultados em todo o domínio dos números reais.

2) Sigmoide: Como neurônios biológicos funcionam de forma binária, a função sigmoide é uma boa forma de modelar esse comportamento, já que assume valores apenas entre 0 e 1. No entanto, se olharmos sua derivada, podemos observar que ela satura para valores acima de 5 e abaixo de -5. A função sigmoide ou logística e sua derivada são dadas, respectivamente, por:

$$\begin{aligned} \text{Sigmoide}(x) &= \sigma(x) = \tfrac{1}{(1+e^{-x})} \\ \sigma'(x) &= \sigma(x)(1-\sigma(x)) \end{aligned} \tag{6}$$

Assim, não é mais recomendado utilizar a função logística com não linearidade de ativação nas redes neurais artificiais. Ela ainda pode ser utilizada na saída da RNA, para modelar variáveis binárias, além disso, alguns modelos probabilísticos, redes neurais recorrentes e alguns modelos não supervisionados têm restrições que tornam uma função sigmoide necessária.

3) Tangente Hiperbólica: Similar a função sigmoide, a função Tangente Hiperbólica (TanH) também tem um formato de *'S'*, mas varia de -1 a 1, em vez de 0 a 1 como na sigmoide. A Tangente Hiperbólica se aproxima mais da identidade, sendo assim uma alternativa mais atraente do que a sigmoide para servir de ativação às camadas ocultas das RNAs. A Tangente Hiperbólica sua derivada é dada, respectivamente, por:

$$\begin{aligned} \text{Tangente Hiperbólica}(x) &= \tanh(x) \\ \tanh'(x) &= 1 - \tanh^2(x) \end{aligned} \tag{7}$$

Podemos ver que as saturações continuam presentes, mas o valor da derivada é maior, chegando ao máximo de 1 quando $x = 0$. Por esse motivo, quando uma função sigmoide precisa ser utilizada, recomenda-se a Tangente Hiperbólica no lugar da sigmoide. A Tangente Hiperbólica tem as mesmas vantagens e ainda resolve um dos problemas da função sigmoide, centrada em zero. Entretanto, sua derivada também converge a zero, e mais rapidamente.

G. K Médias

O k-médias é um tipo de algoritmo de classificação (agrupamento), que pode ser utilizado para a classificação não-supervisionada. No Algoritmo 2, apresentados os passos envolvidos no método k-médias, onde k é o número de clusters (grupos) desejado e informado inicialmente. Como critério de convergência pode ser empregado um número máximo de iterações ou executar o algoritmo até que os centros não se movam mais, ou apresentam uma mudança muito pequena. No último caso, deve ser definido um erro mínimo no início da execução.

Entrada: dados de entrada, valor de $k > 0$
início
  Determinar as posições iniciais dos k centroides dos clusters;
  repita
    Alocar cada elemento ao cluster do centroide mais próximo;
    Recalcular os centros dos clusters a partir dos elementos alocados;
  até *o critério de convergência ser atendido*;
fim

Algoritmo 2: K Médias.

H. Web Scraping

Web scraping[4], também conhecido como extração de dados da web, é o nome dado ao processo de coleta de dados estruturados da web de maneira automatizada. É uma técnica usada para extrair informações de sites da web e transformá-las em um formato estruturado que possa ser armazenado e analisado, conforme apresentado no Algoritmo 3.

I. Valores discrepantes (Outliers)

Outlier, valor aberrante ou valor atípico, é uma análise que apresenta um grande deslocamento das demais da série, ou que é incoerente. A presença de *outliers* pode prejudicar a interpretação dos resultados dos testes estatísticos aplicados às amostras. Em análises estatísticas o efeito do *outlier* pode ser

[4]raspagem de rede, em tradução livre

Entrada: Endereço ou URL do site que você deseja realizar a raspagem;
início
  Inspecionar a página HTML em busca dos elementos de imóveis;
  Encontrar os dados que deseja extrair (valores, quartos, localização, etc);
  Extrair os dados da páginas 1 até n;
  Armazenar os dados no formato exigido (.CSV);
fim

Algoritmo 3: *Web Scraping*.

facilmente observado, um dos métodos de se calcular *outlies* é o *escore z*, *z-score* ou desvio padrão Equação 8, neste método, será considerado *outlier* o valor que se encontrar deslocado em uma determinada proporção dos desvios padrões da média, a dimensão desses desvios pode variar conforme o tamanho da amostra.

