ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

LUÍS FERNANDO BITTENCOURT

ALÉM DOS NÚMEROS: UMA PERSPECTIVA MULTIMODAL NA AVALIAÇÃO IMOBILIÁRIA

Porto Alegre
2024

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ALÉM DOS NÚMEROS: UMA PERSPECTIVA MULTIMODAL NA AVALIAÇÃO IMOBILIÁRIA

LUÍS FERNANDO BITTENCOURT

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz

Porto Alegre
2024

Ficha Catalográfica

B624a Bittencourt, Luís Fernando

Além dos números : uma perspectiva multimodal na avaliação imobiliária / Luís Fernando Bittencourt. – 2024.
66 p.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.

1. Avaliação Imobiliária. 2. Aprendizado de Máquina. 3. Multimodal. I. Ruiz, Duncan Dubugras Alcoba. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

LUÍS FERNANDO BITTENCOURT

ALÉM DOS NÚMEROS: UMA PERSPECTIVA MULTIMODAL NA AVALIAÇÃO IMOBILIÁRIA

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 18 de Janeiro de 2024.

BANCA EXAMINADORA:

Profª. Drª. Karin Becker (PPGC/UFRGS)

Prof. Dr. Dalvan Jair Griebler (PPGCC/PUCRS)

Prof. Dr. Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Dedico este trabalho aos meus professores de Ensino Médio da E.E.E.M. Virgilino Jayme Zinn, o CIEP, por sempre terem acreditado que a gente da vila poderia alçar voos mais altos. Cá estamos, a gente da vila, rompendo fronteiras.

“Eu cheguei de muito longe
E a viagem foi tão longa
E na minha caminhada
Obstáculos na estrada
Mas enfim aqui estou
[...]
É preciso dar um jeito, meu amigo”
Erasmo Carlos — É Preciso Dar Um Jeito, Meu Amigo

AGRADECIMENTOS

Concluir um programa de mestrado não é uma tarefa fácil, especialmente quando se tem que conciliar um trabalho em tempo integral. Por isso, fico feliz em reconher e agradecer a todas as pessoas que tornaram essa jornada um pouco mais tranquila. A Cinthia Becker, William Becker, Fabiana Lorenzi e Greice Laureano pelas imprescindíveis cartas de recomendação. A Rodrigo Schaurich da Pmweb, à TAG Livros e a Neal Bram da PWN, por flexibilizarem meus horários de trabalho para eu cumprir meus compromissos acadêmicos. Aos professores do PPGCC, em especial Isabel Manssour, Rodrigo Barros e Soraia Musse, que me deram a honra de serem coautores do meu primeiro artigo publicado. Ao amigo André Giordani, pelo companheirismo que tornou esses três anos mais leves. A Otávio Parraga, por ter sido um parceiro gigante em todos os momentos. Ao querido professor Duncan, meu mentor, por toda a orientação, no sentido mais amplo da palavra. E sobretudo à minha parceira de vida Alana Porto, por ter estado ao meu lado com paciência e amor durante todos esses anos; sem ela, nada disso teria sido possível.

Esta dissertação foi apoiada pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23/10/1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado no Aditivo à Residência em TIC 02, DOU 01245.012095/2020-56.

ALÉM DOS NÚMEROS: UMA PERSPECTIVA MULTIMODAL NA AVALIAÇÃO IMOBILIÁRIA

RESUMO

A avaliação imobiliária é um problema clássico de pesquisa que abrange diversas áreas do conhecimento. Na Ciência da Computação, esse tópico tem sido abordado principalmente pela proposição de Modelos de Avaliação Automática (MAA), que buscam automatizar esse processo. No entanto, a maioria desses modelos baseia suas estimativas apenas nas características estruturais e na localização geográfica dos imóveis, negligenciando fatores cruciais como estado de conservação e posição solar. Para servir como apoio à tomada de decisão, é essencial que um MAA avalie todo o espectro de informações levado em conta por uma pessoa avaliadora, incluindo fotos e textos. Nesse contexto, este trabalho apresenta uma arquitetura multimodal de componentes intercambiáveis que facilita a rápida prototipagem de modelos multimodais. Essa arquitetura foi aplicada a um estudo de caso no Brasil, um país cuja paisagem urbana ainda é pouco explorada na literatura de avaliação imobiliária multimodal. Utilizando dados provenientes de 212.076 anúncios e 1.702.960 imagens, os resultados observados enfatizam a relevância das modalidades de imagem e texto na otimização de desempenho dos MAA. O uso de descrições textuais, em particular, ocasionou reduções significativas na RMSE, variando de 7 a 27% em comparação com dois modelos de linha de base distintos.

Palavras-chave: avaliação imobiliária, aprendizado de máquina, multimodal.

BEYOND THE FIGURES: A MULTIMODAL PERSPECTIVE ON REAL ESTATE VALUATION

ABSTRACT

Real estate valuation is a classical research problem that spans several fields of knowledge. In Computer Science, this topic has been mainly addressed through the proposition of Automated Valuation Models (AVM), which aim to automate this process. However, most of these models base their estimates solely on the structural characteristics and geographical location of the properties, neglecting crucial factors such as condition and solar orientation. To support decision-making, an AVM needs to evaluate the full spectrum of information considered by a human appraiser, including photos and text. In this context, this work presents a multimodal architecture of interchangeable components that facilitates the rapid prototyping of multimodal models. This architecture was applied to a case study in Brazil, a country whose urban landscape is still underexplored in multimodal real estate valuation literature. Using data from 212,076 listings and 1,702,960 images, the observed results emphasize the relevance of image and text modalities in optimizing AVM performance. The use of textual descriptions, in particular, led to significant reductions in RMSE, ranging from 7 to 27% compared to two different baseline models.

Keywords: real estate valuation, machine learning, multimodal.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API – *Application Programming Interface* (Interface de Programação de Aplicação)

AVA – *Aesthetic Visual Analysis*

AVM – *Automated Valuation Model* (Modelo de Avaliação Automática)

AWS – Amazon Web Services

BRNN – *Bidirectional Recurrent Neural Network*

CNN – *Convolutional Neural Network* (Rede Neural Convolucional)

ERT – *Extremely Randomized Trees*

EUA – Estados Unidos da América

GB – Gigabyte

HTML – *HyperText Markup Language* (Linguagem de Marcação de HiperTexto)

IAAO – International Association of Assessing Officers

IBGE – Instituto Brasileiro de Geografia e Estatística

IQR – *Interquartile Range* (Intervalo Interquartil)

LSTM – *Long Short-Term Memory*

MAA – Modelo de Avaliação Automática

MAE – *Mean Absolute Error* (Erro Absoluto Médio)

MAPE – *Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio)

MDAPE – *Median Absolute Percentage Error* (Erro Percentual Absoluto Mediano)

MSE – *Mean Squared Error* (Erro Quadrático Médio)

PCA – *Principal Component Analysis* (Análise de Componentes Principais)

PPGC – Programa de Pós-Graduação em Computação

PPGCC – Programa de Pós-Graduação em Ciência da Computação

PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul

Q1 – Primeiro Quartil

Q3 – Terceiro Quartil

RAM – *Random Access Memory* (Memória de Acesso Aleatório)

RMSE – *Root Mean Squared Error* (Raiz do Erro Quadrático Médio)

SURF – *Speeded Up Robust Features*

SVD – *Singular Value Decomposition*

UFRGS – Universidade Federal do Rio Grande do Sul

URL – *Uniform Resource Locator* (Localizador Uniforme de Recursos)

VGG – *Visual Geometry Group*

SUMÁRIO

1. INTRODUÇÃO

Avaliar corretamente o valor de um imóvel é uma atividade fundamental para diversos setores da sociedade e para o poder público. Definir um preço que reflita adequadamente suas características físicas e os valores associados a sua localização é uma etapa imprescindível em operações de compra e venda, no cálculo de impostos e em laudos de avaliação para fins de seguro, processos judiciais ou pedidos de financiamento. Dada sua importância, não surpreende que a avaliação imobiliária seja um problema clássico de aprendizado de máquina.

Desse modo, existe vasta literatura propondo Modelos de Avaliação Automática (MAA), cujo objetivo é estimar o preço de anúncio ou o valor efetivamente pago por cada imóvel. Esses modelos são capazes de analisar milhares de informações e reconhecer padrões sutis, levando para isso uma fração do tempo que seria necessário se a mesma tarefa fosse feita manualmente por uma pessoa. Como ferramenta de apoio à tomada de decisão, os MAA reduzem tempo, custo e subjetividade da avaliação de imóveis.

