UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

THIAGO CÉSAR DE MIRANDA SILVA

Uso de Machine Learning para Previsão de Valores de Apartamentos no Município do Recife

Recife

Setembro de 2023

THIAGO CÉSAR DE MIRANDA SILVA

Uso de Machine Learning para Previsão de Valores de Apartamentos no Município do Recife

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco - UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Cleviton Vinicius Fonsêca Monteiro
Co-orientador: Rodrigo Gabriel Ferreira Soares

Recife
Setembro de 2023

THIAGO CÉSAR DE MIRANDA SILVA

Uso de machine learning para previsão de valores de apartamentos no município do Recife

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 12 de Setembro de 2023.

BANCA EXAMINADORA

Cleviton Vinicius Fonsêca Monteiro (Orientador)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Rodrigo Gabriel Ferreira Soares
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Victor Wanderley Costa de Medeiros
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Uso de machine learning para previsão de valores de apartamentos no município do Recife

Thiago César de Miranda Silva [1]
Orientador: Cleviton Vinicius Fonsêca Monteiro [1]
Co-Orientador: Rodrigo Gabriel Ferreira Soares [1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

thiago.cesar@ufrpe.br, cleviton.monteiro@ufrpe.br, rodrigo.gfsoares@ufrpe.br

Resumo. *A pandemia de COVID-19 trouxe consigo uma série de efeitos econômicos e transformações relacionadas ao comportamento e à forma de morar, que, por sua vez, tiveram repercussões nos preços dos imóveis e na demanda de imóveis. Nesse contexto, a previsão de preços de imóveis assume um papel de extrema importância, contribuindo para decisões mais informadas, atenuando os riscos e promovendo uma maior transparência no setor imobiliário. A implementação da automação na previsão de preços amplia ainda mais essa dinâmica, aprimorando significativamente a precisão, a eficiência e a confiabilidade das previsões, além de proporcionar ajustes às flutuações do cenário econômico com mais agilidade. Usando anúncios disponíveis na OLX, foi criada uma base de dados georreferenciada para gerar um modelo de previsão de preços de apartamentos residenciais, em Recife - por meio de modelos de aprendizagem de máquina em AutoML. Essa ferramenta automatiza o desenvolvimento de modelos de aprendizado de máquina, permitindo experimentação rápida e foco na resolução do problema. O trabalho indica que a má distribuição geográfica dos dados tendenciaram os resultados dos modelos, além disso, foi concluído que os dados encontrados em plataformas de compra e venda online são insuficientes para a geração de um modelo de aprendizado de máquina que apresente um nível de acuracidade aceitável, em Recife, principalmente porque não são apresentados valores de transação do imóvel, apenas o preço anunciado. Contudo, o presente trabalho apresenta importantes contribuições para o avanço em pesquisas relacionadas à automação na previsão de preços de imóveis.*

Abstract. *The COVID-19 pandemic has brought with it a series of economic effects and transformations related to behavior and the way people live, which, in turn, have had repercussions on property prices and real estate demand. In this context, property price forecasting assumes an extremely important role, contributing to more informed decisions, mitigating risks, and promoting greater transparency in the real estate sector. The implementation of automation in price forecasting further enhances this dynamic, significantly improving accuracy, efficiency, and reliability of predictions, while providing adaptability to economic fluctuations with greater agility. Utilizing listings available on OLX, a georeferenced database was created to generate a residential apartment price prediction model in Recife, using machine learning models in AutoML. This*

tool automates the development of machine learning models, enabling rapid experimentation and a focus on problem-solving. The work indicates that the poor geographical distribution of the data has biased the results of the models. Furthermore, it was concluded that the data found on online buying and selling platforms are insufficient for generating a machine learning model that achieves an acceptable level of accuracy in Recife, mainly because transaction values for the properties are not provided, only the advertised prices. However, this current work provides significant contributions to the advancement of research related to automation in real estate price prediction.

1. Introdução

A pandemia de Covid-19 causou uma crise econômica global sem precedentes, afetando inclusive o mercado imobiliário, e segundo Fidelis (2023), resultando em aumento de custos de construção, dificuldades na renegociação de contratos de aluguel e redução de investimentos em ativos imobiliários. Um dos motivos desse impacto é o fato do setor ser influenciado diretamente por fatores geográficos e macroeconômicos, como taxas de juros e níveis de emprego (Ferreira Wisniewski, 2022; Cintra, 2021; Fernandes et al., 2021; Dias e da Silva, 2021; Nunes et al., 2020; Pereira et al., 2014).

A pandemia trouxe ainda algumas mudanças em diversos aspectos da vida urbana, como habitação, trabalho, mobilidade, educação, lazer e entretenimento. Em consequência, o panorama do mercado imobiliário no pós-pandemia é marcado por novas tendências e oportunidades (Balemi et al., 2021; Guimarães, 2022; Kaklauskas et al., 2021; Jovanović-Milenković et al., 2020; Gupta, 2023; Liu e Su, 2021).

Entre as novas tendências estão a busca por mais conforto, como a necessidade de mais espaço, com mais dormitórios, mais banheiros, assim como varanda, vista e oferta de mais área de lazer, e com a localização mais próxima ao local de trabalho, com oferta de serviços e comércio, como pode ser visto na Figura 1, que apresenta o nível de importância para os consumidores sobre cada item de um imóvel.

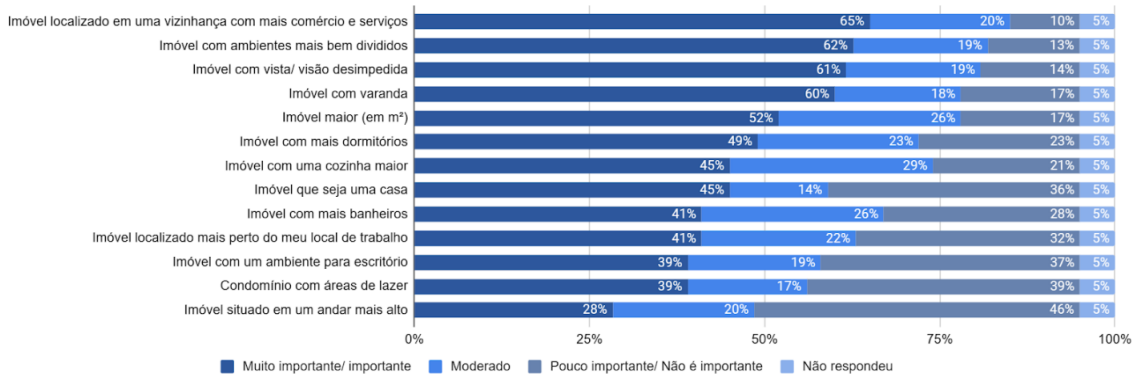

Figura 1. Nível de importância que cada item na aquisição de um imóvel para morar. Fonte: DataZAP (Abril 2021).

Após três anos, algumas tendências tendem a se consolidar como o trabalho home office e as mudanças em torno da moradia, que atualmente centraliza a vida familiar,

profissional e social, passando a ser lugar de trabalho, de estudo, de lazer, de exercícios físicos e de relacionamento social.

O mercado imobiliário vem apresentando números positivos (Figura 2), no entanto, a perspectiva para 2023 ainda é desafiadora, mesmo com a recuperação do mercado de trabalho e da queda da inflação. Dados do Fundação Instituto de Pesquisas Econômicas (FIPE) (2023) mostram que, após uma queda temporária durante a pandemia, os preços de venda de imóveis na Região Metropolitana de Recife se recuperaram e valorizaram.

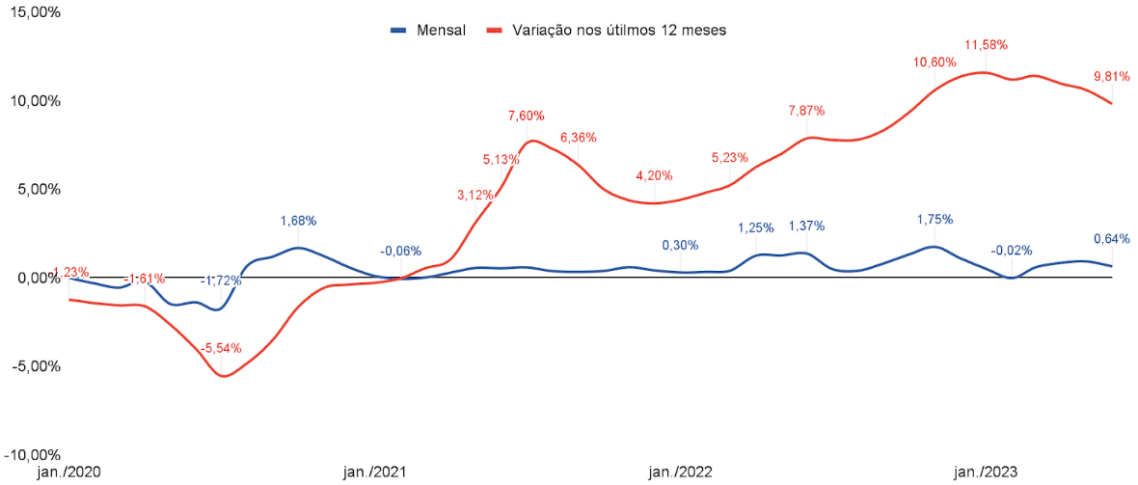

Figura 2. Índice de preços no mercado imobiliário da Região Metropolitana do Recife. Fonte: Fundação Instituto de Pesquisas Econômicas (FIPE) (2023).

A Engenharia de Avaliações é a área da engenharia responsável por estimar o valor de um bem. Sua abrangência a torna uma ciência ampla e com diversos campos de atuação no mercado (Barros, 2018). Segundo Dantas (2005), as aplicações mais comuns estão associadas a transações de compra, venda e locação de imóveis, investimentos imobiliários, processos de financiamentos, na determinação de impostos, seguros e no âmbito judicial.

A previsão de preços no mercado imobiliário é importante por diversas razões, tanto para os profissionais do setor quanto para os consumidores. Aqui estão alguns motivos pelos quais a predição de preços é relevante:

MOTIVOS	**DESCRIÇÃO**
Tomada de decisão baseada em dados	Prever os preços dos imóveis permite que compradores, vendedores e investidores tomem decisões mais informadas sobre a compra, venda ou investimento em propriedades, considerando as tendências de mercado e os possíveis movimentos de preços.