$$z = \frac{x - \mu}{\sigma} \tag{8}$$

J. Normalização

A normalização dispõe os dados em intervalos com 0 e 1 ou -1 e 1, caso haja valores negativos, sem alterar as diferenças nas faixas de valores. Ou seja, ela não retira os *outliers*. Se a distribuição não é Gaussiana ou o desvio padrão é muito pequeno, normalizar os dados é uma escolha melhor em comparação a padronização. A padronização dispõe os dados em uma mesma escala. Esse método é mais bem utilizado quando a nossa distribuição é gaussiana.

$$x_{\text{norm}} = \frac{x - x_{j_{\min}}}{x_{j_{\min}} - x_{j_{\max}}} \tag{9}$$

IV. METODOLOGIA

Todos os algoritmos foram elaborados em linguagem Python, utilizando o IDE (ambiente de desenvolvimento integrado) PyCharm 2023.1.3 (*Community Edition*), sistema operacional Windows 11 Home e processados em um computador com a seguinte configuração: processador, Intel® Core™ i7-8550U CPU @ 1,80GHz 1,99 GHz, RAM 8,00 GB, 64-bit.

A. Conjuntos de dados

Nessa Seção apresentamos as metodologias usadas para coletar os conjuntos de dados usados nesse trabalho.

1) Taipei 2018: Nesse trabalho foi usado o conjunto de dados históricos do mercado de avaliação imobiliária coletado da cidade de Taipei, capital de Taiwan, disponibilizado em 2018. Essa base de dados foi obtida em formato *.csv* no site

repositório da UCI[5]. O conjunto de dados, Taiwan (*Real estate valuation data set* (2018)) tem 414 imóveis, com 7 atributos, sendo eles: data de transação, idade do imóvel, distância até a estação MRT[6] mais próxima, Lojas de conveniência, latitude, Longitude e Preço do imóvel por área.

2) Conjuntos de dados Norte-MG: A intenção de usar os dados de Taipei era ter uma referência de estudo, já que são dados conhecidos e extremamente usados em vários trabalhos[7]. Um dos objetivos deste trabalho específico é usar dados imobiliários do Norte de MG, por isso foi necessária uma base real mais atual e com a necessidade dos imóveis localizarem-se no norte de MG. Sendo assim, foram extraídos dados de anúncios de dois sites da web através do método *Web Scraping* (ver Seção III-H).

Nesse trabalho utilizaram-se os pacotes do Python *Requests* e *Beautiful Soup* para poder obter dados das páginas web. O módulo Requests lhe permite integrar seus programas Python com *web services*, enquanto o módulo Beautiful Soup é projetado para fazer com que a captura de tela ou *screen-scraping* seja simplificada. Para fazer *web scraping* em páginas da *web*, foram seguidas as etapas básicas do *web scraping* com Python (ver Seção III-H).

Durante a construção das bases de dados reais, alguns atributos foram gerados, inspirados pelos atributos presentes na basa de *Benchmark* (Taipei). Os atributos Latitude e Longitude foram acrescentados as bases reais, através da utilização do pacote *Geopy* que possui funções que nos entrega tais coordenadas conforme o endereço do local.

Logo após, foi aplicado os processos de tratamento de dados faltantes (ver Seção IV-B1) e valores discrepantes (ver Seção III-I) (essa etapa foi realizada com o método *k médias* (ver Seção III-G), que usou os atributos dos dados para encontrar os melhores agrupamentos, e assim possibilitou-se o tratamento dos valores discrepantes), resultando em dois conjuntos de dados, um para cada site onde foi efetuada a extração, nomeados com o Norte-MG1[8] e Norte-MG2. Os conjuntos de dados *Norte-MG1* tem 1859 imóveis, sendo obtidos em uma plataforma web onde o proprietário (usuário) negociante do imóvel é o responsável para inserir os dados do imóvel que cogita vender. O *Norte-MG2* tem 1149 imóveis, foram obtidos em uma plataforma *web* especializada em negociações imobiliária, onde a inserção de dados é feita por funcionários especializados. Ambos possuem 7 atributos, sendo eles: área, quartos, garagem, banheiro, latitude, longitude e preço.