A maioria desses modelos baseia suas predições em duas modalidades de informação: localização geográfica e características estruturais, tais como área e número de quartos. Como essas modalidades não abrangem uma série de fatores que afetam o preço de um imóvel, como estado de conservação e posição solar, esses modelos tradicionais podem produzir resultados questionáveis. Por exemplo, eles podem estimar um único preço para todos os apartamentos que compartilham a mesma planta em um edifício, dada a grande probabilidade de que um número tão restrito de variáveis leve a amostras repetidas. Assim, incorporar uma gama maior de características é determinante para que um MAA seja verdadeiramente aplicável ao mundo real.

Na ausência de conjuntos de dados de imóveis que contemplem todos essas características como variáveis individuais, uma alternativa viável é explorar a utilização de informações não estruturadas. Nesse sentido, as modalidades de imagem e texto são duplamente interessantes, pois fotos e descrições compreendem muitos dos detalhes que influenciam o preço, ao mesmo tempo em que são encontradas facilmente em portais agregadores de anúncios imobiliários. Ao contrário das modalidades tradicionais, porém, imagem e texto precisam ser transformados em vetores numéricos antes de poderem ser usados no treinamento de algoritmos de aprendizado de máquina, o que faz com que modelos multimodais sejam necessariamente mais complexos.

Por conta dessa complexidade, a avaliação imobiliária multimodal é um tópico de pesquisa relativamente recente, tendo seus primeiros trabalhos publicados somente em 2016, durante o renascimento das CNNs (LeCun et al., 1989) provocado pela rede AlexNet (Krizhevsky et al., 2012; Tan e Lim, 2018). Ainda assim, como será discutido no Capítulo 3, esses estudos de caso exploram uma única modalidade não estruturada (imagens) e

concentram-se, em termos de abrangência, em uma paisagem urbana bastante específica (casas de subúrbio na América do Norte).

1.1 Problema de pesquisa

A experiência de avaliação imobiliária é essencialmente multimodal. Uma pessoa procurando uma casa para comprar, por exemplo, dispõe de vários formatos de informação que a ajudam a tomar uma decisão embasada. Em momentos distintos, ela pode:

- Restringir sua busca aos imóveis que correspondem a algumas características estruturais, como uma certa área mínima, número de quartos ou vagas de garagem;
- Verificar a disposição dos cômodos, a necessidade de reforma e o aspecto geral de cada propriedade por fotos e vídeos;
- Ler a descrição textual elaborada pela pessoa corretora de imóveis a fim de identificar amenidades, posição solar, existência de vista livre, andar etc;
- Consultar um serviço de mapas para descobrir quais estações de metrô ou pontos de interesse estão próximos ao imóvel.

Cada uma dessas ações está associada a um tipo de informação: características estruturais, imagens, texto e localização geográfica. Cada um desses tipos de informação, por sua vez, é uma modalidade distinta que exerce influência sobre o preço dos imóveis. Assim, uma pessoa avaliadora não faz avaliações dispondo apenas das características estruturais das propriedades. Afinal, o preço de um imóvel é formado por uma miríade de fatores que vai muito além de área, número de quartos e número de banheiros.

Apesar desse senso comum, foi só a partir de 2016 que a multimodalidade começou a ser explorada em tarefas de avaliação imobiliária. Mais especificamente, foram os trabalhos apresentados no Capítulo 3 os primeiros a indicar que a utilização de informações originalmente desestruturadas pode aumentar a acurácia de modelos de predição de preço. Ainda assim, esses estudos de caso concentram-se em uma única modalidade (imagens), um tipo de propriedade (casas) e uma região do mundo (América do Norte).

Desse modo, identificam-se oportunidades de pesquisa tanto na 1) aplicação de estudos de caso sobre paisagens urbanas distintas e heterogêneas quanto na 2) exploração da modalidade de texto, um tipo de informação que ainda não foi explorado na literatura de avaliação imobiliária.

1.2 Objetivos

Nesse contexto, este trabalho tem o objetivo de preencher duas lacunas da pesquisa sobre avaliação imobiliária multimodal. A primeira delas é a falta de diversidade nos estudos de caso existentes, que se concentram majoritariamente em propriedades do tipo casa localizadas no hemisfério norte. A segunda lacuna é a falta de pesquisas explorando a modalidade de texto, um tipo de informação facilmente acessível que, contudo, ainda é pouco explorado como um meio viável para a otimização de MAAs.

Mais do que apresentar um único MAA, esta pesquisa propõe, inicialmente, uma arquitetura multimodal composta por componentes intercambiáveis. O propósito dessa arquitetura é possibilitar a rápida prototipagem de diferentes modelos multimodais, permitindo compreender com agilidade como diferentes modalidades impactam uma determinada tarefa de aprendizado de máquina e estabelecendo um ponto de partida consistente para o desenvolvimento de modelos otimizados para produção.

Além disso, esta pesquisa tem como objetivo aplicar essa arquitetura a um estudo de caso inovador focado na cidade de Porto Alegre, Brasil. Como diferenciais, esse estudo abrangerá mais de um tipo de propriedade, englobando tanto apartamentos quanto casas, e explorará pela primeira vez a utilização simultânea de imagem e texto em tarefas de predição de preços de imóveis.

1.3 Questões de pesquisa

No contexto do estudo de caso apresentado na seção anterior, o trabalho busca responder às seguintes questões de pesquisa:

1. Modelos baseados em imagem propostos em outros estudos de caso são aplicáveis à diversidade da paisagem urbana brasileira, especialmente considerando a inclusão de diferentes tipos de propriedade? (QP1)

2. Como o uso da modalidade de texto impacta na capacidade preditiva de modelos de avaliação imobiliária? (QP2)

3. Quais métodos se destacam como os mais promissores para extrair informação relevante de descrições textuais? (QP3)

4. Quais aspectos visuais e termos de texto influenciam a capacidade preditiva de modelos multimodais de avaliação imobiliária? (QP4)

1.4 Estrutura da dissertação

O restante desta dissertação está organizado da seguinte forma. O Capítulo 2 contém um breve referencial teórico da pesquisa. O Capítulo 3 discute trabalhos que exploram o uso de modalidades não estruturadas na predição de preços de imóveis. O Capítulo 4 apresenta a metodologia empregada na pesquisa, introduzindo a proposta de arquitetura multimodal e sua aplicação em um estudo de caso na cidade de Porto Alegre. Os resultados quantitativos e qualitativos observados são relatados no Capítulo 5 e discutidos no Capítulo 6. Por fim, o Capítulo 7 sumariza as contribuições deste trabalho, aborda suas limitações e propõe possíveis direções para pesquisas futuras.

2. REFERENCIAL TEÓRICO

Pressupõe-se que as teorias e conceitos fundamentais deste trabalho são conhecidos de forma geral. Dessa maneira, este capítulo oferece um breve referencial teórico, concentrando-se nas particularidades da avaliação imobiliária multimodal. Mais especificamente, a Seção 2.1 estabelece as bases dos Modelos de Avaliação Automática, enquanto a Seção 2.2 define o Aprendizado de Máquina Multimodal.

2.1 Modelos de Avaliação Automática

A International Association of Assessing Officers (IAAO) define um MAA ou *Automated Valuation Model* (AVM, em inglês) como um programa de computador que provê o valor estimado de um imóvel em um determinado período de tempo. O diferencial dos MAA em relação à avaliação imobiliária tradicional é a modelagem estatística, utilizada para analisar, a partir de dados coletados previamente, localização, condições de mercado e características da propriedade (International Association of Assessing Officers, 2018). O termo também é utilizado para descrever produtos e serviços que empregam esse tipo de modelo, como Redfin[1] e Zestimate[2].

É importante destacar que todo MAA é um modelo de predição de preços de imóveis, mas nem todo modelo de predição de preços de imóveis é um MAA, pois essa classificação pressupõe sua utilização como ferramenta de apoio à avaliação imobiliária. Inicialmente, defendia-se que a saída de um MAA não fosse tomada como uma estimativa final, dependendo da revisão de uma pessoa avaliadora (Snook, 1998). No entanto, o surgimento de técnicas avançadas de modelagem e a maior disponibilidade de dados (Kok et al., 2017) fez com que o envolvimento de uma pessoa analista de mercado qualificada passasse a ser “altamente recomendável”, mas não obrigatório (International Association of Assessing Officers, 2018).