MOTIVOS	DESCRIÇÃO
Planejamento Financeiro	Comprar ou vender uma propriedade é uma decisão financeira significativa. A previsão de preços ajuda as partes envolvidas a planejar suas finanças com antecedência, evitando surpresas desagradáveis.
Avaliação de Investimentos	Investidores imobiliários utilizam previsões de preços para avaliar o potencial de retorno de seus investimentos. Isso os ajuda a decidir quais propriedades têm maior probabilidade de valorização ao longo do tempo.
Negociações Eficientes	Uma estimativa precisa do valor de mercado de um imóvel facilita as negociações entre compradores e vendedores. Isso ajuda ambas as partes a chegar a um acordo mais rapidamente, com base em informações confiáveis sobre os preços.
Identificação de Oportunidades	A previsão de preços pode revelar oportunidades de compra em momentos de baixa no mercado e oportunidades de venda em momentos de alta, permitindo que os participantes aproveitem os melhores momentos para agir.
Gestão de Riscos	Instituições financeiras que concedem empréstimos hipotecários precisam avaliar o risco de inadimplência. Previsões precisas de preços podem ajudar nesse processo, permitindo uma melhor gestão de riscos.
Análise de Mercado	As previsões de preços também contribuem para a análise macroeconômica do mercado imobiliário. Isso é importante para governos, reguladores e economistas, pois ajuda a compreender a saúde e a estabilidade do setor.
Definição de Estratégias de Marketing	Agentes imobiliários e empresas podem usar as previsões de preços para desenvolver estratégias de marketing mais eficazes, destacando propriedades com boa perspectiva de valorização.
Transparência e Confiança	Fornecer previsões de preços baseadas em dados sólidos aumenta a transparência do mercado imobiliário, construindo confiança entre os participantes do mercado, reduzindo incertezas e evitando especulações e bolhas imobiliárias.
Inovação Tecnológica	A previsão de preços estimula a inovação tecnológica no setor imobiliário, promovendo o desenvolvimento de algoritmos avançados, análises de dados e ferramentas de automação.

Tabela 1. Importância da previsão de preços no mercado imobiliário. Fonte: Geltner et al. (2001).

Resumindo, a previsão de preços no mercado imobiliário proporciona informações cruciais que auxiliam as partes envolvidas a tomar decisões mais inteligentes, gerenciar riscos e aproveitar oportunidades. A automação na previsão de preços de imóveis, nesse sentido, é fundamental para o mercado imobiliário, pois proporciona benefícios como

precisão, consistência e eficiência.

Algoritmos e modelos estatísticos analisam dados para reduzir avaliações subjetivas, economizando tempo e permitindo análises avançadas de fatores que afetam o valor dos imóveis. A automação apoia a tomada de decisões, otimizando processos decisórios no mercado imobiliário.

Segundo Delmiro et al. (2022), uma das formas mais utilizadas para se determinar o valor de mercado de um bem é através da regressão linear múltipla com o auxílio da inferência estatística. No entanto, a escassez de dados avaliativos e a falta de dados históricos têm limitado os estudos na área. Para abordar essa lacuna, um modelo de previsão para valores de imóveis residenciais em Recife é proposto, usando dados disponíveis na internet e variáveis do conhecimento público. O objetivo é facilitar a precificação de imóveis e permitir sua replicação em outras localidades.

Este estudo considera a disponibilidade abundante de dados de anúncios, propondo a adoção de uma abordagem baseada em aprendizado de máquina. Essa abordagem é amplamente utilizada no mercado de trabalho e tem se tornado cada vez mais relevante em cenários em que há muitas variáveis e nos quais a aplicação de algoritmos específicos não é viável (Brynjolfsson et al., 2019; Boselli et al., 2017; Colombo et al., 2018; Boselli et al., 2018; Perera e Kamalaruban, 2021; Sidey-Gibbons e Sidey-Gibbons, 2019).

1.1. Formulação do problema

O cenário do mercado imobiliário foi um dos mercados mais cresceu pelo cenário pós covid-19, os valores de imóveis subiram de maneira exagerada com a vinda da pandemia (Baptista et al., 2022; Cintra, 2021; Peev, 2022), além disso, houve um impacto direto na forma de morar e nas variáveis que são consideradas na busca por um imóvel.

O setor representa uma importante área de interesse de estudos, que usa a estatística e a computação, devido à sua complexidade. O uso de aprendizado de máquina pode proporcionar a otimização de processos como a automação na precificação de imóveis, que propicia maior agilidade na adaptabilidade às mudanças de mercado, a redução de viés em avaliações e maior transparência do setor.

Ferramentas como web scraping e AutoML permitem uma predição mais precisa ao lidar com grandes volumes de dados e algoritmos, impulsionando a agilidade na precificação de imóveis e facilitando a tomada de decisões para compradores e vendedores. O presente trabalho pretende responder a questão sobre como automatizar a precificação de imóveis, a partir de uma ferramenta de AutoML.

1.2. Objetivos

Este trabalho tem como objeto principal o desenvolvimento de um modelo de previsão de valor do imóvel, do tipo apartamento, em Recife. Para atingir este objetivo, três etapas deverão ser realizadas, caracterizando-as como os seguintes objetivos específicos:

- Realizar a coleta de dados de apartamentos ofertados para venda, em Recife;
- Realizar o tratamento dos dados extraídos para a criação de uma base de dados, com o intuito de viabilizar o uso desses dados para a criação de um modelo de aprendizado de máquina; e
- Criação e validação do modelo de aprendizado de máquina para previsão do valor de imóveis.

1.3. Organização do trabalho

A Seção 2 apresenta a fundamentação teórica dos assuntos abordados, como a coleta de dados por meio de web scraping, o pré-processamento de dados com georreferenciamento e o AutoML, além dos algoritmos de Aprendizado de Máquina utilizados. Os trabalhos relacionados estão apresentados na Seção 3. A Seção 4 descreve a abordagem proposta, desde a coleta, assim como o pré-processamento dos dados, a modelagem e a avaliação dos modelos de aprendizado de máquina utilizados. Os resultados obtidos são descritos na Seção 6. Por fim, a Seção 7 trata das considerações finais e de sugestões para trabalhos futuros.

2. Referencial teórico

Essa seção aborda uma revisão da literatura pertinente ao tema do presente trabalho com uma síntese dos conceitos e estudos relevantes já realizados, proporcionando um embasamento para a pesquisa, ajudando a contextualizar o problema e fundamentando a metodologia e a análise dos resultados.

2.1. Mineração de dados

A coleta de dados foi realizada por meio de mineração de dados, que segundo Hall et al. (2016) é o processo de descobrir padrões em grandes quantidades de dados, que nos permitem fazer previsões não triviais sobre novos dados, como ajudar com novos insights, melhorar a tomada de decisão e, em algumas configurações, ter vantagens competitivas.

Conforme Hall et al. (2016), existem dois extremos para a expressão de um padrão como uma caixa preta cujo interior é efetivamente incompreensível e como uma caixa cuja construção revela a estrutura do padrão. Ambos, estamos assumindo, fazer boas previsões. A diferença é se os padrões que são minerados são representados em termos de uma estrutura que pode ser examinada, fundamentada sobre, e usada para informar decisões futuras. Tais padrões chamamos de estruturais porque capturam a estrutura de decisão de forma explícita. Em outras palavras, eles ajudam a explicar algo sobre os dados.

2.1.1. Web Scraping

A web nos fornece mais dados do que podemos ler e entender e para trabalhar com essas informações é utilizada a mineração de dados por meio do web scraping, que é um método automatizado de extração de dados de sites, muito usado para coletar grandes números de informação.

Segundo Lawson (2023), o web scraping é a prática de coletar dados por qualquer meio que não seja um programa interagindo com uma Application Programming Interface (API). Isso é mais comumente realizado escrevendo um programa automatizado que consulta um servidor web, solicita dados (geralmente na forma de HTML e outros arquivos que incluem páginas da web) e, em seguida, analisa esses dados para extrair aqueles que são necessários.

Esse método abrange uma ampla variedade de técnicas de programação e tecnologias, como análise de dados e segurança da informação, que são essenciais para coletar e

processar grandes quantidades de dados. Por meio de um script Python, é possível visualizar os dados em seu navegador, acessá-los em um script e armazená-los em um banco de dados, para em seguida realizar qualquer análise com esses dados. Dessa forma, podem ser visualizados bancos de dados que abrangem de milhares a milhões de páginas de uma só vez.

Para realizar o web scraping em um site, primeiro é necessário baixar suas páginas da web contendo os dados de interesse em um processo conhecido como crawling, neste caso, existem várias abordagens que podem ser usadas, a escolha apropriada dependerá da estrutura do site de destino. As três principais abordagens, segundo Lawson (2023), para realizar o crawling de um website são:

- Buscando por um sitemap;
- Iterando os IDs do banco de dados de cada página da web; e
- Seguindo os links de páginas da web.

No presente trabalho foi feito o uso da última abordagem apresentada, a de seguir os links de páginas da web. Essa decisão foi tomada com base no funcionamento da plataforma online utilizada no trabalho, em que é necessário acessar uma página de busca para recuperar os links das páginas que contém os dados.

2.2. Georreferenciamento de dados

O georreferenciamento é definido, segundo Wade et al. (2006) como o alinhamento de dados geográficos a um sistema de coordenadas conhecido para que possam ser visualizados, consultados e analisados com outros dados geográficos.

Conforme Hill (2009), a ciência e a engenharia, além de práticas governamentais, comerciais e políticas estão incorporando, cada vez mais, sistemas de informação geográfica (GIS) para armazenar e analisar dados georreferenciados, levando à descoberta de padrões de distribuição geográfica, que apoiam os responsáveis por tomadas de decisão.

Segundo Hill (2009), o escopo do georreferenciamento inclui os meios informais de referência a locais, como nomes de lugares, e as representações formais baseadas em coordenadas de longitude e latitude, além de outros sistemas de referência espacial, que são usados em atividades como cartografia e navegação.

Ainda de acordo com Hill (2009), a maneira mais comum de georreferenciar pontos no planeta, é com o uso de latitude e longitude. A referência primária para as latitudes é a Linha do Equador, enquanto para as longitudes é o Meridiano de Greenwich. Este meridiano se estende de polo a polo, passando pelo local original do Observatório Real de Greenwich no Reino Unido. As latitudes têm um intervalo de valores de até 90°, abrangendo a distância do equador até os pólos, tanto ao norte quanto ao sul. As longitudes, por sua vez, têm um intervalo de 180°, circundando o globo a partir do meridiano de Greenwich em direção leste e oeste.

O georreferenciamento desempenha um papel fundamental no presente trabalho, pois fornece um contexto espacial valioso, ao permitir a visualização dos preços dos imóveis em um mapa e possibilitando a identificação de padrões geográficos de distribuição de preços, o que pode revelar desde a distribuição espacial dos anúncios até o conhecimento de áreas de alta e baixa valorização imobiliária.

2.2.1. H3

A ferramenta H3, criada pela Uber Technologies (2018), representa um avanço na indexação geoespacial, ao possibilitar a segmentação de regiões geográficas em hexágonos, ampliando a precisão da análise de dados. Essa ferramenta foi criada para combinar os benefícios de um sistema de grade global hexagonal com um sistema de indexação hierárquica. Os hexágonos foram uma escolha importante porque permitem aproximar raios facilmente.

Um sistema de grade global envolve uma projeção cartográfica para mapear a Terra em duas dimensões e uma sobreposição de grade no mapa resultante. Diferentes projeções e grades, como a projeção de Mercator com uma grade quadrada, podem ser combinados para criar esse sistema. Entretanto, essa abordagem tem limitações, como a distorção de tamanho na projeção de Mercator e as desvantagens das grades quadradas, que possuem dois tipos de vizinhos (arestas e vértices), resultando na necessidade de múltiplos conjuntos de coeficientes para análise.