A Tabela I apresenta as legendas dos atributos para os três conjuntos de dados de Taipei, Norte-MG1 e Norte-MG1.

Tabela I
LEGENDA DOS ATRIBUTOS DOS CONJUNTOS DE DADOS.

Atributos	Taipei 2018	Norte-MG1	Norte-MG2
x_1	Transação (ano).	Área (m^2).	Área (m^2).
x_2	Idade (ano).	Quartos.	Quartos.
x_3	Distância até a estação MRT* (m).	Garagem.	Garagem.
x_4	Lojas de conveniência.	Banheiros.	Banheiros.
x_5	Latitude (°).	Latitude (°).	Latitude (°).
x_6	Longitude (°).	Longitude (°).	Longitude (°).
y	Preço por área**.	Preço ($R$$).	Preço ($R$$).

* Estação de Metrô;
** 10000 Novo Dólar de Taiwan/Ping, onde Ping é uma unidade local, 1 Ping = 3,3 metros quadrados.
Fonte: Autor.

B. Pré-processamento dos dados

1) Dados faltantes: Nesse trabalho os dados de atributos omitidos, com erro de preenchimento, fora de escala, foram considerados dados faltantes[9], onde não se tem informação consistente da característica (Exemplo: *"um anúncio de uma casa com 3 quartos, 2 banheiros, 50 m^2 localizado no centro de BH,* ***com um valor de*** *R**1,00**"*). Esse valor no anúncio não é o valor real do imóvel, trata-se de um método que anunciantes usam para atrair possíveis compradores de seu imóvel, e muitas vezes o valor é combinado com o comprador se houver interesse do mesmo. Esse tipo de dados pode ser prejudicial na etapa de treinamento dos modelos de regressão, que propagam em um erro que afeta os resultados. Para tratar esse tipo de dado, com essas características duvidosas foram considerados nulos (Exemplo: vários imóveis tinham valores muito inferiores a 17 mil reais e com 4 m^2). Após a anulação desses dados, corrigiu-se os mesmo com a média para encontrar os dados referentes a cada atributos que foi anulado. O próximo passo é tratar os valores discrepantes sem o efeito dos dados faltantes.

2) Normalização dos dados: Antes da etapa de treinamento os dados foram normalizados aplicando-se a Equação 9. Normalizar os dados é uma técnica usada para colocar os dados em uma escala comum. Isso é feito para que as características dos dados tenham a mesma escala e para que o processo de aprendizado seja acelerado. Além disso, normalizar os dados ajuda a cuidar de diferentes características de maneira justa sem importar a sua escala.

[5]UCI *Machine Learning Repository*: O UC *Irvine Machine Learning Repository* é um repositório de conjuntos de dados para aprendizado de máquina mantido pela Universidade da Califórnia em Irvine.

[6]A estação de metro de Taipé que é um sistema metropolitano que serve a cidade de Taipé desde 1996.

[7]como em H.-Y. Lin e KUENTAI (2015)

[8]Norte de Minas Gerias

[9]Dados faltantes não são tratados como **Valores discrepantes** ou ***Outliers***.

3) Ajustes de Hiper-parâmetros: Os hiper-parâmetros são propriedades que controlam o treinamento de um modelo de aprendizado de máquina, ajustando-os, podemos tornar o modelo mais bem preparado para resolver um problema realista. À vista disso, os dados foram testados e validados nos quarto métodos apresentados nesse trabalho, porém, para encontrar os melhores resultados temos que ajustar parâmetros e também testar as funções de ativação Sigmoide (ver Seção III-F2) e Tangente Hiperbólica (ver Seção III-F3) no Perceptron e nas camadas ocultas da rede MLP[10] em cada modelo preditivo. Para isso, utilizou-se a validação cruzada para validar os melhores modelos encontrados. Para ter um modelo generalizado que não sofra com *Overfitting*[11] ou *Underfitting*[12], aplicou-se a validação cruzada com $k = 10$ *Folds* dos conjuntos de dados, os resultados dos modelos após a validação estão no Seção V.

V. RESULTADOS E DISCUSSÃO

Foram realizados experimentos com os quatro modelos de aprendizado de Máquina. Todos os resultados foram obtidos utilizando validação cruzada com o $k = 10$ *Folds* e a melhor escolha de função de ativação foi a Sigmoide para o Perceptron e nas camadas ocultas da MLP.