2.1.1 Métricas de avaliação

As métricas mais comumente utilizadas para mensurar o desempenho de MAAs são o erro percentual absoluto médio (MAPE, em inglês), o erro percentual absoluto mediano (MdAPE, em inglês), o coeficiente de determinação (R^2), o erro absoluto médio (MAE, em inglês), o erro quadrático médio (MSE, em inglês) e a raiz do erro quadrático médio (RMSE, em inglês), representadas pelas Equações 2.1, 2.2, 2.3, 2.4, 2.5 e 2.6, respectiva-

[1] Disponível em https://www.redfin.com/what-is-my-home-worth.
[2] Disponível em https://www.zillow.com/z/zestimate/.

mente. Nas equações, y é o vetor dos valores reais, $\overline{y}$ é a média dos valores reais, $\hat{y}$ é o vetor de valores preditos pelo modelo e n é número de amostras.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2.1}$$

$$MdAPE = median\left(\frac{|y_1 - \hat{y}_1|}{y_1}, \frac{|y_2 - \hat{y}_2|}{y_2}, \ldots, \frac{|y_n - \hat{y}_n|}{y_n}\right) \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})} \tag{2.3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.5}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.6}$$

MAE, MSE e RMSE são métricas dependentes de escala, padrões em tarefas de regressão. Como sua saída tem relação direta com o intervalo de valores presente nos dados, essas métricas são utilizadas principalmente para comparação entre modelos de um mesmo estudo de caso. Por outro lado, MAPE, MdAPE e R^2 são independentes de escala, sendo úteis, portanto, para comparação de resultados entre diferentes estudos de caso. R^2 indica a robustez do modelo — quanto maior seu valor, melhor o resultado. Todas as outras métricas medem erro e operam no sentido contrário, isto é, quanto menor seu valor, melhor o resultado.

No domínio da avaliação imobiliária, cada amostra corresponde a um imóvel avaliado. A variável preditora pode ser qualquer representação numérica do valor do imóvel, como preço de venda, preço de anúncio ou mesmo o preço estimado por uma pessoa avaliadora. Idealmente, utiliza-se o preço de venda, pois tanto preço de anúncio quanto preço estimado apresentam diferenças relevantes (10% e 30%, respectivamente) em relação ao valor efetivamente pago (Loft, 2020; G et al., 2020), o que introduz ruído na avaliação.

2.2 Aprendizado de máquina multimodal

Um problema de pesquisa é considerado multimodal quando envolve várias modalidades, agrupamentos de informações de acordo com o modo como um fenômeno

acontece ou é experimentado. Nesse contexto, Baltrusaitis et al. (2019) afirmam que o objetivo do Aprendizado de Máquina Multimodal é “construir modelos que possam processar e relacionar informações de múltiplas modalidades”. A taxonomia proposta por esses autores introduz cinco desafios centrais desse campo de pesquisa (representação, tradução, alinhamento, fusão e coaprendizagem), dos quais dois (representação e fusão) são especialmente importantes para os MAA.

O desafio da representação relaciona-se às técnicas empregadas para sumarizar a informação de cada modalidade em vetores numéricos representativos e complementares, de forma que possam ser utilizadas em algoritmos de aprendizado de máquina. No domínio da avaliação imobiliária, a modalidade de imagem já foi sumarizada pelo algoritmo SURF (Bay et al., 2006), em um dos primeiros modelos multimodais propostos (H. Ahmed e Moustafa, 2016). Os trabalhos posteriores, no entanto, adotaram unanimamente as redes neurais convolucionais (LeCun et al., 1989).

O desafio da fusão relaciona-se às dificuldades técnicas de unir as informações de várias modalidades a fim de realizar uma predição. A heterogeneidade, própria da multimodalidade, faz com que esses modelos precisem ser construídos levando-se em conta aspectos como complementariedade e redundância das informações, diferentes níveis de capacidade preditiva e ruído, e possíveis dados faltantes (Baltrusaitis et al., 2019).

Por fim, é interessante ressaltar que o estado da arte em modelos multimodais explora cenários onde o texto descreve a imagem (Hossain et al., 2019). No domínio da avaliação imobiliária, porém, as diferentes modalidades cumprem um papel muito mais de complementariedade do que alinhamento.

3. TRABALHOS RELACIONADOS

Este capítulo explora trabalhos relacionados no campo da avaliação imobiliária multimodal, oferecendo um panorama abrangente das pesquisas anteriores que contribuíram para o entendimento atual do tema.

3.1 Modelos baseados em imagem

A utilização da modalidade de imagem em MAAs é um tópico de pesquisa relativamente recente, inaugurado por Bessinger e Jacobs (2016) e H. Ahmed e Moustafa (2016). Esses artigos apresentam uma estrutura comum que foi adotada sem grandes variações em trabalhos posteriores: estudos de caso aplicados a uma região geográfica específica, contendo um ou mais tipos de imagem, um algoritmo para extração de características dessas imagens e um algoritmo para a predição de preço.

3.1.1 Modelos baseados em fotos internas e externas

Bessinger e Jacobs (2016) usaram a CNN VGG-16 (Liu e Deng, 2015) treinada sobre o banco de dados Places2 (Zhou et al., 2017) para extrair características das imagens das fachadas de 83.140 casas do Condado de Fayette, no Kentucky (Estados Unidos). Essas características foram adicionadas a um modelo de Random Forest (Breiman, 2001), reduzindo sua RMSE em 6,14%. Já H. Ahmed e Moustafa (2016) utilizaram imagens internas e externas de 535 casas da Califórnia (Estados Unidos). O algoritmo de extração escolhido foi o SURF (Bay et al., 2006) e a predição de preço foi dada por uma rede neural com o algoritmo de Levenberg-Marquardt, resultando em um aumento do R^2 por um fator de 3 e redução do MSE por uma ordem de magnitude.

Em You et al. (2017), foram avaliadas 4.564 casas de San José, Califórnia, e Rochester, Nova Iorque, ambas nos Estados Unidos. As características visuais foram extraídas através da CNN GoogLeNet (Szegedy et al., 2015) e o algoritmo de predição de preço escolhido foi uma LSTM bidirecional, um tipo especial de BRNN. Esse modelo obteve um MAPE de 16,11% para San José e 22,69% para Rochester.

O estudo de caso de Liu et al. (2018) se concentrou no estado da Califórnia (Estados Unidos). Nele, os autores coletaram cerca de 900.000 imagens de 30.141 casas e propuseram um modelo chamado Multi-instance Deep Ranking and Regression (MiDRR), uma rede neural que executa tanto a extração de características quanto a predição de preço. O principal diferencial do MiDRR é a ordenação prévia dos imóveis baseada em

regras derivadas do senso comum (uma casa com piscina tende a ser mais cara do que uma casa sem piscina, por exemplo). Os autores utilizaram essa classificação como um regulador do MiDRR e relataram um MAPE de menos de 1%.

Poursaeed et al. (2018) treinaram uma DenseNet (Huang et al., 2017) sobre imagens obtidas do banco de dados Places (Zhou et al., 2014), Houzz[1] e Google Image Search[2] para avaliar o nível de requinte de fotos internas e externas de 1.000 casas nos Estados Unidos. O nível de requinte e os atributos tradicionais das casas foram utilizados como dados de entrada de um algoritmo Kernel SVM, resultando em um MdAPE de 5,8%.

Zhao et al. (2019) utilizaram o algoritmo MobileNet (Howard et al., 2017), treinado com imagens do banco de dados AVA (Murray et al., 2012), para extrair características visuais de 248 casas de Camberra, Austrália. Utilizando XGBoost (Chen e Guestrin, 2016) como algoritmo de aprendizado, seu modelo obteve 8,70% de MAPE.

Wu e Zhang (2021) destacam-se por terem avaliado diversas CNNs, pré-treinadas com o banco de dados ImageNet (Deng et al., 2009), como algoritmos de extração. Sua arquitetura final incluiu uma CNN profunda (ResNet-50 [He et al., 2016]) e uma CNN rasa como algoritmo de extração, bem como uma rede neural como algoritmo de aprendizado, obtendo 18,01% de MAPE na avaliação de 1.000 imóveis de Los Angeles, Califórnia (EUA).

Em seu estudo de caso, Srirutchataboon et al. (2021) utilizaram como algoritmo de extração a rede VGG-19 (Liu e Deng, 2015), também pré-treinada com o banco ImageNet. A predição de preço, dada por Random Forest e calibrada por regressão linear, obteve 17,83% de MAPE para 4.472 casas na Tailândia.

3.1.2 Modelos baseados em imagens de satélite

Dois trabalhos destacam-se pela introdução de imagens de satélite. Em Bency et al. (2017), foram avaliados 51.253 imóveis à venda e 55.700 imóveis para locação das cidades inglesas de Londres, Birmingham e Liverpool. As características visuais foram extraídas com a CNN Inception V3 (Szegedy et al., 2016) pré-treinada com o banco de dados ImageNet e o algoritmo de predição escolhido foi Random Forest, obtendo um R^2 de 92,51% para os imóveis à venda e 90,77% para os imóveis para locação. Já Muhr et al. (2017) criaram uma CNN chamada SatNet-8 a partir da VGG (Liu e Deng, 2015) para avaliar 2.739 imóveis da região do Tirol, na Áustria. Seu modelo de preço hedônico (Liao e Wang, 2012) resultou em um R^2 de 66%.

Através de uma rede chamada CNN for United Mining (UMCNN), Yao et al. (2018) extraíram características visuais de 4.331 imóveis da cidade de Shenzhen, na China. A predição de preço, dada pelo algoritmo Random Forest, resultou em um MAPE de 26,2%.