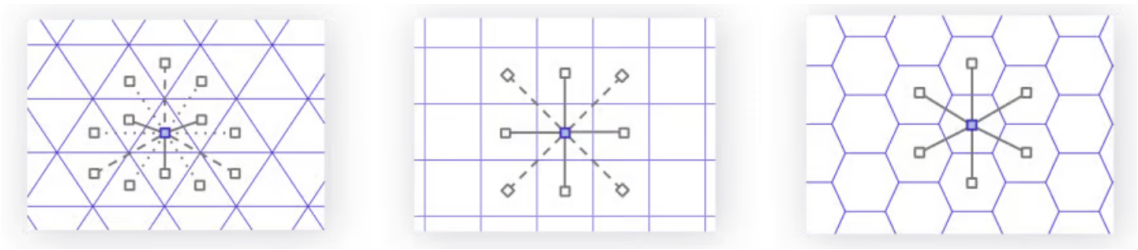

Figura 3. Representação das distâncias das formas geométricas aos seus vizinhos. Fonte: Uber Technologies (2018).

A figura 3 representa as distâncias de um triângulo aos seus vizinhos (esquerda), de um quadrado aos seus vizinhos (centro) e de um hexágono aos seus vizinhos (direita), conforme ilustrado na figura, hexágonos exibem uma notável singularidade: a distância entre o ponto central de um hexágono e aqueles de seus vizinhos é reduzida a apenas uma única medida, em contraste com as duas distâncias necessárias em quadrados ou as três em triângulos. Essa característica simplificada desempenha um papel fundamental na facilitação da realização de análises e na suavização dos gradientes, aprimorando significativamente a aplicabilidade do H3.

Segundo Uber Technologies (2018), o H3 suporta 16 resoluções diferentes, que determinam o tamanho do hexágono. A cada incremento na resolução, as células hexagonais se subdividem em células menores, apresentando uma área aproximadamente sete vezes menor do que as células na resolução anterior. A estrutura hierárquica das células permite uma eficiente indexação espacial, possibilitando a representação de locais geográficos em diferentes níveis de detalhe.

2.3. Aprendizado de Máquina

Aprendizado de máquina, ou machine learning, é um subcampo da engenharia e da ciência da computação que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial.

Segundo Hall et al. (2016), historicamente o conhecimento de aprendizado de máquinas e a estatística possuem uma relação muito próxima, padrão e métodos estatísticos se aplicam como na visualização de dados, na seleção de atributos, no descarte de outliers. Muitos algoritmos de aprendizado usam testes estatísticos ao construir regras e corrigir modelos. Testes estatísticos são usados para validar modelos e avaliar algoritmos de aprendizado de máquina. Hall et al. (2016) comprova que as técnicas de aprendizado de máquina possuem uma grande quantidade de pensamento estatístico.

Segundo Murphy (2012), o aprendizado de máquina geralmente é dividido em três tipos. Na modalidade preditiva ou abordagem de aprendizagem supervisionada, o objetivo é aprender um mapeamento de entradas x para saídas y, dado um rotulado conjunto de pares de entrada-saída (1), onde D é chamado de conjunto de treinamento e N é o número de exemplos de treinamento.

$$D = \{Xi\}_{i=1}^{N} \tag{1}$$

No aprendizado descritivo ou abordagem não supervisionada são recebidas apenas entradas, com o objetivo de encontrar "interessantes padrões" nos dados. Neste caso, não é dito que tipos de padrões procurar e não há métrica de erro a ser usada, diferente do aprendizado supervisionado, onde podemos comparar nossa previsão de y para um dado x ao valor observado (Murphy, 2012).

O terceiro tipo, o aprendizado por reforço, é uma abordagem do aprendizado de máquina em que, segundo Murphy (2012), um agente interage com um ambiente para aprender a tomar ações que maximizem uma recompensa cumulativa ao longo do tempo. De acordo com Sutton e Barto (2018), o aprendizado por reforço é particularmente adequado para situações em que não se dispõe de um conjunto de dados de treinamento supervisionado, mas o agente pode explorar e aprender com sua própria interação com o ambiente. Isso o torna uma abordagem poderosa para a automação de tarefas complexas e a tomada de decisões em tempo real.

A área de aprendizado de máquina abrange uma ampla gama de modelos e algoritmos, cada modelo tem suas vantagens e limitações, e a escolha do modelo certo depende da natureza da tarefa e dos dados em questão. A quantidade de modelos de aprendizado de máquina existentes é vasta e com o avanço contínuo de pesquisas na área, novas abordagens são desenvolvidas.

Os modelos de regressão são uma categoria de algoritmos que se enquadram no campo do aprendizado supervisionado. Segundo Murphy (2012), esses modelos são projetados para lidar com tarefas em que a saída desejada é um valor numérico contínuo, ou seja, a previsão é uma estimativa numérica ao invés de uma classificação em categorias.

2.3.1. Modelos de Regressão

A regressão para predição é uma técnica estatística que visa prever um valor de saída com base em um ou mais valores de entrada, importante ferramenta utilizada para fazer previsões com base em dados históricos e é amplamente utilizada em ciência de dados, análise estatística e aprendizado de máquina para modelar relacionamentos entre

variáveis e fazer previsões, em diversas áreas do conhecimento.

Segundo Montgomery et al. (2012), essa técnica envolve a criação de um modelo matemático que descreve a relação entre uma variável dependente (a variável que você deseja prever) e uma ou mais variáveis independentes (as variáveis que você usa para fazer a previsão). O modelo de regressão busca encontrar os melhores coeficientes para as variáveis independentes, de modo que a soma ponderada delas se aproxime o máximo possível do valor da variável dependente.

2.3.1.1. Random Forest Regressor

O Random Forest Regressor é um algoritmo de aprendizado de máquina baseado em árvores de decisão. Esse algoritmo faz parte da família de algoritmos conhecidos como "ensemble learning", que combinam as previsões de vários modelos para melhorar o desempenho e a generalização. O Random Forest é altamente versátil e eficaz em uma variedade de cenários, desde problemas de regressão até classificação e detecção de anomalias.

Segundo Breiman (2001), uma floresta aleatória é composta por várias árvores de decisão individuais, onde cada árvore é construída usando uma técnica chamada bootstrapping, em que amostras aleatórias com substituição são extraídas do conjunto de treinamento. Durante a construção de cada árvore, a seleção do melhor recurso de divisão é restrita a um subconjunto aleatório de recursos. Essas técnicas introduzem aleatoriedade nas árvores individuais, o que ajuda a evitar o overfitting e torna a floresta mais robusta.

Para prever a saída em um Random Forest Regressor, Breiman (2001) afirma que cada árvore individual prevê um valor e, em seguida, a média (ou outra combinação, dependendo do problema) das previsões de todas as árvores que é considerada como a previsão final. Essa abordagem, também conhecida como agregação de previsões, reduz a variância e melhora o desempenho geral do modelo.

De acordo com Breiman (2001), a construção de múltiplas árvores e o uso de amostragem aleatória de recursos ajudam a lidar com ruído e características irrelevantes nos dados, permitindo que o modelo generalize bem para novos dados.

Esse algoritmo possui vários parâmetros ajustáveis, como o número de árvores na floresta, a profundidade máxima das árvores individuais e outros parâmetros que afetam a construção das árvores. É importante destacar que a seleção adequada desses parâmetros é essencial para evitar o overfitting e obter um desempenho ideal do modelo.

O Random Forest Regressor é amplamente usado em diversas aplicações, como previsão de preços de imóveis, análise financeira e previsão de vendas. Sua capacidade de lidar com dados complexos e robustez, além de facilidade de uso, tornam-no uma escolha popular entre os algoritmos de regressão.

2.3.1.2. Extreme Gradient Boosting

O Extreme Gradient Boosting (XGBoost) é um algoritmo avançado de aprendizado de máquina que pertence à família de métodos de boosting, amplamente utilizado para problemas de regressão, classificação e detecção de anomalias.

O XGBoost, segundo Chen e Guestrin (2016), é conhecido por sua capacidade de lidar com dados complexos, grande quantidade de características e sobreajuste. Além dessas vantagens, o modelo possui ainda uma funcionalidade de validação cruzada incorporada que permite avaliar o desempenho do modelo ao longo do processo de treinamento e ajustar os hiperparâmetros de acordo com a necessidade.

De acordo com Freund e Schapire (1997), o método de Boosting é uma técnica de aprendizado de máquina que combina múltiplos modelos fracos para criar um modelo forte e de melhor desempenho, que envolve treinar os modelos fracos em sequência, onde cada modelo subsequente é ajustado para corrigir os erros do modelo anterior. Em outras palavras, o foco é dado às instâncias que foram classificadas ou previstas incorretamente pelo modelo anterior. O resultado é um modelo final que é capaz de realizar previsões precisas, mesmo que os modelos fracos individualmente não sejam tão bons.

Segundo Chen e Guestrin (2016), o XGBoost incorpora várias técnicas de regularização para evitar o overfitting, como penalização L1 (Lasso) e L2 (Ridge) nos termos das funções de custo. Ele também tem uma funcionalidade chamada ”pruning”que permite limitar a profundidade das árvores individuais, o que é crucial para evitar que as árvores se tornem excessivamente complexas. Os autores afirmam que o algoritmo permite ajustar tanto a profundidade quanto a largura das árvores. Isso oferece uma flexibilidade adicional para capturar padrões complexos em diferentes conjuntos de dados.

2.3.1.3. K Neighbors Regressor

O K Neighbors Regressor é um algoritmo que faz parte da família de algoritmos baseados em instância, ou seja, as previsões são feitas com base nas instâncias mais próximas do conjunto de treinamento. Esse modelo é uma versão regressora do algoritmo K Nearest Neighbors (KNN), que é amplamente utilizado para classificação. O algoritmo possui uma abordagem simples, mas pode ser eficaz em problemas de regressão quando há padrões locais claros nos dados.

Segundo Scikit-learn (2021), o modelo faz previsões para um ponto de dados desconhecido calculando a média (ou outra medida de tendência central) dos valores-alvo das k instâncias mais próximas desse ponto no conjunto de treinamento. A proximidade é geralmente medida usando a distância euclidiana, isto é, a distância entre dois pontos em um espaço euclidiano.

O parâmetro k é crucial no algoritmo K Neighbors Regressor porque, de acordo com Scikit-learn (2021), isso irá determinar quantas instâncias vizinhas serão consideradas para fazer a previsão. Um valor muito pequeno de k pode levar a previsões instáveis, enquanto um valor muito grande pode suavizar demais as previsões e ignorar padrões locais. Esse modelo permite atribuir pesos diferentes aos vizinhos com base na sua proximidade. Isso significa que as instâncias mais próximas terão um impacto maior na previsão do que as instâncias mais distantes.

É importante ressaltar que, conforme Scikit-learn (2021), o desempenho do K Neighbors Regressor pode ser afetado pela dimensionalidade dos dados pois, em espaços de alta dimensionalidade, o conceito de ”vizinhos próximos”pode se tornar menos significativo. Outra consideração dos autores é que, devido à dependência da distância Euclidi-

ana, é importante escalar os dados antes de aplicar o K Neighbors Regressor para garantir que todas as características tenham o mesmo impacto na medida de distância.