A. Análise descritiva dos dados

A estatística descritiva visa descrever e resumir um conjunto de dados. Para isso, são utilizadas medidas descritivas, como a média e o desvio padrão, além de gráficos e tabelas que permitem visualizar e entender aspectos importantes dos dados. A estatística descritiva é a etapa inicial da análise de dados e tem por objetivo descrever os dados observados.

A Tabela II, apresenta as estatísticas descritivas dos conjuntos de dados onde é possível analisarmos: média, desvio padrão, valor mínimo e valor máximo de cada atributo dos três conjuntos de dados.

B. Dispersão dos dados

As Figuras 1(a), 1(b), 1(c) apresentam gráficos de dispersão dos atributos entre si e na diagonal, um histograma para cada atributo onde podemos analisar como os dados se comportam entre si.

[10]Perceptron de Múltiplas Camadas (*Multi Layer Perceptron* em inglês)

[11]*Overfitting* (sobre-ajuste ou superajuste) ocorre quando um modelo é muito complexo e produz boas previsões para pontos de dados no conjunto de treinamento, mas tem um desempenho ruim em novas amostras.

[12]*Underfitting* (sub-ajuste) ocorre quando o modelo de machine learning não está bem ajustado ao conjunto de treinamento e o modelo resultante não está capturando bem a relação entre entrada e saída.

Tabela II
ESTATÍSTICAS DESCRITIVAS DO CONJUNTO DE DADOS.

Atributos	Média	Desvio padrão	Mínimo	Máximo
		Taipei		
x_1	2013,149	0,282	2012,667	2013,583
x_2	17,713	11,392	0,000	43,800
x_3	1083,886	1262,110	23,383	6488,021
x_4	4,094	2,946	0,000	10,000
x_5	24,969	0,012	24,932	25,015
x_6	121,533	0,015	121,474	121,566
y	37,980	13,606	7,600	117,500
		Norte-MG1		
x_1	189,544	123,308	10,000	9,0e+2
x_2	2,903	0,721	1,000	5,000
x_3	2,238	1,231	0,000	5,000
x_4	1,901	1,012	1,000	5,000
x_5	-16,691	0,315	-18,445	-15,15
x_6	-43,908	0,406	-45,834	-42,22
y	495523,581	378361,955	9500,000	1,8e+6
		Norte-MG2		
x_1	204,409	177,685	30,500	1991,410
x_2	2,877	0,737	1,000	5,000
x_3	1,691	0,959	0,000	4,000
x_4	1,674	0,774	1,000	4,000
x_5	-16,612	0,304	-17,887	-14,898
x_6	-43,805	0,281	-44,991	-42,314
y	422327,735	211047,619	8000,00	9,9e+5

Fonte: Autor.

C. Gráficos de Georreferência

As Figuras 2(a), 2(b), 2(c) apresentam gráficos com subplotes: em seu eixo horizontal (x) é plotado o atributo Longitude e vertical (y) é plotado Longitude; e a escala de cor (Azul), é determinada pelo atributo qualitativo de cada imóvel.

D. Resultados dos modelos

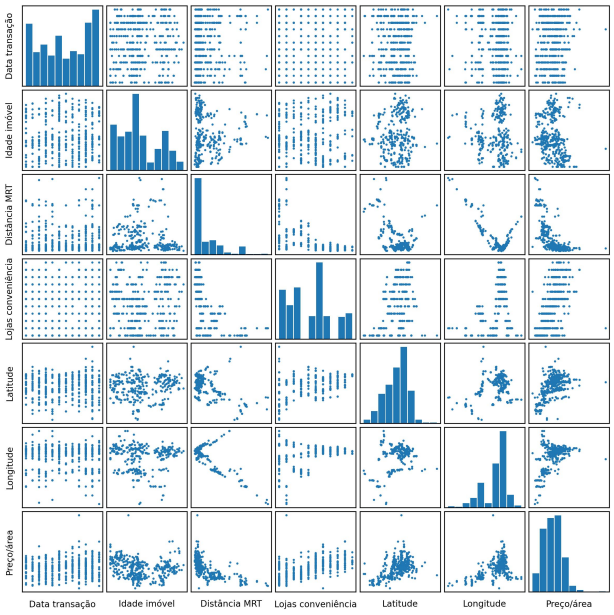
Apresentamos aqui os resultados obtidos dos métodos de aprendizado de máquina propostos nesse trabalho.