[1] Disponível em https://www.houzz.com/.
[2] Disponível em https://www.google.com/imghp.

Como algoritmo de extração, Chen et al. (2020) combinaram Rank Siamese Network (Burges et al., 2005) e a CNN ResNet-50 pré-treinada com o banco ImageNet. As características visuais extraídas foram utilizadas para estimar o preço de 54.000 imóveis da Região Metropolitana de Toronto, no Canadá, com 55,2% das predições registrando MAPE de até 10%. O algoritmo de aprendizado utilizado foi Gradient Boosting (Friedman, 2001).

3.1.3 Modelos baseados em imagens ao nível da rua

Zhang e Dong (2018) foram os primeiros autores a explorar imagens ao nível da rua, um tipo especial de imagens panorâmicas de 360° capturadas e disponibilizadas por serviços como Google Maps[3] e Tencent Maps[4]. O objetivo dessa pesquisa foi verificar se os preços dos apartamentos de uma região de Pequim, na China, eram influenciados pela quantidade de vegetação no entorno de seu edifício. Para isso, o "nível de verde" de cada rua foi extraído com o algoritmo SegNet (Badrinarayanan et al., 2017) — pré-treinado com o banco de dados CamVid (Brostow et al., 2008) — e adicionado a um modelo de preço hedônico, obtendo um R^2 de 75,36%.

Bin et al. (2020) utilizaram o mesmo tipo de imagem para predizer o preço de avaliação de 15.815 casas da Filadélfia, na Pensilvânia (Estados Unidos). Seu algoritmo de extração, uma versão modificada da rede VGG-16, foi pré-treinado sobre parte das imagens do banco de dados Places365 (Zhou et al., 2017). Para selecionar as imagens de treino, os autores utilizaram a rede PSPNet (Zhao et al., 2017) para remover imagens cuja cena contivesse mais de 20% de objetos irrelevantes. Utilizando XGBoost como algoritmo de aprendizado, seu modelo obteve 82,30% de R^2.

O trabalho de Lee e Park (2020), por sua vez, propôs um MAA composto por uma CNN como algoritmo de extração e uma rede neural como algoritmo de aprendizado. Essa arquitetura aumentou a acurácia das predições de preço de 3.007 casas e edifícios da cidade de Guri, na província de Gyeonggi, Coreia do Sul.

3.1.4 Outros modelos baseados em imagem

Bin et al. (2019) concluíram que a utilização de aspectos visuais extraídos de mapas, através de uma CNN chamada Attention-Based Multi-Modal Fusion (AMMF), aumentou o R^2 de seu modelo para 85,10%, enquanto o MAPE foi reduzido a 17,50%. Para isso, foram avaliadas as predições de preço feitas para 3.873 casas de Los Angeles, Califórnia (Estados Unidos). O algoritmo de aprendizado utilizado foi XGBoost.

[3] Disponível em https://www.google.com/maps.
[4] Disponível em https://map.qq.com/.

Já Law et al. (2019) destacam-se por terem combinado dois tipos de imagem e comparado várias arquiteturas experimentais. Em seu estudo de caso, os autores utilizaram imagens de satélite e imagens ao nível da rua para predizer o preço de imóveis localizados em 40.470 ruas de Londres, na Inglaterra. Seu melhor modelo, cujo R^2 foi de 85%, foi composto de uma CNN como algoritmo de extração e um perceptron multicamadas como algoritmo de aprendizado.

Kostic e Jevremovic (2020) também optaram pela estratégia de combinação, utilizando fotos internas e externas, imagens de satélite e imagens ao nível da rua de 19.942 imóveis dos estados de Massachusetts e Nova Iorque (Estados Unidos). Seu trabalho destaca-se por combinar as representações numéricas geradas por quatro algoritmos de extração distintos: entropia (Shannon, 1948), centro de gravidade, segmentação e ResNet152-hybrid1365 (Zhou et al., 2017) pré-treinada com os banco de dados ImageNet e Places365. Utilizando LightGBM (Ke et al., 2017) como algoritmo de aprendizado, seu melhor modelo obteve R^2 de 90% e MAPE de 11%.

3.2 Modelos baseados em texto

Se a utilização da modalidade de imagem em tarefas de avaliação imobiliária já reúne um número razoável de trabalhos, a modalidade de texto, por outro lado, segue praticamente inexplorada, apenas mencionada como uma possibilidade por Law et al. (2019). Existem, no entanto, pesquisas que exploram o uso de características textuais em problemas similares, como a predição de preços de acomodações.

Em Peng et al. (2020), por exemplo, o conteúdo textual de avaliações de usuários foi utilizado para predizer o preço de locação de acomodações disponibilizadas pela plataforma Airbnb[5]. Para isso, os autores computaram uma pontuação de sentimento para cada avaliação e agregaram as pontuações de cada acomodação. Como conclusão do trabalho, identificou-se que utilizar o resultado dessa agregação em conjunto com outras informações da acomodação, tais como características estruturais e localização geográfica, contribuiu positivamente em todas as métricas de avaliação.

3.3 Discussão

Alguns trabalhos (H. Ahmed e Moustafa, 2016; Wu e Zhang, 2021; Nouriani e Lemke, 2022) usam o termo “características textuais” (nos originais, *textual features* ou *textual information*) para referir-se, na verdade, a características estruturais *numéricas* dos imóveis, como área e número de quartos. Dessa forma, até onde sabe-se pela re-

[5] Disponível em https://www.airbnb.com.

visão da literatura, a modalidade de texto (informações textuais desestruturadas, como descrições de anúncios) ainda não foi explorada na avaliação imobiliária.

Em relação à modalidade de imagem, os trabalhos apresentados na Subseção 3.1 relatam aumento de desempenho em uma ou mais métricas de avaliação, como R^2, MAPE ou MdAPE. Essas melhorias indicam que a utilização de aspectos visuais pode, de fato, aumentar a capacidade de modelos de predição de preços de imóveis. No entanto, o mapa da Figura 3.1 evidencia que esses estudos de caso estão concentrados, em termos de tamanho de amostra, em apenas três países: Inglaterra, Estados Unidos e Canadá.

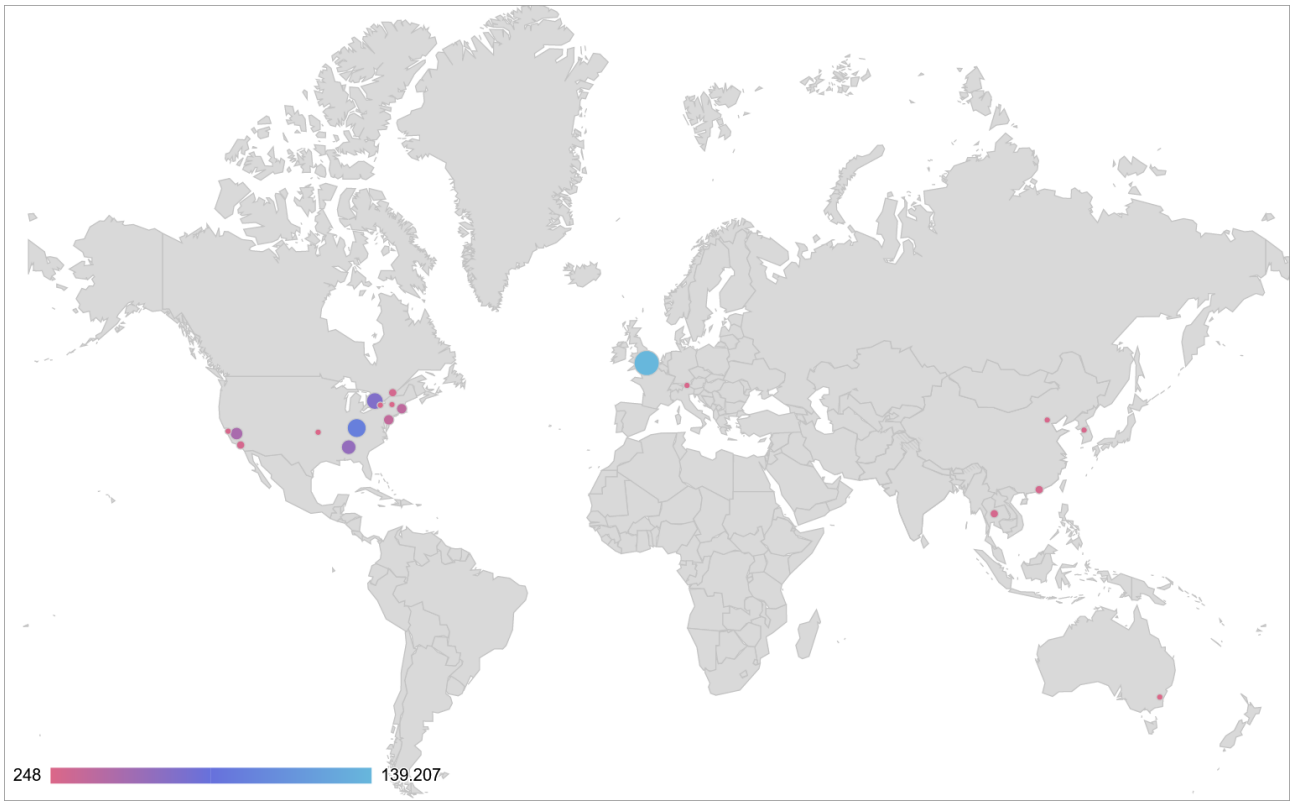


Figura 3.1: Distribuição dos estudos de caso. O tamanho de cada círculo corresponde ao número de imóveis analisados

Consequentemente, existem vastas regiões geográficas sem estudos de caso que mensurem o potencial dessa modalidade para a avaliação imobiliária. A solução para essas lacunas não pode ser a pressuposição de que o melhor modelo proposto para imóveis norte-americanos, por exemplo, também será o melhor para qualquer região não estudada, pois “cada país terá uma cultura e uma experiência diferentes, que vão determinar os métodos adotados para qualquer avaliação” (Pagourtzi et al., 2003).