2.3.1.4. Extra Trees Regressor

O algoritmo de árvores extremamente aleatórias (Extra Trees Regressor) é uma técnica de aprendizado de máquina relativamente recente. De acordo com Ahmad et al. (2018) esse algoritmo emprega o mesmo princípio da floresta aleatória e usa um subconjunto aleatório de recursos para treinar cada estimador base. No entanto, ele seleciona aleatoriamente o melhor recurso junto com o valor correspondente para dividir o nó. O Extra Trees Regressor usa todo o conjunto de dados de treinamento para treinar cada árvore de regressão, por outro lado, usa uma réplica de bootstrap para treinar o modelo.

2.3.1.5. Gradient Boosting Regressor

O modelo Gradient Boosting Regressor é uma técnica de aprendizado de máquina que pertence à categoria de algoritmos de conjunto (ensemble), amplamente utilizado para tarefas de regressão, onde o objetivo é prever um valor numérico com base em um conjunto de características de entrada. Esse modelo é uma extensão do algoritmo Gradient Boosting.

A principal ideia por trás do Gradient Boosting Regressor é construir um modelo de regressão forte a partir de modelos de regressão fracos, de forma iterativa, onde novos modelos são treinados para corrigir os erros cometidos pelos modelos anteriores. Segundo Friedman (2001), o processo pode ser resumido da seguinte maneira:

- Inicialização: Um modelo de regressão fraco, geralmente uma árvore de decisão rasa, é treinado com os dados de treinamento. Isso serve como ponto de partida;
- Etapas iterativas: Em cada iteração, um novo modelo de regressão fraco é treinado com base nos resíduos (diferença entre os valores reais e as previsões atuais) do modelo anterior. O novo modelo é ajustado para minimizar esses resíduos; e
- Agregação: Os modelos fracos são combinados de forma ponderada para criar o modelo final. Os pesos são atribuídos com base no desempenho de cada modelo fraco, de modo que modelos que performam melhor têm mais influência no modelo final.

O Gradient Boosting Regressor utiliza uma função de perda específica, como o erro quadrático médio (MSE) ou o erro absoluto médio (MAE), como afirma Friedman (2001), para medir o quão bem os resíduos estão sendo ajustados a cada iteração. Segundo o autor, esse modelo também utiliza um hiperparâmetro chamado ”taxa de aprendizado”(learning rate) para controlar o tamanho das correções feitas a cada iteração. Um valor menor de taxa de aprendizado torna o processo mais robusto, mas requer mais iterações para alcançar um desempenho ótimo.

De acordo com Friedman (2001), o Gradient Boosting Regressor possui algumas desvantagens, como a sensibilidade aos hiperparâmetros e o risco de overfitting se não for ajustado corretamente. Em contrapartida, esse modelo de regressão tem as seguintes vantagens:

- Alta precisão: É capaz de criar modelos altamente precisos, muitas vezes superando outros algoritmos de regressão;
- Robustez: Lida bem com outliers e dados ruidosos, graças à sua abordagem de correção iterativa; e
- Flexibilidade: Pode ser combinado com diferentes funções de perda e modelos fracos, como árvores de decisão, tornando-o adaptável a uma variedade de problemas.

2.3.1.6. Decision Tree Regressor

A ideia básica do algoritmo da árvore de decisão ou Decision Tree Regressor (DT) é dividir um problema complexo em vários problemas mais simples, a fim de chegar a uma solução que seja mais fácil de interpretar.

Segundo Ahmad et al. (2018), o modelo representa um conjunto de condições, que são hierarquicamente organizadas e aplicadas sucessivamente da raiz à folha da árvore. Essas árvores produzem um modelo treinado para representar regras lógicas, que podem então ser usados para prever um novo conjunto de dados por meio do processo repetitivo de divisão.

De acordo com Breiman (2001), em um método de árvore de decisão, os recursos dos dados são referidos como variáveis preditoras, enquanto a classe a ser mapeada é a variável de destino. Para problemas de regressão, as variáveis alvo são contínuas.

Para treinar um modelo de árvores de decisão, Ahmad et al. (2018) afirmam que são realizados particionamento recursivo e regressões múltiplas a partir do conjunto de dados de treinamento. A partir do nó raiz da árvore, o processo de divisão de dados em cada nó interno de uma regra da árvore é repetido até que o critério de parada seja atendido. No algoritmo DT, cada nó folha da árvore contém um modelo de regressão simples, que se aplica apenas a essa folha. Após o processo de indução, a poda pode ser aplicada para melhorar a capacidade de generalização do modelo, reduzindo a complexidade da árvore.

Ahmad et al. (2018) ressaltam que a árvore de decisão é apenas para fins de demonstração e o DT real usado na análise é mais complexo (ou seja, mais do que dois recursos são considerados ao procurar a melhor divisão e a árvore é mais profunda).

2.3.2. AutoML

A quantidade de dados disponíveis e de informações geradas automaticamente por big data sobre demografia, comportamentos e necessidades do usuário, é cada vez maior, o que gera interesse em empresas, de todos os domínios, para melhorar seus produtos e serviços e com isso, surge, consequentemente, a necessidade de exploração desses dados.

O rápido crescimento da infraestrutura do aprendizado de máquina e a computação de alto desempenho contribuíram significativamente para esse processo, por este motivo, segundo Karmaker et al. (2021), as empresas estão contratando cada vez mais cientistas de dados engenheiros de aprendizado de máquina para entender melhor seus dados, enquanto a indústria e a academia fazem das pesquisas de aprendizado de máquina e inteligência

artificial suas prioridades.

À medida que essas partes interessadas tentam aproveitar ao máximo seus dados, a demanda por ferramentas de aprendizado de máquina estimula os pesquisadores a explorar as possibilidades do aprendizado de máquina automatizado, como afirma Karmaker et al. (2021).

O AutoML é uma abordagem que utiliza técnicas de automação para otimizar todo o processo de construção e seleção de modelos de aprendizado de máquina, desde o pré-processamento dos dados até o ajuste dos hiperparâmetros. Ele visa simplificar e acelerar a criação de modelos de alta qualidade, permitindo que mesmo pessoas sem experiência profunda em aprendizado de máquina possam aproveitar os benefícios dessa tecnologia. O funcionamento do AutoML pode ser dividido em várias etapas, como:

MÉTODO	**DESCRIÇÃO**
Pré-processamento de Dados	Nessa etapa, os dados brutos são limpos, tratados e transformados em um formato adequado para a construção do modelo. Isso pode envolver a remoção de valores ausentes, a normalização dos dados e a codificação de variáveis categóricas.
Seleção de Algoritmos	O AutoML seleciona automaticamente os algoritmos de aprendizado de máquina mais adequados para o conjunto de dados em questão. Isso é feito com base em características do conjunto de dados, como o tipo de dados (tabulares, imagens, texto), o tamanho do conjunto de dados e outras propriedades.
Extração de Recursos (Feature Engineering)	Nesta etapa, o AutoML pode realizar automaticamente a criação de novas características ou a seleção das características mais relevantes para o modelo.
Ajuste de Hiperparâmetros	Essa é uma das etapas mais importantes do processo, onde os hiperparâmetros do modelo são ajustados para otimizar o desempenho. O AutoML utiliza técnicas como busca aleatória ou busca em grade para explorar diferentes combinações de hiperparâmetros e encontrar a que leva ao melhor desempenho do modelo.
Avaliação e Validação	Durante todo o processo, a avaliação do desempenho do modelo é realizada usando técnicas de validação cruzada ou separação do conjunto de dados em treinamento, validação e teste. Isso garante que o modelo seja avaliado de maneira robusta e evita problemas como overfitting.
Seleção do Melhor Modelo	Ao final do processo, o AutoML seleciona automaticamente o modelo que teve o melhor desempenho de acordo com a métrica escolhida, que pode ser a acurácia, a precisão, o F1-score, entre outras.

Tabela 2. Etapas do funcionamento do AutoML. Fonte: He et al. (2021); Hutter et al. (2019); Feurer et al. (2015).

As ferramentas de AutoML (Automated Machine Learning) visam tornar o aprendizado de máquina acessível para não especialistas, melhorar a eficiência do aprendizado

de máquina e acelerar as pesquisas sobre o assunto. Entretanto, Karmaker et al. (2021) afirmam que, embora a automação e a eficiência sejam os principais pontos do AutoML, esse processo ainda requer envolvimento humano em vários pontos importantes como:

- na compreensão dos atributos de dados específicos do domínio;
- na definição de problemas de previsão;
- na criação de um conjunto de dados de treinamento adequado; e
- na seleção de um modelo que melhor atende a situação.

2.3.3. Avaliação de Modelos de Aprendizagem de Máquinas

A avaliação dos modelos é uma importante etapa no desenvolvimento de algoritmos de aprendizado de máquina, ela permite medir o desempenho e a eficácia de um modelo em relação aos dados de teste ou dados não vistos.

2.3.3.1. Métricas de Avaliação

Diversas métricas e técnicas são empregadas para avaliar a qualidade das previsões e a capacidade de generalização do modelo. Segundo James et al. (2013), o Erro Médio Absoluto (MAE) e Erro Médio Absoluto Percentual (MAPE) são métricas usadas em problemas de regressão para medir o quão próximas as previsões estão dos valores reais.

O Erro Médio Absoluto (MAE) é usado para medir a precisão de um modelo de previsão de valores numéricos. Segundo Hastie et al. (2009), ele calcula a média das diferenças absolutas entre as previsões do modelo e os valores reais (rótulos) nos dados de teste. O MAE fornece uma visão direta do quão distantes as previsões do modelo estão dos valores reais, independentemente da direção do erro (positivo ou negativo).

O Erro Médio Absoluto Percentual (MAPE) é uma métrica de avaliação que expressa o erro médio absoluto como uma porcentagem da média dos valores reais. Segundo Hyndman e Koehler (2006), ele mede a magnitude do erro relativo em relação ao tamanho dos valores reais, tornando-se útil para interpretar o erro em termos proporcionais. No entanto, é importante notar que o MAPE pode ser sensível a divisores próximos a zero.

Essas métricas são amplamente utilizadas para avaliar o desempenho de modelos de previsão e regressão, permitindo comparar diferentes abordagens e escolher o modelo mais adequado para a tarefa em questão.

2.3.3.2. Validação Cruzada (Cross-Validation)

No AutoML, o uso da validação cruzada é essencial para a avaliação de modelos de aprendizagem de máquinas. Segundo Kohavi e Provost (1998) a validação cruzada é usada para estimar a precisão de um indutor dividindo os dados em k subconjuntos (ou dobras) mutuamente exclusivos de tamanho aproximadamente igual. O indutor é treinado e testado k vezes, cada vez que é treinado no conjunto de dados menos um subconjunto e testado nesse subconjuntos, assim a estimativa de precisão é a precisão média para o k subconjuntos.

Conforme Kohavi e Provost (1998), o número de subconjuntos pode variar de acordo com a quantidade de dados disponíveis e a natureza do problema. No final do processo de treinamento, a validação cruzada é usada para comparar o desempenho dos diferentes modelos gerados. O modelo que apresentar o melhor desempenho médio em todos os subconjuntos da validação cruzada é geralmente selecionado como o modelo final.

2.3.3.3. Validação em Conjunto de Teste Separado

A divisão de conjuntos de dados em treino, teste e validação é uma prática essencial no campo de aprendizado de máquina para avaliar, ajustar e selecionar modelos de forma adequada. Essa abordagem ajuda a garantir que os modelos sejam capazes de generalizar bem para dados não vistos e evitar problemas de overfitting. Esses dados são divididos conforme descritos na Tabela 3, que apresenta os conjuntos de dados e sua descrição.