A Tabela III apresenta os melhores resultados da regressão polinomial em ordem crescente, o desvio padrão da validação cruzada e o grau do polinômio usado no treinamento do modelo.

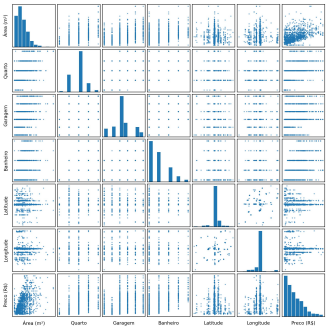
Para obter os resultados da Tabela III, foi preciso ajustes dos modelos de regressão Polinomial, foi testado o grau do polinômio entre 1 e 20 e notou-se que valores superiores a 7 aumentam consideravelmente o erro e o desvio padrão.

A Tabela IV apresenta os melhores resultados do Perceptron Multicamadas em ordem crescente, o desvio padrão da validação cruzada e a quantidade de neurônios usados no treinamento do modelo.

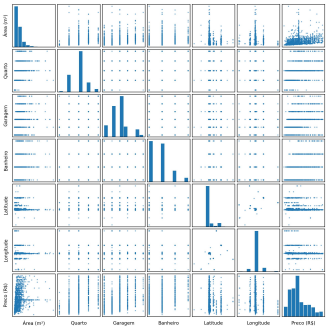
Encontrar os resultados na Tabela IV exigiu a determinação empírica de parâmetros ideais, como o número de neurônios a serem definidos antes do teste. O erro e o desvio padrão da rede MLP tiveram bom desempenho até cerca de 20 neurônios,



(a)



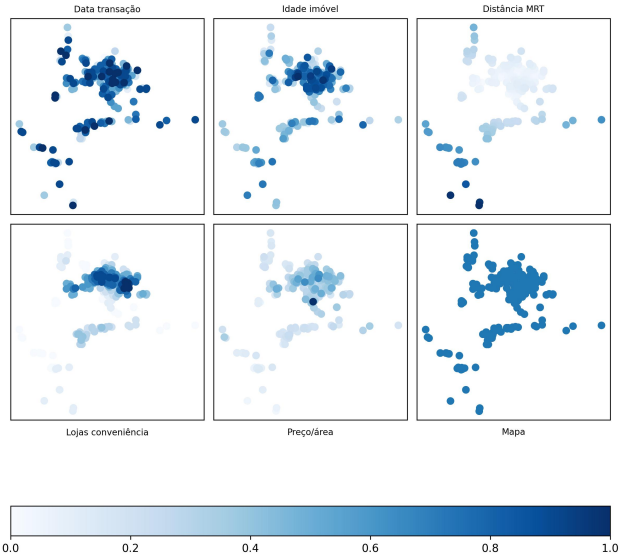
(b)



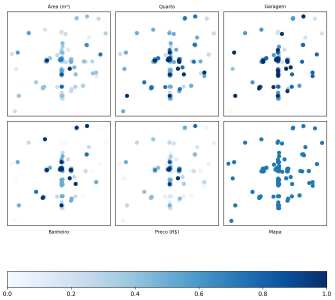
(c)

Fonte: Gráfico criado com a biblioteca Matplotlib em Python.

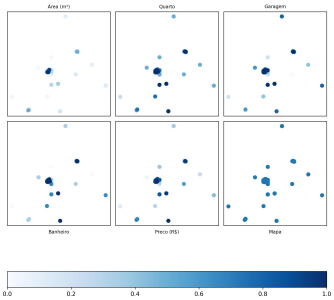
Figura 1. Gráfico de dispersão do conjunto de dados (a) **Taipei 2018** (b) **Norte-MG1** (c) **Norte-MG2**.



(a)



(b)



(c)

Fonte: Gráfico criado com a biblioteca Matplotlib em Python.

Figura 2. Latitudes e Longitude do conjunto de dados (a) **Taipei 2018** (b) **Norte-MG1** (c) **Norte-MG2**.

mas acima desse valor a variância dos resultados aumentou significativamente.