Adicionalmente, os tipos de dado utilizados em um estudo de caso podem não estar disponíveis ou mesmo não se aplicar a outra região. Como exemplo, Lee e Park (2020) mencionam que alguns bancos de imagem sobre os quais muitas CNNs são pré-treinadas são enviesados por características arquitetônicas comuns na América do Norte

e na Europa. Logo, o que essas redes entendem por templo, por exemplo, baseia-se mais na concepção ocidental de igreja do que nos traços de um pagode asiático.

Nesse contexto, a paisagem urbana do Brasil possui ao menos duas características que representam oportunidades de pesquisa. Primeiramente, as casas brasileiras são cercadas por muros e/ou grades; casas de subúrbio dos Estados Unidos, sobre as quais baseia-se grande parte dos estudos de caso, não são. Além disso, 14,2% da população brasileira vive em apartamentos (Instituto Brasileiro de Geografia e Estatística, 2020), tipo de imóvel abordado apenas por Zhang e Dong (2018).

Por fim, a pesquisa de avaliação imobiliária multimodal pode avançar também em direção à interpretabilidade, já que uma parcela significativa dos trabalhos relacionados não faz nenhuma análise qualitativa que busque determinar, por exemplo, quais aspectos visuais mais contribuíram para os resultados relatados.

4. METODOLOGIA DE PESQUISA

Este trabalho adotou um protocolo experimental sistemático. Nesse contexto, este capítulo expõe em detalhes toda a metodologia empregada na pesquisa, apresentando a proposta de arquitetura multimodal e demonstrando sua aplicação em um estudo de caso focado na cidade de Porto Alegre, Brasil.

4.1 Visão geral da arquitetura multimodal

Todos os experimentos conduzidos nesta pesquisa foram implementados em uma arquitetura multimodal desenvolvida exclusivamente para este trabalho. Seu objetivo é permitir a rápida prototipagem de modelos multimodais e avaliar as interações entre as modalidades, bem como a contribuição de cada uma para a tarefa em questão. Os componentes de todas as camadas dessa arquitetura são intercambiáveis, possibilitando a rápida ativação e desativação de modalidades e a exploração de novos algoritmos de aprendizado supervisionado, por exemplo. Essa flexibilidade permite, em última instância, a criação de um número virtualmente infinito de MAAs.

A Figura 4.1 ilustra as quatro camadas conceituais da arquitetura: transformação de dados, submodelos de modalidade, camada de fusão e preditor. Na primeira camada, os conjuntos de dados são submetidos a transformações, expansões e enriquecimentos. Na segunda camada, cada modalidade de informação é processada de forma independente por um submodelo de aprendizado de máquina. A camada de fusão, por sua vez, visa integrar as saídas desses submodelos de forma com que elas se complementem e resultem em uma predição mais robusta. Por fim, os dados resultantes da fusão são utilizados como entrada do preditor, camada responsável por estimar um valor, seja ele contínuo ou uma classe, conforme a natureza do problema.

As próximas subseções apresentam uma descrição detalhada de cada uma dessas camadas, bem como a implementação técnica da arquitetura multimodal.

4.1.1 Camada de transformação de dados

A camada de transformação de dados é responsável por realizar todas as etapas de pré-processamento necessárias antes do início do treinamento do modelo. Nessa fase, o conjunto de dados pode passar por alterações tanto no número de variáveis quanto no número de amostras. A lista abaixo apresenta alguns exemplos de suas aplicações, muitos dos quais foram empregados no estudo de caso deste trabalho:

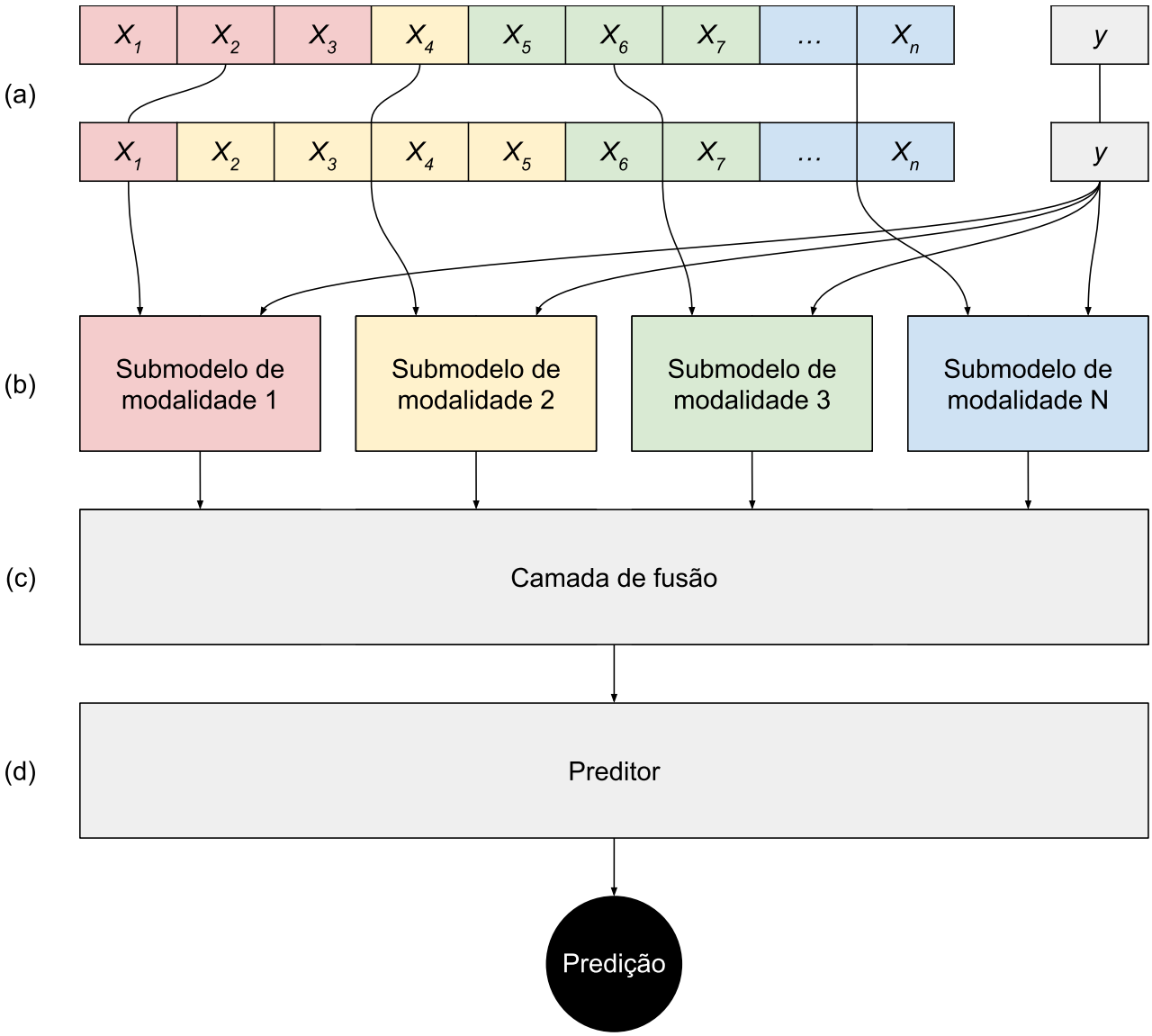


Figura 4.1: Visão geral da arquitetura multimodal mostrando a) a camada de transformação de dados, b) os submodelos de modalidade, c) a camada de fusão e d) o preditor

- **Normalização:** ajuste dos valores para que estejam dentro de uma escala específica, geralmente entre 0 e 1;
- **Padronização:** ajuste dos valores para que tenham média 0 e desvio padrão 1;
- **Tratamento de dados ausentes:** preenchimento ou remoção de valores ausentes, garantindo que o conjunto de dados esteja completo;
- **Codificação de variáveis categóricas:** conversão de variáveis categóricas em uma forma numérica adequada para o modelo, como *one-hot encoding*;
- **Transformação logarítmica ou exponencial:** aplicação de logaritmo ou exponenciação aos dados para lidar com distribuições assimétricas ou para reduzir a escala de valores extremos;

- **Redução de dimensionalidade:** utilização de técnicas como PCA para reduzir a dimensionalidade e preservar as características mais importantes;
- **Engenharia de recursos:** criação de novas variáveis ou combinações de variáveis que possam representar melhor os padrões presentes nos dados;
- **Remoção de *outliers*:** identificação e remoção de valores atípicos que podem distorcer o treinamento do modelo;
- **Discretização:** conversão de variáveis contínuas em categorias discretas, o que pode facilitar a interpretação e o treinamento em alguns casos;
- **Agrupamento de dados:** agrupamento de dados em categorias ou intervalos, especialmente útil para variáveis contínuas.