CONJUNTO DE DADOS	DESCRIÇÃO
Dados de treinamento	Usado para treinar o modelo.
Dados de validação	Usado para comparação de diferentes modelos e hiperparâmetros.
Dados de teste	Usado para comprovar que aquele modelo realmente funciona.

Tabela 3. Conjunto de dados usados na aprendizagem supervisionada. Fonte: Suniga (2020).

Os dados de treinamento são usados para alimentar o modelo durante a fase de construção ou treinamento. Segundo Hastie et al. (2009), o modelo ajusta seus parâmetros com base nessas informações para aprender as relações entre as entradas (variáveis independentes) e as saídas (variáveis dependentes). Um conjunto de treinamento deve ser representativo e abranger a variabilidade dos dados do mundo real, pois influenciam diretamente a capacidade do modelo de generalizar para novos dados.

Os dados de validação são usados para ajustar hiperparâmetros e tomar decisões sobre a arquitetura e configuração do modelo. Goodfellow et al. (2016) afirmam que eles ajudam a evitar o overfitting, onde o modelo se ajusta excessivamente aos dados de treinamento.

Segundo Goodfellow et al. (2016), o overfitting ocorre quando um modelo se ajusta aos dados de treinamento com tanta precisão que captura até mesmo os ruídos e variações aleatórias presentes nesses dados e como resultado, o modelo se torna altamente especializado nos dados de treinamento, mas não consegue generalizar bem para novos dados que não foram usados durante o treinamento. Ainda segundo os autores, o overfitting é um problema sério porque prejudica a capacidade do modelo de fazer previsões precisas e úteis, levando a resultados com desempenho ruim em dados de teste ou em situações do mundo real.

Os dados de teste são usados para avaliar o desempenho final do modelo. De acordo com Géron (2022), eles não devem ser usados durante o processo de treinamento

ou validação, garantindo uma avaliação imparcial. Os resultados nos dados de teste fornecem uma estimativa de como o modelo poderá generalizar novos dados do mundo real. A escolha de um conjunto de teste independente é crucial para evitar vieses.

2.3.3.4. Regularização e Ajuste de Hiperparâmetros

A busca aleatória é uma abordagem amplamente usada no contexto do AutoML para otimizar os hiperparâmetros dos modelos de aprendizado de máquina e encontrar configurações que levem a melhores resultados.

De acordo com Bergstra e Bengio (2012), esse processo busca identificar, de forma eficaz, combinações de hiperparâmetros que otimizem o desempenho dos modelos de aprendizado de máquina, funcionando conforme pode ser visto na Tabela 4, que apresenta as etapas da busca aleatória.

ETAPAS	DESCRIÇÃO
Definição do Espaço de Hiperparâmetros	Inicialmente, o espaço contendo os hiperparâmetros é delineado.
Amostragem Aleatória	A seleção de conjuntos de hiperparâmetros é feita através de amostragem aleatória dentro desse espaço.
Avaliação do Desempenho	Cada conjunto é empregado para treinar um modelo e avaliar seu desempenho por meio de métricas específicas.
Iteração	Através de iterações, múltiplas combinações de hiperparâmetros são testadas, refinando gradualmente a busca.
Seleção do Melhor Conjunto	Ao término das iterações, o conjunto que demonstrou o melhor desempenho é escolhido como a configuração final.

Tabela 4. Funcionamento da busca aleatória (random search). Fonte: Bergstra e Bengio (2012).

Segundo Bergstra e Bengio (2012), a busca aleatória tem a vantagem de ser simples de implementar e de não estar restrita a padrões pré-concebidos, permitindo explorar uma ampla variedade de combinações de hiperparâmetros; no entanto, essa abordagem também pode ser limitada em encontrar conjuntos de hiperparâmetros ótimos para problemas complexos.

3. Trabalhos Relacionados

Pai e Wang (2020) usaram 4 modelos de aprendizado de máquina: least squares support vector regression (LSSVR), classification and regression tree (CART), general regression neural networks (GRNN) e backpropagation neural networks (BPNN) para previsão de preços imobiliários, além de utilizarem algoritmos genéticos para selecionar os parâmetros para os modelos usados. Em seu trabalho, o LSSVR superou os outros 3 modelos em termos de precisão, e foram melhores, inclusive do que os estudos relacionados anteriormente, em termos de erro percentual médio absoluto. É importante destacar, no entanto, que esses resultados baseiam-se em dados de transações imobiliárias realizadas em Taiwan, que estão inseridos em um contexto com características muito distintas em relação aos municípios brasileiros.

No Brasil, Isaac et al. (2022) criaram um modelo de precificação de aluguel baseado nas características do imóvel pelo método de Mínimos Quadrados Ordinários (MQO) e utilizaram modelos de Aprendizado de Máquina por meio de AutoML, como o CatBoost Regressor, Extra Trees Regressor e Light Gradient Boosting Machine, onde o Catboost se mostrou mais eficiente, apresentando melhores resultados. Os autores usaram dados minerados do site QuintoAndar, de imóveis para aluguéis da cidade de Belo Horizonte.

A fim de conhecer melhor o comportamento de precificação de imóveis na mesma região abordada neste estudo, foram conduzidas investigações fundamentadas em informações provenientes da cidade do Recife. Apesar de não apresentarem métricas semelhantes às utilizadas no presente trabalho para que fossem feitas comparações com os resultados de suas pesquisas, Bezerra e Reithler (2022) e Delmiro et al. (2022) mostram análises de predição na avaliação de imóveis por regressão, no contexto da realidade local.

Durante a pandemia do Covid-19, Bezerra e Reithler (2022) realizaram uma análise sobre o comportamento dos valores de venda de imóveis residenciais na cidade de Recife. O modelo apresenta um resultado razoável, no entanto como existem diversas variáveis independentes, tanto macroeconômicas como microeconômicas que influenciam no comportamento dos preços dos imóveis, os autores apontam a necessidade de estudos complementares e afirmam que esse trabalho abre precedente para estudos não apenas no mercado imobiliário da cidade do Recife, como também nas demais cidades do Brasil.

Em Recife, Delmiro et al. (2022), usaram dados de imóveis de Casa Forte para avaliação de imóveis, os autores não utilizaram algoritmos de aprendizado de máquina, fizeram uma análise por regressão linear múltipla pelo método dos mínimos quadrados a partir dos requisitos referenciados na ABNT NBR 14653-2 (2011) para predição de avaliação de imóveis.

4. Abordagem proposta

A pesquisa foi conduzida por meio da coleta de dados via web scraping. Posteriormente, esses dados passaram por processos de tratamento e georreferenciamento, preparando-os para a implementação de modelos de aprendizado de máquina por meio de AutoML, conforme explicado nos próximos tópicos.

Na etapa de coleta de dados, foi empregada a técnica de web scraping, fundamentada no uso das bibliotecas ”requests”e ”beautiful soup”. Essas ferramentas permitiram a obtenção eficiente das informações contidas nas páginas web pertinentes.

O tratamento de dados envolveu inicialmente o uso das bibliotecas ”pandas”e ”numpy”, possibilitando a manipulação e organização das informações de maneira adequada. Em seguida, procedeu-se ao georreferenciamento das informações, por meio da utilização de recursos como a API Nominatim e o sistema de indexação espacial H3, conferindo às informações um contexto geográfico relevante.

A modelagem dos dados e a avaliação dos modelos foram conduzidas empregando uma abordagem de AutoML, através da ferramenta ”pycaret”. Essa metodologia permite a automatização de diversos processos envolvidos na criação, ajuste e avaliação de modelos de aprendizado de máquina.

4.1. Coleta de dados

A coleta de dados foi feita por meio de mineração de dados, pelo método de web scraping, a partir da abordagem que segue os links de páginas da web e usando as ferramentas: requests e beautifulsoup. A biblioteca Requests é responsável por realizar o mapeamento do protocolo HTTP na semântica orientada a objetos do Python. Enquanto a biblioteca BeautifulSoup ajuda a formatar a página HTML, apresentando-nos objetos Python facilmente percorríveis representando estruturas XML, segundo Mitchell (2015).

4.2. Tratamento e Georreferenciamento de dados

O tratamento de dados é uma etapa importante, momento para aplicar processamentos e preparações nos dados que foram obtidos, a fim de torná-los próprios para o uso em modelos de aprendizado de máquina. Neste trabalho foram usadas para a manipulação e tratamento de dados as ferramentas pandas e numpy.

O pandas é um dos pacotes da linguagem Python, largamente utilizado no aprendizado de máquina e inteligência artificial. Segundo Grus (2019), essa ferramenta fornece estruturas para trabalhar com conjuntos de dados, onde é possível fatiar, agrupar e manipular os dados, que junto com o pacote Numpy (que tem seu foco em operações numéricas), facilita o trabalho com arrays que funcionam melhor que nossos vetores de lista, matrizes de listas e muitas funções numéricas para trabalhar com conjuntos de dados.

O georreferenciamento foi utilizado no presente trabalho para transformar as variáveis que identificam a localização dos anúncios, no par de variáveis latitude e longitude. Isso foi feito para mitigar possíveis problemas causados pela informação de localização fornecida pela plataforma, pois imóveis em uma mesma rua podem estar em bairros distintos, assim como imóveis no mesmo bairro podem estar em localidades com características bem distintas.

A biblioteca H3 foi empregada devido à sua capacidade de facilitar o agrupamento geográfico, permitindo a formação de grupos que possuem áreas e distâncias iguais entre si.

4.3. Modelagem e avaliação dos modelos

A escolha de uma ferramenta de AutoML para o presente trabalho foi feita com a intenção de testar vários modelos de aprendizado de máquina e visando acelerar o tempo de modelagem da pesquisa, levando em consideração que o tempo de desenvolvimento do trabalho seria consumido em grande parte na extração e tratamento dos dados.

A PyCaret, tem como parâmetros obrigatórios, data, que corresponde ao dataset utilizado, e target, que corresponde a variável a ser prevista. Além desses, foram utilizados alguns parâmetros opcionais, conforme descritos na tabela 5.

PARÂMETRO	DESCRIÇÃO	VALOR UTILIZADO
session_id	Controla a aleatoriedade do experimento. Isso pode ser usado posteriormente para garantir a reprodução completa do experimento.	123
train_size	Proporção do conjunto de dados a ser utilizado para treinamento e validação.	0.8
max_encoding_ohe	Colunas categóricas com o número de valores únicos menor ou igual a max_encoding_ohe são codificadas usando OneHotEncoding. Se houver mais valores únicos, o encoding_method é utilizado.	0
ordinal_features	Lista de variáveis categóricas que devem ser codificadas de maneira ordinal.	quartos, banheiros e vagas na garagem

Tabela 5. Variáveis opcionais utilizadas no trabalho. Fonte: Ali (2020).

Os modelos de aprendizado de máquina selecionados para o trabalho são todos de regressão, pelo fato do problema do trabalho ser relacionado à previsão de uma variável numérica (dependente) a partir de um conjunto de uma ou mais variáveis independentes.

5. Experimentos

Nesta seção, está detalhado como foi realizada as etapas do trabalho. O que inclui a coleta, o tratamento e georreferenciamento dos dados, assim como a criação e a avaliação de modelos de aprendizado de máquina em uma ferramenta de AutoML.