A Tabela V mostra os melhores resultados para os quatro modelos de aprendizado supervisionado (regressão linear, regressão polinomial, perceptron e perceptron multicamada), o desvio padrão da validação cruzada e o tempo médio de treinamento dos modelos em ordem crescente.

Os modelos de regressão linear e Perceptron não requerem ajuste de parâmetros, portanto, não são mostrados na tabela de comparação resultante.

As Figuras 3(a) e 3(b) apresentam gráficos: em seu eixo horizontal (x) é plotado Área (m^2); na vertical (y) é plotado o atributo Preço do Imóvel ($R\$$), esse eixo esta na escala Logarítmica para melhorar visualização da curva da previsão. Essa é uma figura ilustrativa, já que existem vários outros atributos influenciando a curva de predição e não apenas a área.

E. Discussão

Analisando a Figura 1(a), podemos observar o comportamento dos imóveis em Taipei. Em **Data de transação**: é demonstrado uma frequência maior nos imóveis com data de transação mais recente, nos dados exitem várias distribuições com média como pode ser observado no histograma em Platô

Tabela III
MELHORES RESULTADOS DA REGRESSÃO POLINOMIAL.

EMQ	Desvio padrão	G°
Taipei		
$4{,}928e{-}3$	$3{,}435e{-}3$	3
$5{,}412e{-}3$	$3{,}451e{-}3$	2
$5{,}725e{-}3$	$3{,}759e{-}3$	4
$6{,}555e{-}3$	$3{,}634e{-}3$	1
$7{,}092e{-}3$	$5{,}500e{-}3$	5
Norte-MG1		
$2{,}688e{-}2$	$3{,}632e{-}2$	4
$2{,}689e{-}2$	$3{,}601e{-}2$	3
$2{,}752e{-}2$	$3{,}786e{-}2$	6
$2{,}816e{-}2$	$3{,}769e{-}2$	2
$2{,}833e{-}2$	$3{,}627e{-}2$	1
Norte-MG2		
$3{,}461e{-}2$	$9{,}755e{-}2$	4
$3{,}473e{-}2$	$9{,}974e{-}2$	3
$3{,}575e{-}2$	$9{,}706e{-}2$	2
$3{,}729e{-}2$	$1{,}060e{-}2$	1
$6{,}366e{-}2$	$8{,}190e{-}2$	5

Fonte: Autor.

Tabela IV
MELHORES RESULTADOS DO PERCEPTRON MULTICAMADAS.

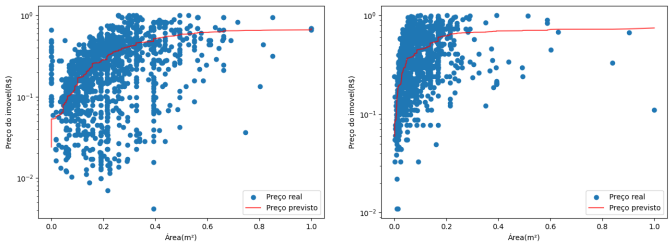
EMQ	Desvio padrão	N
Tapei		
$4{,}734e{-}3$	$3{,}256e{-}3$	5
$4{,}868e{-}3$	$3{,}330e{-}3$	3
$4{,}889e{-}3$	$3{,}310e{-}3$	7
$4{,}897e{-}3$	$3{,}363e{-}3$	8
$4{,}906e{-}3$	$3{,}267e{-}3$	4
Norte-MG1		
$3{,}664e{-}2$	$5{,}146e{-}2$	4
$3{,}671e{-}2$	$5{,}186e{-}2$	6
$3{,}671e{-}2$	$5{,}172e{-}2$	3
$3{,}674e{-}2$	$5{,}166e{-}2$	5
$3{,}687e{-}2$	$5{,}257e{-}2$	7
Norte-MG2		
$2{,}689e{-}2$	$1{,}126e{-}2$	17
$2{,}740e{-}2$	$1{,}054e{-}2$	16
$2{,}741e{-}2$	$1{,}133e{-}2$	12
$2{,}775e{-}2$	$1{,}067e{-}2$	15
$2{,}836e{-}2$	$1{,}186e{-}2$	9

Fonte: Autor.