4.1.2 Camada de submodelos de modalidade

A partir da segunda camada da arquitetura, as variáveis deixam de ser tratadas individualmente e são agrupadas em conjuntos maiores — as modalidades. A composição específica dessas modalidades varia conforme a tarefa a ser realizada e as relações entre as informações que se deseja explorar, aproveitando a facilidade da arquitetura para ativar e desativar esses conjuntos. Cada submodelo é um algoritmo de aprendizado de máquina responsável por processar de forma independente os dados de uma modalidade. Em geral, o submodelo guarda uma relação próxima com a camada de transformação de dados, pois pode depender de um pré-processamento específico das variáveis.

Nessa camada, dados não estruturados como imagem e texto (que não podem ser utilizados diretamente como entrada de modelos de aprendizado de máquina) são processados para um formato adequado, geralmente um vetor numérico que representa a informação. Modalidades de texto podem ter submodelos baseados em estratégias de *ensemble* como *bagging* (Breiman, 1996) e *boosting* (Freund et al., 1996). Imagens, por sua vez, são tradicionalmente processadas por meio de CNNs (LeCun et al., 1989).

Independentemente do submodelo escolhido, é importante ressaltar que a arquitetura não impõe nenhuma restrição ao formato de saída dos submodelos, permitindo que eles produzam desde novos vetores numéricos até mesmo predições intermediárias, saídas que são então encaminhadas à camada de fusão.

4.1.3 Camada de fusão

O papel da camada de fusão é unificar os dados provenientes dos submodelos de modalidade de forma a gerar uma representação consolidada de cada amostra. Na prática, essa função se traduz na geração de um vetor numérico único, posteriormente encaminhado à camada de predição. Essa tarefa, no entanto, não é trivial, pois as modalidades envolvidas podem variar significativamente entre si. A escolha da estratégia de fusão, seja por concatenação, ponderação ou outras abordagens, deve levar em conta a complementaridade das informações e evitar redundâncias.

Cabe destacar que uma abordagem específica pode produzir resultados distintos dependendo da combinação de submodalidades utilizada, influenciando diretamente na capacidade preditiva do modelo. Não surpreende, portanto, que a fusão seja um dos desafios centrais do aprendizado de máquina multimodal (Baltrusaitis et al., 2019).

4.1.4 Camada de predição

Em termos conceituais, essa é a camada mais direta da arquitetura multimodal. Sua função básica é aplicar a representação consolidada gerada pela camada de fusão como parâmetro de entrada em diferentes preditores, permitindo assim a comparação do desempenho individual de cada um. A saída dessa camada é a predição de um valor contínuo em problemas de regressão ou de uma categoria em problemas de classificação.

4.1.5 Implementação técnica

Embora as camadas da arquitetura proposta representem etapas comuns no desenvolvimento de modelos de aprendizado de máquina, sua implementação técnica se destaca ao garantir que seus componentes sejam verdadeiramente intercambiáveis. Na prática, isso significa que é possível criar diversos modelos distintos com o mínimo de alterações de código, proporcionando uma avaliação abrangente e metódica do impacto da multimodalidade nas mais variadas tarefas.

Para assegurar essa flexibilidade e garantir a compatibilidade com um número maior de projetos, optou-se por implementar a arquitetura como uma leve extensão à classe *Pipeline* da biblioteca scikit-learn (Buitinck et al., 2013). Essa classe fornece uma maneira conveniente de encadear diversas etapas, desde o pré-processamento de dados até o treinamento do modelo. Além disso, facilita a repetição desses passos em diferentes arranjos de componentes, alinhando-se inteiramente ao propósito da arquitetura

proposta. Por fim, essa implementação dá acesso, por padrão, a todas as ferramentas da scikit-learn, como diferentes métodos de seleção de variáveis, validação de modelo e otimização de hiperparâmetros (por exemplo, *grid search*).

A extensão proposta, denominada MP, provê funções adicionais através de pequenas classes utilitárias. Essas adições incluem o cálculo de MdAPE, a capacidade de expandir variáveis do tipo lista e a possibilidade de utilizar dados de diferentes modalidades em um mesmo submodelo. Além disso, a arquitetura aprimora o suporte a redes neurais de duas maneiras distintas. Em primeiro lugar, permitindo a utilização de redes PyTorch (Paszke et al., 2019) por meio da biblioteca skorch (Tietz et al., 2017). Em segundo lugar, incluindo um adaptador de redes neurais que resolve, por padrão, diversos problemas comuns na integração desse tipo de algoritmo, como correções de tipos de dados e a capacidade de determinar, em tempo de execução, o número de dimensões da camada de entrada da rede. O código-fonte e exemplos de aplicação da arquitetura estão disponíveis em https://github.com/lfbittencourt/mp.

4.2 Coleta de dados

O conjunto de dados utilizado neste trabalho foi coletado em maio de 2022, aplicando-se a técnica de *scraping* a um *site* agregador de anúncios imobiliários. Ao todo, foram coletadas 212.076 amostras, correspondentes a todos os anúncios de venda para fins residenciais então disponíveis para a cidade de Porto Alegre, Brasil. Nessa primeira etapa, foram selecionadas todas as variáveis listadas na Tabela 4.1.

Tabela 4.1: Variáveis originais do conjunto de dados

Variável	Tipo
Tipo da propriedade	Texto
Área	Real
Número de quartos	Inteiro
Número de banheiros	Inteiro
Latitude	Real
Longitude	Real
Imagens	Lista de textos
Descrição	Texto
Preço	Real

Após duas semanas, realizou-se a coleta das imagens indicadas na variável homônima, que continha a lista das URLs de imagem associadas a cada amostra. Ao todo, foram baixadas 1.702.960 imagens, totalizando 91 GB de informação adicional. Vale ressaltar que nem todas as imagens puderam ser obtidas, o que pode ter ocorrido devido à remo-

ção ou edição dos anúncios entre as duas coletas, tornando as imagens indisponíveis, ou a falhas simples no processo de *download*.

As próximas seções detalham como a aplicação desse conjunto de dados à arquitetura proposta culminou no estudo de caso deste trabalho.

4.3 Pré-processamento dos dados

Concluídas ambas as coletas, deu-se início ao processo de limpeza do conjunto de dados. Inicialmente, os tipos de propriedade presentes na variável *Tipo da propriedade* foram padronizados, mantendo-se apenas as amostras relacionadas a apartamentos ou casas. Em seguida, os métodos detalhados no Apêndice A foram empregados para identificar e remover anúncios duplicados, preservando-se a amostra com a descrição mais longa. Por fim, foram removidas amostras contendo quaisquer valores faltantes, incluindo imóveis sem nenhuma imagem baixada e/ou com descrições em branco[1].

Após uma análise gráfica por meio de histogramas e diagramas de caixa (*boxplots*), identificou-se a presença de *outliers* nas variáveis *Área*, *Número de quartos*, *Número de banheiros* e *Preço*. Esses valores aberrantes caracterizam-se por estar significativamente distantes das demais observações, o que pode prejudicar o desempenho do MAA. Por essa razão, optou-se por remover as amostras associadas a esses *outliers*. Para evitar a aplicação de critérios arbitrários de exclusão, adotou-se a técnica de IQR (Whaley III, 2005), na qual consideram-se aberrantes valores que estejam abaixo de $Q1 - 1,5$ IQR ou acima de $Q3 + 1,5$ IQR, sendo $IQR = Q3 - Q1$.