5.1. Coleta de dados

A coleta dos dados dos anúncios, realizada com métodos de web scraping, seguiu um processo composto por três etapas sequenciais. Primeiramente, executou-se uma busca na página de listagem de anúncios, com o intuito de identificar e coletar os links correspondentes aos anúncios individuais. Em seguida, utilizando esses links, efetuou-se uma nova busca para acessar as páginas individuais de cada anúncio. Por fim, a partir dessas páginas de anúncios, foram extraídos os dados específicos de cada anúncio para posterior análise, incluindo título, descrição, preço, detalhes e endereço.

A coleta de anúncios ocorreu diariamente no período entre 23 de março de 2023 e 7 de maio de 2023, resultando na identificação de um total de 77.092 anúncios únicos. A obtenção das páginas HTML dos anúncios foi realizada por meio da biblioteca ”Requests”, enquanto a transformação dessas páginas em objetos manipuláveis em Python foi realizada mediante a utilização da biblioteca ”Beautiful Soup”.

Cabe ressaltar que, durante essa etapa, foi encontrada uma limitação técnica na plataforma OLX, que limitava a possibilidade de acessar páginas a partir do número 100, via URL. Essa restrição somente se fazia contornável por meio de navegação convencional. Consequentemente, optou-se por adotar uma estratégia de busca baseada na segmentação dos imóveis por faixas de preço. Após uma série de experimentações,

definiu-se o montante de R$50.000 como o mais apropriado para o comprimento dessas faixas, de modo a garantir que todas as faixas retornem no máximo 100 páginas.

Para efetuar a busca nos limites das faixas de preço predefinidas, foram empregados parâmetros de pesquisa específicos, nomeadas de "ps" e "pe" para indicar os valores mínimo (preço inicial) e máximo (preço final) respectivamente, bem como o parâmetro "o" para denotar o número da página.

5.2. Tratamento e Georreferenciamento de dados

Inicialmente foram removidos os anúncios que possuíam os termos "repasse" e "repasso" dentro do título ou da descrição, para desconsiderar anúncios que não apresentavam o preço total do ímovel, o que é feito na modalidade de repasse.

As variáveis detalhes e endereço, que eram listas de tuplas, foram convertidas para as respectivas variáveis.

As variáveis detalhes e endereço, que eram listas de tuplas, foram convertidas para as respectivas variáveis. Posteriormente as colunas "Detalhes do imóvel" e "Detalhes do condomínio" foram convertidas em variáveis booleanas, para indicar a presença de características no anúncio, indicado na tabela 6 abaixo.

permitido_animais	area_de_servico	academia
salao_de_festas	armarios_na_cozinha	seguranca_24h
piscina	condominio_fechado	portaria
varanda	mobiliado	ar_condicionado
churrasqueira	elevador	quarto_de_servico
armarios_no_quarto	area_murada	porteiro_24h
portao_eletronico		

Tabela 6. Variáveis booleanas.

As variáveis de preço, condomínio, IPTU e área útil foram convertidas para número, pois estavam em formato de texto com o "R$", para as variáveis monetárias, e "m²", para a variável de área.

Ao analisar as variáveis de Condomínio e IPTU, foi percebido que elas apresentavam uma inconsistência muito grande, quando visto o seu maior valor e desvio padrão, como pode ser visto na tabela 7 abaixo. Por esse motivo elas foram removidas da base de dados.

VARIÁVEL	**MAIOR VALOR**	**DESVIO PADRÃO**
Condomínio	R$15.000.000	R$67.999
IPTU	R$11.488.630	R$71.219

Tabela 7. Maior valor e desvio padrão das variáveis Condomínio e IPTU.

Foram removidas também os anúncios de imóveis que estavam em desacordo com o Código de Edificações do Recife (Recife, 1997), onde são especificados número mínimo de banheiros e área mínima do apartamento, necessitando assim de 1 banheiro e, pelo menos, 18m² área, conforme os artigos 48 e 49, respectivamente.

Mesmo após remover esses anúncios ainda era possível identificar que haviam muitos valores absurdos, por isso foi realizado uma investigação nos dados para encontrar um valor que pudesse ser considerado o menor preço aceitável, resultando assim, no valor de R$40.000.

No aprendizado de máquinas, os outliers representam pontos de dados que se afastam consideravelmente do padrão predominante dentro de um determinado conjunto de dados. Estes são valores que prejudicam a análise estatística e o desempenho dos modelos. A identificação e a correção de outliers são cruciais para garantir que os modelos sejam treinados com dados de alta qualidade e produzam previsões confiáveis.

No contexto da análise de dados, Witten e James (2013) afirma que o Z-score, a Distância de Mahalanobis e a distância interquartil, descritas na Tabela 8, são ferramentas estatísticas amplamente utilizadas para detectar os outliers em conjuntos de dados. Cada uma dessas técnicas tem sua aplicação específica e oferece informações valiosas sobre a estrutura subjacente dos dados. Neste trabalho foram testadas essas três técnicas, para remover os registros incorretos da base de dados de anúncios.

TÉCNICA	**DESCRIÇÃO**
Z-score	É uma medida estatística usada para identificar outliers calculando quantas vezes um valor está desviando da média em termos de desvio padrão. Valores de Z-score significativamente altos ou baixos podem indicar a presença de outliers.[1]
Distância de Mahalanobis	Esse método leva em conta a relação entre as variáveis em um conjunto de dados, considerando a covariância entre elas. Ela é usada para quantificar a distância entre um ponto de dados e uma distribuição multivariada. Valores altos de distância de Mahalanobis podem indicar a presença de outliers.[2]
Distância Interquartil	É uma medida de dispersão estatística que se concentra na diferença entre o terceiro quartil ($Q3$) e o primeiro quartil ($Q1$) de um conjunto de dados. Ela é usada para identificar outliers como valores que se encontram significativamente abaixo de $Q1{-}1.5 * IQR$ ou acima de $Q3 + 1.5 * IQR$.[3]

Tabela 8. Técnicas de remoção de outliers. Fonte: [1]Rudin (2011), [2]Chandola et al. (2009) e [3]Tukey et al. (1977).

O tratamento foi realizado considerando um agrupamento ideal, com as variáveis bairro e tipo de imóvel, uma vez que essas variáveis agregam imóveis com características mais similares. Contudo, esse agrupamento ideal resultou em um grupo com apenas um anúncio (o que impediria a cálculo das técnicas citadas) e, a fim de evitar isso, foram criadas alternativas de agrupamentos para que fossem realizados os cálculos e aplicá-los no grupo ideal, resultando nos seguintes agrupamentos dos anúncios:

- por bairro e tipo de imóvel (ideal);
- apenas por bairro; e
- apenas por tipo de imóvel.

Para os casos em que não era viável realizar o tratamento para o agrupamento ideal, uma vez que resultou em um único anúncio, foi explorada a segunda alternativa.

Assim, explorou-se o agrupamento considerando apenas o bairro em questão, independentemente do tipo de imóvel, para calcular os valores limites para este agrupamento, com o intuito de identificar os anúncios discrepantes no grupo ideal.

Nos casos em que essa segunda abordagem ainda resultou em apenas um anúncio, então uma terceira estratégia foi adotada, foi feito um agrupamento levando em consideração apenas o tipo do imóvel, abrangendo todos os bairros da amostra, para calcular os valores limites para este grupo, a fim de identificar os anúncios discrepantes no grupo ideal.

A distância interquartil foi utilizada, por apresentar melhores resultados na remoção de anúncios que estavam em discrepância com a base de dados. A quantidade de registros removidos por este método foi de 7.550 anúncios.

A variável "permitido_animais" diz respeito a permissão do condomínio a animais dentro das áreas do condomínio, porém esta proibição acaba indo de encontro com alguns pontos legais, como o artigo 5º da constituição nos incisos XV e XXII, o artigo 146 do Decreto-lei Nº 2.848/40 e a decisão da 3ª turma do STJ que reconhece que não é permitido a proibição da guarda de um pet pelo condômino.

Para transformar os endereços dos anúncios (textos) em coordenadas geográficas, foi usado a API Nominatim para busca nos dados do OpenStreetMap, que oferece suporte a consultas de pesquisa estruturadas e de formato livre. Neste caso, recebendo uma string (sequência de caracteres, no caso, o endereço) e retornando a latitude e longitude desta posição, em um processo chamado de geocodificação.

Durante o tratamento também foi identificado que há uma grande diferença entre a quantidade de registros localizados no bairro de Boa Viagem, com 38,64% dos anúncios, em relação a outros bairros da cidade do Recife. Essa discrepância causou uma má distribuição dos dados, como pode ser observado na Tabela 9, onde estão os 10 bairros com mais anúncios, o número de anúncios e o percentual que eles representam do total.

BAIRRO	**ANÚNCIOS**	**%**
Boa Viagem	26328	38,64
Madalena	3687	5,41
Pina	2977	4,36
Casa Amarela	2775	4,07
Torre	2639	3,87
Graças	2392	3,51
Imbiribeira	2376	3,48
Boa Vista	2359	3,46
Encruzilhada	2086	3,06
Casa Forte	1803	2,64

Tabela 9. Distribuição dos anúncios nos bairros da cidade do Recife.

5.3. Modelagem e Avaliação dos modelos de machine learning

Para realizar a seleção do modelo, foi necessário antes realizar testes preliminares, com o intuito de identificar quais configurações dos dados iriam auxiliar o modelo a ter

um melhor desempenho. Três configurações foram testadas neste momento:

- Como seria apresentada o hexágono, se por meio do seu identificador hexadecimal ou utilizando duas variáveis numéricas que indicam a distância de um hexágono configurado como origem;
- Se todas as variáveis seriam fornecidas para o modelo ou se um conjunto de variáveis seriam removidos; e
- Qual resolução dos hexágonos do H3 seria utilizada.

Com essas configurações foi possível gerar 4 cenários, que podem ser vistos na Tabela 10 abaixo, que foram testados com diversas resoluções diferentes dos hexágonos do H3.

CENÁRIOS	**DESCRIÇÃO**
Cenário 1	Todas as variáveis e identificador do H3
Cenário 2	Todas as variáveis e coordenadas
Cenário 3	Variáveis removidas e identificador do H3
Cenário 4	Variáveis removidas e coordenadas

Tabela 10. Descrição dos cenários de teste.

Para as resoluções do H3, foram selecionadas 6 resoluções, igualmente espaçadas, para que fosse possível identificar em qual delas o modelo iria ter um melhor resultado. As resoluções selecionadas foram as seguintes: 0, 3, 6, 9, 12 e 15.

Para as variáveis, foram listadas aquelas que poderiam criar algum viés que atrapalhasse o modelo e, por fim, as variáveis removidas da amostra para os cenários 3 e 4 estão presentes na Tabela 11.

vagas_na_garagem	armarios_na_cozinha	categoria
tipo_venda	armarios_no_quarto	ar_condicionado
area_murada		

Tabela 11. Variáveis removidas nos cenários 3 e 4.