Tabela V
MELHORES RESULTADOS DOS MODELOS DE APRENDIZADO DE MÁQUINA.

Modelos	EMQ	Desvio padrão	Tempo médio (s)
Taipei			
MLP	$4{,}734e{-}3$	$3{,}256e{-}3$	48,956
RP	$4{,}928e{-}3$	$3{,}433e{-}3$	$1{,}655e{-}3$
P	$5{,}821e{-}3$	$3{,}494e{-}3$	$4{,}142e{-}1$
RL	$6{,}555e{-}3$	$3{,}633e{-}3$	$5{,}507e{-}4$
Norte-MG1			
RP	$2{,}688e{-}2$	$3{,}632e{-}2$	$6{,}988e{-}3$
RL	$2{,}833e{-}2$	$3{,}627e{-}2$	$2{,}089e{-}3$
MLP	$3{,}664e{-}2$	$5{,}146e{-}2$	16,278
P	$3{,}718e{-}2$	$4{,}128e{-}2$	20,995
Norte-MG2			
MLP	$2{,}689e{-}2$	$1{,}126e{-}2$	11,444
RP	$3{,}461e{-}2$	$9{,}755e{-}3$	$6{,}089e{-}3$
RL	$3{,}729e{-}2$	$1{,}060e{-}2$	$7{,}984e{-}4$
P	$4{,}335e{-}2$	$2{,}182e{-}2$	17,450

Fonte: Autor.



(a) (b)

Fonte: Gráfico criado com a biblioteca Matplotlib em Python.

Figura 3. Previsão do preço dos imóveis (a) Norte MG1 (b) Norte MG2.

[13], isto ocorre quando existem várias distribuições com médias diferentes; **Idade do imóvel:** é possível analisar até três grupos nesse histograma, entre imóveis mais novos (até 5 anos), mais velhos ($\approx$ 20 anos) e mais antigos ($\approx$ 40 anos), como é possível observar no histograma Multimodal [14], dessa forma sabemos que em dois momentos diferentes há uma concentração de frequência que se destaca; **Distância MRT:** os imóveis em sua maioria estão próximos às estações MRT, como é possível observar no histograma assimetria à direita [15]; **Lojas de Conveniência:** esse atributo está dividido em quatro grupos, 1º (nenhuma loja), 2º (1 à 2 lojas), 3º (4 à 6 lojas), 4º (9 à 10 lojas); **Latitude:** os imóveis estão centrados em relação à média, isso pode acontecer, pois os imóveis estão próximos em relação à latitude, como é possível observar no histograma simétrico [16]. **Longitude:** a formação de dois grupos se dá pelo fato os imóveis serem de suas regiões verticalmente diferentes em relação à longitude. **Preço por área:** os valores dos imóveis se concentra entre (22 à 44 Taiwan/Ping), como é possível observar no histograma quase-simétrico.

Com a Figura 2(a) é possível analisar como os atributos qualitativos, se comportam em relação à localização geográfica, nessa figura podemos analisar algumas correlações não lineares, comparando preço por área em relação aos outros atributos: quanto menor a distância até a estação MRT e quanto mais lojas de conveniência, maior será a valorização do imóvel para a maioria dos imóveis, porém a data de transação e idade do imóvel são menos expressivos.

Na Tabela III, podemos analisar que a regressão polinomial é um bom modelo de previsão. Por ser possível obter bons resultados em todas as bases de dados, isso nos mostra que os imóveis podem ser bem representados por este método, além disso, seu parâmetro de treinamento depende apenas do grau de seu polinômio, facilitando muito o processo.

Em comparação com os resultados da Tabela IV, os resultados de Taipei são melhores em comparação com outros dados porque é uma base de referência com tratamento diferente e trata-se apenas de uma cidade; Os resultados obtidos para a base Norte-MG1 podem ter sido mais afetados pelos dados duvidosos na fase de coletados, apesar de terem sidos tratados na etapa de pré-processamento; Nos resultados do Norte-MG2, a quantidade de neurônios pode ser associado aos clusters observados na Figura 2(c), por serem dados de mais de vinte cidades diferentes;

[13] Quando suas barras têm praticamente as mesmas alturas.
[14] Quando há o aparecimento de vários picos.
[15] Quando a distribuição de dados indica a ocorrência de altos valores com baixa frequência.
[16] Apresenta uma frequência mais alta no centro, que vai diminuindo conforme se aproxima das bordas.