Em seguida, foram removidas amostras que, porventura, estivessem localizadas fora dos limites geográficos do município de Porto Alegre. Essa delimitação foi feita de maneira aproximada traçando-se um retângulo do ponto definido pela latitude e pela longitude mínimas (-30,26945 e -51,30344, respectivamente) até o ponto estabelecido pela latitude e longitude máximas da cidade (-29,932474 e -51,018852, respectivamente).

Iniciando a fase de transformação dos dados, as variáveis *Área*, *Número de quartos* e *Número de banheiros* foram normalizadas para um intervalo entre 0 e 1. A variável *Tipo da propriedade*, por sua vez, foi convertida para valores inteiros, sendo casas e apartamentos representados pelos números 0 e 1, respectivamente.

[1] Uma alternativa à remoção de amostras com dados faltantes é a substituição desses dados por valores plausíveis, como a média das demais amostras. Essa imputação preserva um número maior de amostras, mas introduz valores estimados que não foram observados de fato. Por esse motivo, optou-se pela remoção.

4.3.1 Transformação de informações geográficas

Embora as variáveis *Latitude* e *Longitude* sejam do tipo real e não necessitem transformações para serem empregadas em modelos de aprendizado de máquina, sua eficácia como atributo de localização só ocorre quando utilizadas em conjunto, indicando uma coordenada geográfica precisa. Por isso, foi preciso convertê-las para um formato que mapeasse como o preço varia no espaço coberto pelo estudo de caso. Para isso, o retângulo imaginário mencionado anteriormente foi subdividido em uma matriz de 28×38 quadrantes de aproximadamente 1 km^2, identificados como A1, A2, B1, B2 etc, categorias que foram então expandidas por *one-hot encoding*[2].

A Figura 4.2 mostra a distribuição dos quadrantes sobre o mapa de Porto Alegre. Observa-se que alguns desses quadrantes ultrapassam os limites geográficos da cidade. Apesar dessa limitação, manteve-se essa abordagem pela facilidade de aplicação a outros estudos de caso que envolvam coordenadas geográficas, proporcionando consistência metodológica e comparação entre diferentes contextos. Independentemente da técnica utilizada, acredita-se que a subdivisão em quadrantes menores pode capturar melhor a variação de preços do que recortes geográficos mais amplos, como bairros.

4.3.2 Classificação de imagens

Assim como as descrições textuais, as imagens são formas desestruturadas de informação que precisam ser convertidas em vetores numéricos representativos, usualmente por meio de CNNs (LeCun et al., 1989). Dada a grande variação na quantidade[3] e diversidade das imagens coletadas para este estudo de caso, foi necessário encontrar uma forma homogênea de sintetizar os aspectos visuais de cada propriedade. Para isso, optou-se por combinar diversas imagens em um *grid*, uma espécie de colagem das imagens mais representativas de cada imóvel. Segundo Rosebrock (2019), essa estratégia evita apresentar à CNN imagens diferentes para preços idênticos, otimizando assim o aprendizado de filtros discriminativos.

Para que tal aprendizado seja possível, no entanto, a montagem desses *grids* não pode ser feita de forma aleatória, já que a CNN depende de padrões visuais para compreender a relação entre imagens e preços. Em termos práticos, isso significa que a colagem de fotos deve seguir uma ordem pré-determinada, de maneira que imagens semelhantes sempre ocupem a mesma posição. Nesse contexto, decidiu-se pela implementação de um *grid* 2×2 medindo 256px × 256px e contendo imagens de sala, cozinha, quarto e

[2] Nessa técnica, cada categoria é mapeada para um vetor binário distinto em que o único elemento marcado como 1 é o que corresponde à própria categoria.

[3] A quantidade de imagens por propriedade variava de 0 a 200.

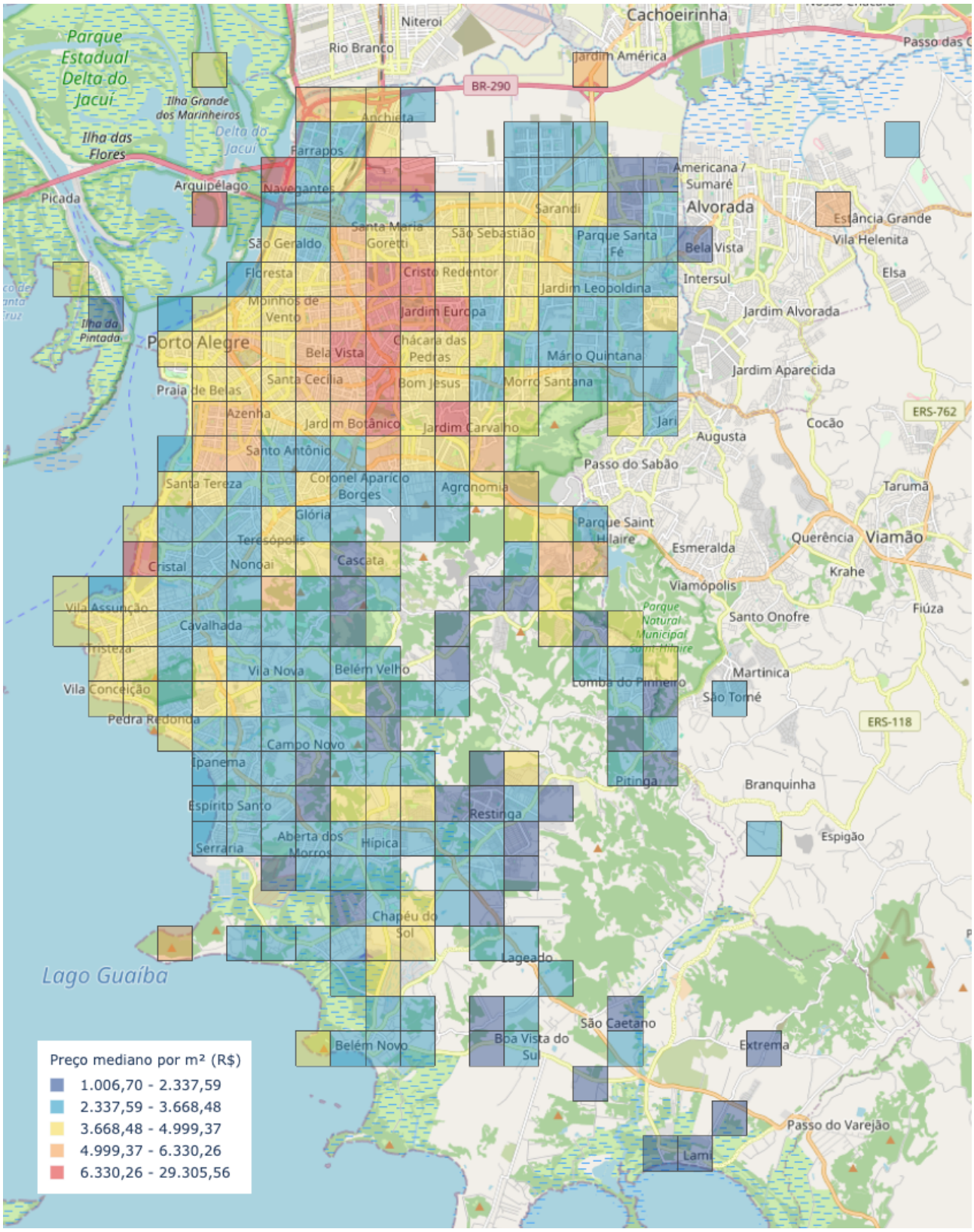


Figura 4.2: Distribuição dos quadrantes geográficos com ao menos uma amostra sobre o mapa de Porto Alegre. As cores indicam a variação do preço mediano do metro quadrado

banheiro. A escolha desse formato baseou-se na sua compatibilidade com uma grande variedade de CNNs que requerem imagens quadradas. Além disso, essa abordagem segue o princípio de reproduzir da melhor forma possível a experiência de avaliação de um ser humano, contemplando uma foto para cada cômodo relevante do imóvel.

Dado que as imagens coletadas não eram rotuladas, foi necessário realizar uma etapa preliminar de classificação. Para viabilizar essa tarefa com custo computacional relativamente baixo, optou-se por empregar transferência de conhecimento a partir de uma ResNet-50 (He et al., 2016) pré-treinada sobre o banco de dados Places365 (Zhou et al., 2017), uma combinação popular utilizada em outros trabalhos de avaliação imobiliária multimodal, como Bin et al. (2020) e Kostic e Jevremovic (2020).