Os testes preliminares proporcionaram informações importantes sobre as configurações de dados mais significativas para o modelo, bem como a resolução ótima do H3. A análise indicou que a resolução 12 do H3 se destacou ao apresentar as menores métricas de erro entre todas as resoluções avaliadas. Quanto aos cenários, observou-se que o Cenário 2, que preserva todas as variáveis e utiliza pares de coordenadas gerados a partir dos hexágonos do H3, exibiu o menor MAPE em quase todos os modelos, com exceção do Gradient Bosting Regressor. A Figura 4 exibe os resultados de MAPE para os quatro cenários testados, todos na resolução 12 do H3, oferecendo uma visão abrangente desses achados.

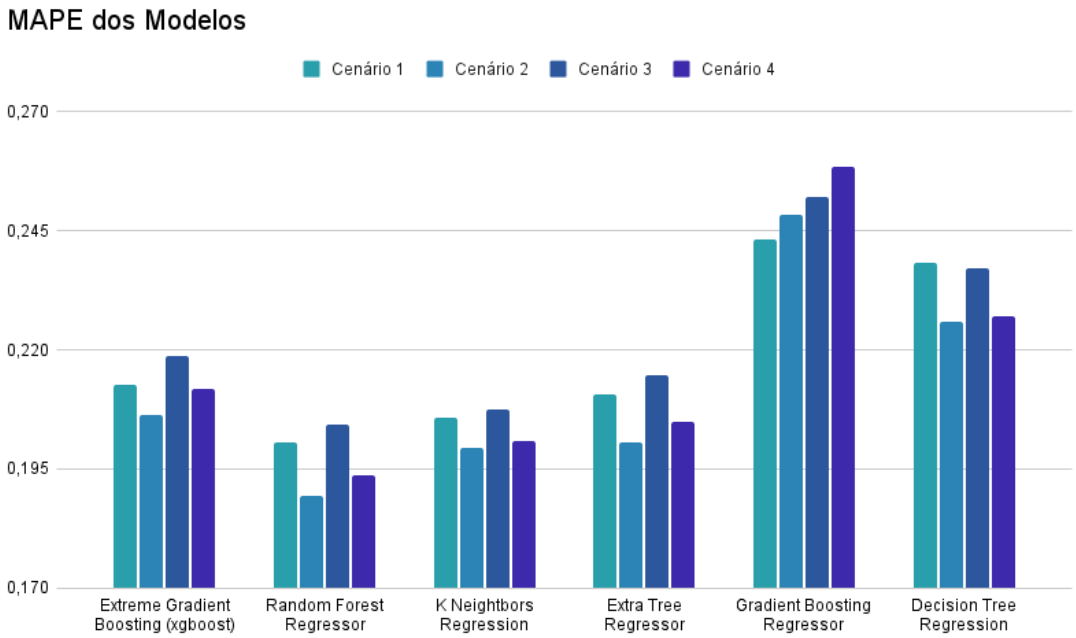

Figura 4. Comparação do MAPE dos modelos pelos cenários, na resolução 12 do H3.

Após a identificação das melhores resolução e configuração dos dados foi realizada uma nova separação entre os dados de treino e teste, com o intuito de garantir que o modelo seria treinado e testado com proporções de anúncios por bairro semelhante. Assim, foram divididos os conjuntos de treino e teste, utilizando o parâmetro stratify do train_test_split do scikit_learn, o que resultou em 70% dos anúncios de cada bairro para o conjunto de treino e 30% dos anúncios de cada bairro para o conjunto de teste. Essa alteração afetou as métricas de erro do modelo, porém garantiu que ele fosse treinado com um conjunto semelhante ao que foi separado para teste.

6. Resultados

Em relação aos modelos de aprendizagem usados para a previsão dos valores dos imóveis, os resultados apontaram o Random Forest como o modelo com a melhor precisão, baseado nas métricas de Erro Médio Absoluto (MAE) e Erro Médio Absoluto Percentual (MAPE), como pode ser visto na Tabela 12

MODELO	MAE	MAPE
Random Forest Regressor	104889.8863	0.1971
Extra Trees Regressor	110104.2614	0.2060
K Neighbors Regressor	109725.5930	0.2082
Extreme Gradient Boosting	114066.8150	0.2141
Decision Tree Regressor	127435.7122	0.2325
Gradient Boosting Regressor	139334.5836	0.2521

Tabela 12. MAE e MAPE dos modelos no melhor cenário de teste.

Avaliando mais a fundo os resultados obtidos pelo Random Forest, foi feita uma análise da distribuição do Erro Médio Absoluto Percentual pelos bairros da município, como pode ser visto na Figura 5, quanto mais escuro o tom de verde, maior o erro (os bairros em vermelho são aqueles que não existem dados).

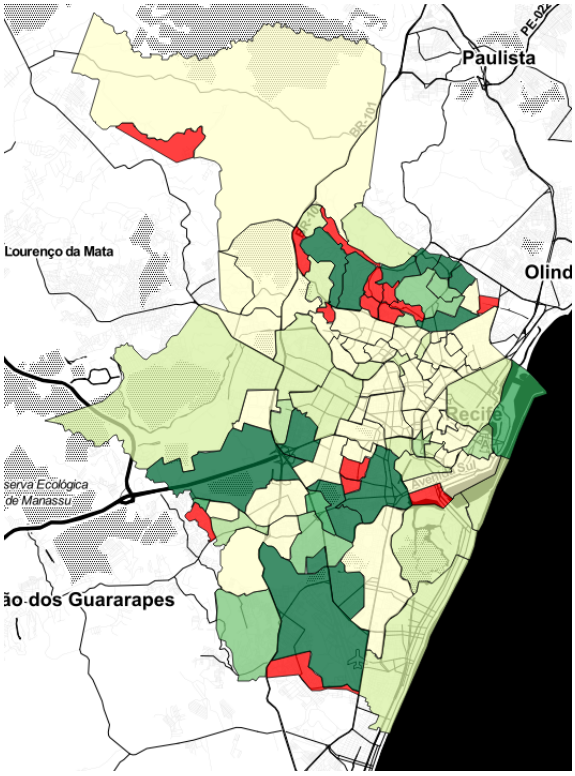

Figura 5. Distribuição do erro percentual médio absoluto nos bairros da cidade do Recife.

Esse resultado foi causado pela má distribuição geográfica dos anúncios, que estavam concentrados principalmente no bairro de Boa Viagem, fazendo com que a generalização do modelo fosse tendenciosa para as características deste bairro. Os bairros que têm características semelhantes a ele foram assimilados corretamente pelo modelo, porém ele não teve capacidade de assimilar os bairros com características mais distintas ao bairro de Boa Viagem.

Mesmo com o desempenho do modelo aquém do esperado, os objetivos do trabalho foram alcançados. O trabalho permitiu a construção de uma base de dados estruturada, com dados de anúncios de apartamentos para venda em Recife, para realização de seu tratamento e posterior uso no treinamento de modelos de aprendizado de máquina.

Ao comparar os resultados dos modelos deste estudo com os resultados dos trabalhos relacionados, notamos que os modelos de Pai e Wang (2020), que empregaram dados reais de transações imobiliárias do município de Taichung, fornecidos pelo governo de Taiwan, apresentaram uma precisão superior. Em contraste, tanto o trabalho de Isaac et al. (2022) quanto o presente estudo utilizaram dados disponíveis publicamente em plataformas online e exibem resultados com um Erro Percentual Absoluto Médio (MAPE) de aproximadamente 20%, como evidenciado na Tabela 13, que compila os modelos de

aprendizado de máquina adotados nas pesquisas e seus respectivos valores de MAPE. Essa diferença nos resultados pode estar relacionada à qualidade dos dados utilizados para a geração dos modelos nos trabalhos citados.

MODELOS DE PREVISÃO	MAPE (%)
Least Squares Support Vector Regression (LSSVR)[1]	0.228
Classification And Regression Tree (CART)[1]	2.278
General Regression Neural Network (GRNN)[1]	8.738
Backpropagation Neural Networks (BPNN)[1]	14.424
CatBoost Regressor[2]	19.03
Extra Trees Regressor[2]	20.26
Light Gradiente Boosting Machine (LGBM)[2]	21.23
Random Forest Regression	19.71

Tabela 13. Comparação dos resultados de modelos de previsão. Fonte: [1]Pai e Wang (2020), [2]Isaac et al. (2022).

No tocante à comparação entre valores dos imóveis buscados (DataZAP, Abril 2021) e os valores dos imóveis coletados neste estudo, observa-se que, enquanto 67% das buscas se concentram em imóveis com valores até R$ 399.999, 62% dos valores anunciados ultrapassam a marca de R$ 400 mil. Essa discrepância é detalhadamente ilustrada na Figura 6, que apresenta uma análise comparativa entre a proporção de imóveis buscados e a proporção de imóveis anunciados, categorizados por faixas de valor.

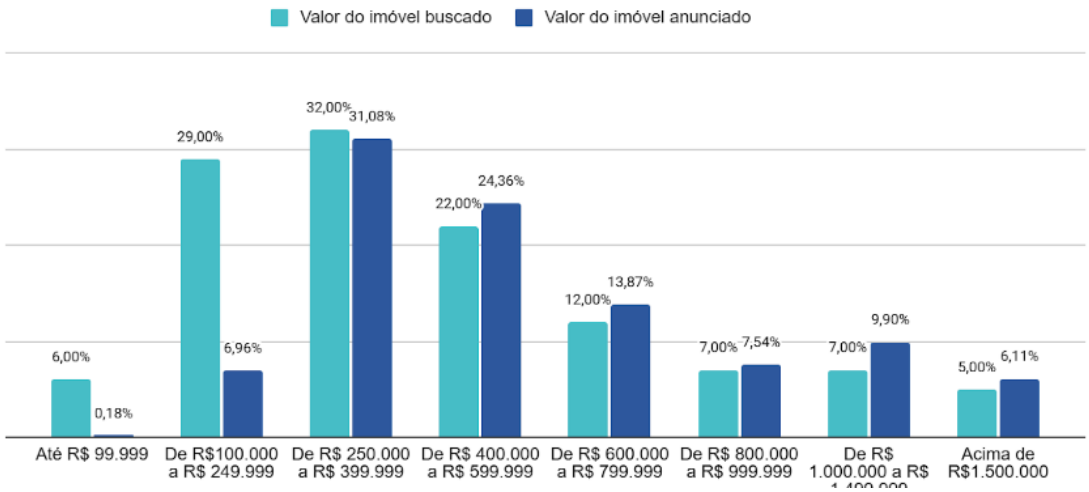

Figura 6. Comparativo entre valor do imóvel buscado e valor anunciado para compra de imóveis. Fonte: DataZAP (Abril 2021).

7. Conclusão

Este trabalho teve como objetivo principal a construção de um modelo de aprendizado de máquina para previsão do valor de imóveis residenciais, do tipo apartamento, no município de Recife. Para alcançar esse objetivo, foram realizadas três etapas específicas: a coleta de dados de apartamentos à venda em Recife, o tratamento desses dados para

criar uma base que pudesse ser usada em um modelo de aprendizado de máquina e, por fim, a criação e teste de um modelo de regressão para estimar os preços dos imóveis.

Os resultados deste trabalho revelaram que o modelo de aprendizagem mais eficaz para a previsão de valores de imóveis foi o Random Forest, apontando as melhores métricas de Erro Médio Absoluto (MAE) e Erro Médio Absoluto Percentual (MAPE), conforme apresentado nos resultados, descritos na seção anterior. No entanto, uma análise mais profunda desse modelo evidenciou que sua performance foi influenciada pela concentração geográfica dos anúncios, principalmente no bairro de Boa Viagem, resultando em uma generalização enviesada para as características desse bairro.