Analisando a Tabela V, é possível analisar que o *Multilayer Perceptron* obteve melhores resultados na base Taipei e Norte-MG2. Quanto a regressão polinomial obteve melhores resultados na base (Norte-MG1), o motivo pode ser devido à rede MLP não estar com os hiper-parâmetros bem ajustados, embora vários dos melhores parâmetros tenham sido tentados. Espera-se que uma rede MLP supere todos os outros métodos propostos nesse trabalho porque seu diferencial é consistente e pode acomodar todas as propriedades em um espaço de variáveis contínuas adaptadas à informação que está tentando aprender.

VI. CONCLUSÃO

As técnicas de *machine learning* utilizadas neste estudo mostraram ser extremamente úteis para a previsão de preços de imóveis. Todos os quatro métodos de aprendizado de máquina revelaram capacidades de grande precisão e qualificação, tendo em vista as bases de dados utilizadas.

Foi constatado que, embora seja uma técnica simples da estatística, a regressão polinomial apresentou o desempenho satisfatório na maioria dos experimentos, ficando atrás somente da rede MLP, uma vez que obteve resultados equivalentes ou superiores.

A criação de bases de dados imobiliários para a região Norte de Minas Gerais se mostrou eficaz, embora ainda fossem necessárias melhorias na etapa de processamento do banco de dados para maximizar seu potencial.

O trabalho atingiu os objetivos pretendidos, e os resultados obtidos podem contribuir para um melhor entendimento do processo de previsão de valor imobiliário. Produzindo estudos de valor de propriedade colaborativamente para diminuir o tempo de diagnóstico usando técnicas de aprendizado de máquina.

Para trabalhos futuros, podemos melhorar os banco de dados norte-MG1 e norte-MG2 e possivelmente obter novos bancos de dados de aglomeração metropolitana para testar melhores dados regionais. Também explorar o refinamento de hiper-parâmetros e arquiteturas de rede MLP, encontrar a melhor estrutura e configuração de rede para o problema descrito, implementar métodos ou modelos que visem reduzir o problema de desequilíbrio de banco de dados.

REFERÊNCIAS

Araújo, Elton Gean et al. (2012). “Proposta de uma metodologia para a avaliação do preço de venda de imóveis residenciais em Bonito/MS baseado em modelos de regressão linear múltipla”. Em: *P&D em Engenharia de Produção* 10.2, pp. 195–207.

Baptistella, Marisa, Maria Teresinha Arns Steiner e Anselmo Chaves Neto (2005). “O uso de redes neurais e regressão linear múltipla na engenharia de avaliações: Determinação dos valores venais de imóveis urbanos”. Em: *Diss., Universidade Federal do Paraná*.

Dantas Rubens Alves e Cordeiro, GM (1988). “Uma nova metodologia para avaliação de imóveis utilizando modelos lineares generalizados”. Em: *Revista Brasileira de Estatística*.

González, Marco Aurélio Stumpf (1997). “Variação qualitativa e índices de preços na análise do comportamento recente dos aluguéis residenciais em Porto Alegre (1994-1997”. Em: *Análise Econômica* 15.28.

Lin, Hong-Yu e CHEN KUENTAI (2015). “The Trend of Average Unit Price in Taipei City”. Em: *Research in World Economy* 6.1, p. 133.

Lin, Hongyu e Kuentai Chen (2011). “Predicting price of Taiwan real estates by neural networks and support vector regression”. Em: *Proc. of the 15th WSEAS Int. Conf. on Syst*, pp. 220–225.

Moreira, Daniela Souza, RS Silva e AMR Fernandes (2010). “Engenharia de avaliações de imóveis apoiada em técnicas de análise multicritério e redes neurais artificiais”. Em: *Revista de Sistemas de Informação da FSMA* 6, pp. 49–58.

Real estate valuation data set (2018). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5J30W.

Russell, Stuart J (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Selim, Sibel (2008). “Determinants of house prices in Turkey: A hedonic regression model”. Em: *Doğuş Üniversitesi Dergisi* 9.1, pp. 65–76.

AGRADECIMENTOS

Os autores agradecem à Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM) pelo apoio ao desenvolvimento deste trabalho.