Os resultados obtidos apresentaram modelos com índices de erro relativamente altos, além de dificuldade de generalização para diferentes áreas do município, assim como o trabalho de Isaac et al. (2022). Essas limitações são atribuídas, principalmente, à distribuição geográfica dos dados, que não é proporcional para as regiões do município, e aos valores adotados serem provenientes de anúncios, ao invés de transações efetivas, o que aponta a insuficiência de dados disponíveis em plataformas virtuais de compra e venda, como fonte isolada para construção de um modelo de aprendizado de máquina que ofereça um desempenho significativo.

Apesar de não ter alcançado os resultados desejados nos modelos, o presente trabalho oferece importantes contribuições. Ele destaca, em primeiro lugar, que a utilização exclusiva de dados provenientes de anúncios não é suficiente para criar um modelo altamente generalizável. Além disso, ajuda a revelar que os preços dos imóveis à venda em Recife não seguem um padrão uniforme de demanda local.

A partir das reflexões apresentadas, este estudo fornece valiosas informações para futuras pesquisas relacionadas à previsão de valores imobiliários. Uma abordagem viável compreende o balanceamento geográfico dos dados, objetivando criar um conjunto de dados geograficamente equilibrado e mais representativo. Além disso, seria interessante realizar uma investigação comparativa que avalie as vantagens e desvantagens entre o uso de informações de bairro e hexágonos do H3 como características. Adicionalmente, uma perspectiva importante reside na adoção de dados de transações de imóveis em substituição aos dados de anúncios, proporcionando uma visão mais precisa do mercado.

Quanto à metodologia, um enfoque singular em um modelo específico, em detrimento do uso de AutoML, é uma alternativa a ser explorada. Por último, mas não menos importante, a inclusão de variáveis adicionais, tais como características socioeconômicas dos bairros, densidade populacional e acessibilidade a serviços, tem o potencial de enriquecer os modelos e resultar em previsões mais precisas e contextualmente embasadas.

Uma relevante contribuição do presente trabalho é o compartilhamento dos dados coletados[1]. Ao disponibilizar uma base de dados estruturada, com os dados de anúncios de venda de apartamentos em Recife, extraídos da plataforma OLX, e devidamente documentada, esse estudo possibilita o acesso a dados organizados para o desenvolvimento de futuras pesquisas. Esse estudo fornece, assim, importantes contribuições e sugere várias direções para o avanço das pesquisas em automação na previsão de valores imobiliários.

[1] Repositório com os dados coletados no trabalho https://github.com/ThiagoCMS/tcc-bsi-data

Referências

Muhammad Waseem Ahmad, Jonathan Reynolds, e Yacine Rezgui. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production*, 203:810–821, 2018.

Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. URL `https://www.pycaret.org`. PyCaret version 1.0.0.

Nadia Balemi, Roland Füss, e Alois Weigand. Covid-19's impact on real estate markets: review and outlook. *Financial Markets and Portfolio Management*, pages 1–19, 2021.

Echiley Dias Baptista, Isabella Silva Marcondes Pereira, Jamylle Prado Oliveira, e Ana Paula Freitas Lima. Setor imobiliário: Um estudo de caso sobre o fechamento de contratos durante a pandemia. *Journal of Technology & Information (JTnI)*, 2(4), 2022.

Matheus Costa Barros. Avaliação do valor de imóveis por análise de regressão linear-estudo da taxa de rendimento do mercado imobiliário em fortaleza. 2018.

James Bergstra e Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Suélliton da Rocha Bezerra e Yuri Almeida Reithler. Análise sobre o comportamento dos valores de venda de imóveis residenciais na cidade de recife/pe durante a pandemia do covid-19. B.S. thesis, 2022.

Roberto Boselli, Mirko Cesarini, Fabio Mercorio, e Mario Mezzanzanica. Using machine learning for labour market intelligence. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10*, pages 330–342. Springer, 2017.

Roberto Boselli, Mirko Cesarini, Fabio Mercorio, e Mario Mezzanzanica. Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, 86:319–328, 2018.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Erik Brynjolfsson, Daniel Rock, e Prasanna Tambe. How will machine learning transform the labor market? *Governance in an Emerging New World*, 619, 2019.

Varun Chandola, Arindam Banerjee, e Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Tianqi Chen e Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Rodrigo Brandão dos Santos Cintra. Efeito da pandemia do covid-19 na variação do valor imobiliário na região metropolitana de recife. B.S. thesis, 2021.

Emilio Colombo, Fabio Mercorio, e Mario Mezzanzanica. Applying machine learning tools on web vacancies for labour market and skill analysis. *Terminator or the Jetsons*, 2018.

Rubens Alves Dantas. *Engenharia de Avaliações: uma introdução à metodologia científica*. Pini, 2005.

DataZAP. A influência do coronavírus no mercado imobiliário brasileiro. Abril 2021. Publicação interna.

Thiago Felipe de Véras Delmiro et al. Avaliação de imóveis por análise de regressão linear múltipla no bairro de casa forte, recife/pe. 2022.

Eric Ivantes Dias e Antônio Carlos Magalhães da Silva. Análise do desempenho dos fundos imobiliários no brasil de 2017 a pandemia covid-19. *Revista Vianna Sapiens*, 12(2):22–22, 2021.

Ana Gabrielly Lucas Fernandes, Samara Eduarda Pereira Estigarribia, e Nilson César Bertóli. O cenário da pandemia do novo corona vírus no setor imobiliário: Um estudo de caso em loteamentos na cidade de sertanópolis-pr the scenario of the new corona virus pandemic on the real estate sector: A case study in allotment in the city of sertanópolis-pr. *Brazilian Journal of Development*, 7(12):120680–120695, 2021.

Marcos Sergio Ferreira Wisniewski. *O setor imobiliário português: uma análise evolutiva da última década e possíveis impactos causados pela pandemia COVID-19*. PhD thesis, 2022.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, e Frank Hutter. Efficient and robust automated machine learning. *Advances in neural information processing systems*, 28, 2015.

Sara Andrade Fidelis. Análise dos impactos da pandemia da covid-19 no mercado imobiliário brasileiro durante o período de 2020 a 2022. 2023.

Yoav Freund e Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Fundação Instituto de Pesquisas Econômicas (FIPE). Índice fipezap - fundação instituto de pesquisas econômicas, 2023. URL `https://www.fipe.org.br/pt-br/indices/fipezap#indice-mensal`.

David Geltner, Norman G Miller, Dr Jim Clayton, e Piet Eichholtz. *Commercial real estate analysis and investments*, volume 1. South-western Cincinnati, OH, 2001.

Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.", 2022.

Ian Goodfellow, Yoshua Bengio, e Aaron Courville. *Deep learning*. MIT press, 2016.

Joel Grus. *Data science from scratch: first principles with python*. O'Reilly Media, 2019.

Vítor Costa França Guimarães. A moradia e a pandemia de covid-19. 2022.

Arpit Gupta. Accelerating remote work after covid-19. *The Center for Growth and Opportunity*, 2023.

Mark A. Hall, Eibe Frank, Ian H. Witten, e Christopher J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Morgan Kaufmann, 4rd edition, 2016.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, e Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Xin He, Kaiyong Zhao, e Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

Linda L Hill. *Georeferencing: The geographic associations of information*. Mit Press, 2009.

Frank Hutter, Lars Kotthoff, e Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

Rob J Hyndman e Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Guilherme Costa Isaac, Ari Francisco de Araujo Junior, e Luiz Carlos Day Gama. Estimativa de preços hedônicos para aluguel em belo horizonte através de econometria e machine learning. 2022.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

Marina Jovanović-Milenković, Ana Đurković, Darko Vučetić, e Borko Drašković. The impact of covid-19 pandemic on the real estate market development projects. *European Project Management Journal*, 10(1):36–49, 2020.

Arturas Kaklauskas, Edmundas Kazimieras Zavadskas, Natalija Lepkova, Saulius Raslanas, Kestutis Dauksys, Ingrida Vetloviene, e Ieva Ubarte. Sustainable construction investment, real estate development, and covid-19: a review of literature in the field. *Sustainability*, 13(13):7420, 2021.

Shubhra Kanti Karmaker, Md Mahadi Hassan, Micah J Smith, Lei Xu, Chengxiang Zhai, e Kalyan Veeramachaneni. Automl to date and beyond: Challenges and opportunities. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021.

Ron Kohavi e Foster Provost. Glossary of terms. *Machine Learning*, 30(2):271–274, 1998.

Richard Lawson. *Web Scraping with Python*. Packt Publishing, Birmingham, England, April 2023.

Sitian Liu e Yichen Su. The impact of the covid-19 pandemic on the demand for density: Evidence from the us housing market. *Economics letters*, 207:110010, 2021.

Ryan Mitchell. *Web scraping with python*. O'Reilly Media, June 2015.

Douglas C Montgomery, Elizabeth A Peck, e G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2012.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Jéssica Martins Nunes, Orlando Celso Longo, Luciane Ferreira Alcoforado, e Gustavo Oliveira Pinto. Analysis of covid-19's impacts on the brazilian real estate market. *Research, Society and Development*, 9:e46891211317, Dec. 2020. doi: 10.33448/rsd-v9i12.11317.

Ping-Feng Pai e Wen-Chang Wang. Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, 10:5832, 08 2020. doi: 10.3390/app10175832.

Alexandre Peev. *O impacto da pandemia na precificação imobiliária*. PhD thesis, 2022.

Paulo Pereira et al. Elevação de preços no mercado residencial no brasil: questões estruturais, desempenho do setor e risco do sistema econômico. Technical report, Latin American Real Estate Society (LARES), 2014.

ATD Perera e Parameswaran Kamalaruban. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137:110618, 2021.

Recife. Lei nº 16292 de 29 de janeiro de 1997. 1997. URL `http://www.legiscidade.recife.pe.gov.br/lei/16292/`.

Cynthia Rudin. Chapter 4: Summarizing numerical data - mit opencourseware, 2011. URL `https://ocw.mit.edu/courses/15-075j-statistical-thinking-and-data-analysis-fall-2011/e8d615e72bb6d384e34fc2a10a8f03cb_MIT15_075JF11_chpt04.pdf`.

Scikit-learn. Nearest neighbors regression, 2021. URL `https://scikit-learn.org/stable/modules/neighbors.html#regression`.

Jenni AM Sidey-Gibbons e Chris J Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19:1–18, 2019.

Abner Suniga. Conjuntos de treino, teste e validação em machine learning (fast.ai), 2020. URL `https://medium.com/`

@abnersuniga7/conjuntos-de-treino-teste-e-valida%C3%A7%C3%A3o-em-machine-learning-fast-ai-5da612dcb0ed.

Richard S Sutton e Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

Inc. Uber Technologies. Uber engineering blog: H3, 2018. URL https://www.uber.com/en-BR/blog/h3/.

Tasha Wade, Shelly Sommer, et al. *A to Z GIS, An illustrated dictionary of geographic information systems*. 2006.

Daniela Witten e Gareth James. *An introduction to statistical learning with applications in R*. springer publication, 2013.