Inicialmente, a classe do banco Places365 atribuída a cada imagem foi aquela com a probabilidade mais alta predita pela ResNet-50. No entanto, uma análise exploratória revelou que imagens de um mesmo cômodo podiam ser associadas a mais de uma classe, evidenciando os vieses no reconhecimento de padrões relatados por Lee e Park (2020). Para atenuar os efeitos desses vieses, identificou-se inicialmente a classe mais atribuída a cada um dos cômodos que comporiam o *grid*, a saber “living_room”, “kitchen”, “bedroom” e “bathroom”. Entre todas as imagens associadas a cada uma dessas classes, observou-se que outras classes eram mais frequentemente apontadas como a segunda mais provável. Para cada uma dessas classes secundárias, conduziu-se uma análise exploratória adicional para confirmar se as imagens classificadas como tal referiam-se, de fato, ao cômodo examinado. Em caso afirmativo, essas classes foram consideradas *aliases*, isto é, classes equivalentes. Sua lista completa pode ser vista na Tabela 4.2.

Tabela 4.2: Classes principais e equivalentes (*aliases*) do banco de dados Places365 para cada um dos cômodos incluídos no *grid* de imagens

Cômodo	Classe principal	*Aliases*
Sala de estar	*living_room*	*dining_room* *television_room* *waiting_room*
Cozinha	*kitchen*	*galley* *restaurant_kitchen*
Quarto	*bedroom*	*bedchamber* *berth* *childs_room* *hotel_room* *youth_hostel*
Banheiro	*bathroom*	*shower*

Ao fim desse processo, constatou-se que 36,15% das imagens foram classificadas como algum dos cômodos do *grid*. Os imóveis que não apresentaram nenhuma imagem

classificada foram removidos do conjunto de dados. Nas demais amostras, os *grids* apresentaram uma média de 2,78 imagens, destacando-se que 82,83% continham imagens de banheiro, 68,17% de cozinha, 67,84% de sala e 59% de quarto[4]. Embora não tenha sido possível mensurar o desempenho do classificador, uma breve análise exploratória indicou que a maioria das imagens foi categorizada corretamente. Nesse sentido, a Figura 4.3 apresenta exemplos de imagens correta e incorretamente classificadas.

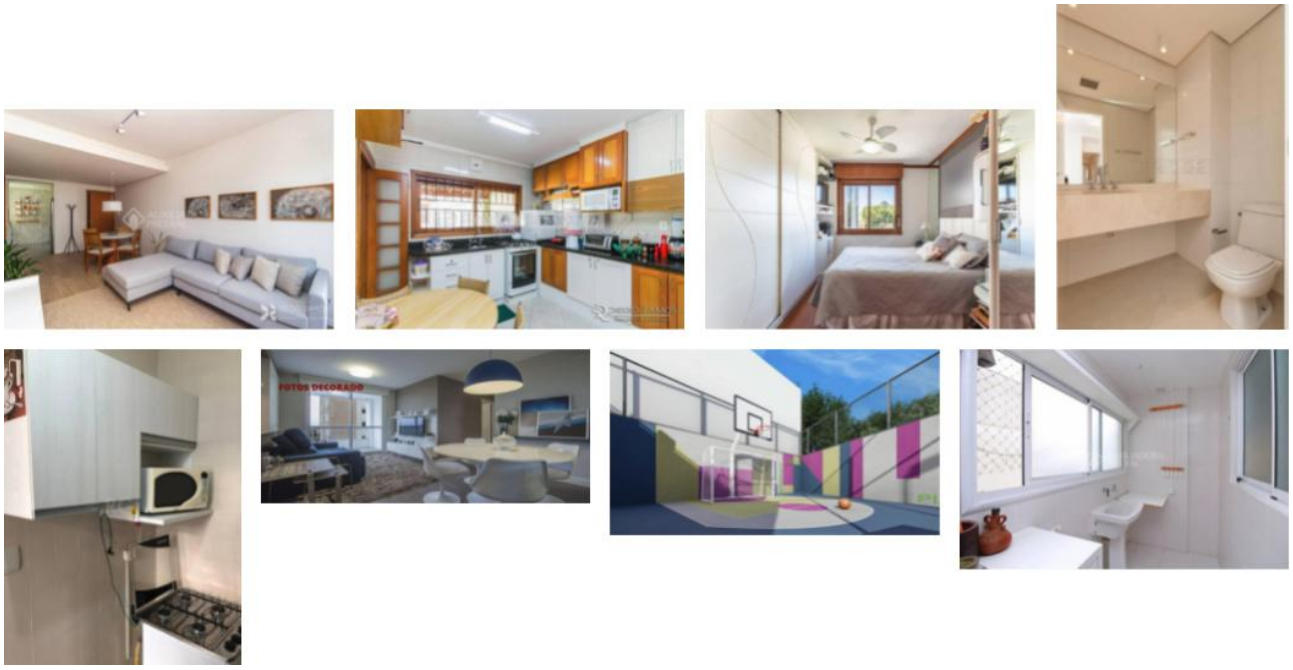


Figura 4.3: Exemplos da classificação de imagens. A linha superior apresenta imagens corretamente identificadas como sala de estar, cozinha, quarto e banheiro. A linha inferior destaca imagens incorretamente classificadas para os mesmos cômodos

4.3.3 Extração de características visuais

Na etapa final no pré-processamento de imagens, recorreu-se novamente à transferência de conhecimento para converter cada um dos *grids* gerados na etapa anterior em um vetor numérico representativo. Para essa tarefa, essencial para identificar padrões e características relevantes nas imagens, optou-se por reutilizar a CNN ResNet-50, dessa vez pré-treinada com o banco de dados ImageNet (Deng et al., 2009), abordagem empregada em trabalhos relacionados como Stivanello e Brignoli (2023). Os *grids* já estavam no formato esperado pela rede[5], que foi então utilizada em modo de inferência, excluindo a

[4] Nos casos em que uma amostra apresentou mais de uma imagem classificada como um determinado cômodo, manteve-se a de maior probabilidade predita para a classe ou *alias*. De forma semelhante, nos casos em que um imóvel possuía múltiplos cômodos do mesmo tipo (dois quartos, por exemplo), assumiu-se que o cômodo cuja imagem foi selecionada não apresentava diferenças visuais significativas em relação aos outros cujas fotos não foram escolhidas.

[5] Nos casos em que o *grid* não possuía imagem para um determinado cômodo, o espaço correspondente foi preenchido por um quadrante preto.

camada de classificação. O vetor de características foi extraído da saída da última camada convolucional, composta por 2.048 elementos, e agregado a cada amostra.

Por fim, a Figura 4.4 ilustra todas as fases do pré-processamento de imagens, apresentando a configuração dos *grids*, um exemplo de *grid* completo após a etapa de classificação e as transformações aplicadas pela biblioteca PyTorch para otimizar a extração de características visuais.

Figura 4.4: Etapas do pré-processamento de imagens: a) configuração dos *grids*, b) *grid* preenchido e c) transformações aplicadas pela biblioteca PyTorch

4.4 Definição de modalidades

Após todas as operações de limpeza e transformação, o conjunto de dados final consistiu em 61.233 amostras, sendo 49.945 relacionadas a apartamentos (81,57%) e 11.288 a casas (18,43%). A Tabela 4.3 apresenta algumas características das amostras de acordo com o tipo de propriedade, enquanto a Tabela 4.4 apresenta as principais estatísticas das variáveis quantitativas do conjunto de dados.

Tabela 4.3: Valores médios das variáveis quantitativas por tipo de propriedade

	Área	Nº quartos	Nº banheiros	Preço (R$)
Apartamentos	79,77	2,18	1,61	490.741,10
Casas	165,37	2,94	2,50	693.342,48

Conforme apresentado na Tabela 4.5, as variáveis do conjunto de dados ou suas derivações foram agrupadas em quatro modalidades: características estruturais, localização geográfica, imagem e texto. As variáveis de localização geográfica foram agrupadas em uma modalidade à parte para permitir avaliar o impacto das modalidades não estruturadas na presença e na ausência de informações geográficas precisas (como os quadrantes de localização apresentados na Subseção 4.3.1).

Tabela 4.4: Estatísticas das variáveis quantitativas do conjunto de dados

Variável	Tipo da propriedade	Área	Nº quartos	Nº banheiros	Preço
Tipo	Inteiro	Inteiro	Inteiro	Inteiro	Real
Média	0,82	95,55	2,32	1,78	528.089,66
Desvio padrão	0,39	56,60	0,77	0,93	340.929,20
Mínimo	0,00	10,00	1,00	1,00	20.000,00
25º percentil	1,00	57,00	2,00	1,00	260.000,00
50º percentil	1,00	75,00	2,00	2,00	430.000,00
75º percentil	1,00	118,00	3,00	2,00	700.000,00
Máximo	1,00	294,00	4,00	6,00	1.705.000,00

Tabela 4.5: Variáveis do conjunto de dados por modalidade

Modalidade	Variáveis
Características estruturais	Tipo da propriedade Área Número de quartos Número de banheiros
Localização geográfica	Latitude Longitude
Imagem	Imagens
Texto	Descrição

4.4.1 Submodelos de modalidade

Os submodelos de características estruturais, localização geográfica e imagem limitaram-se a transferir à camada de fusão os valores transformados durante o pré-processamento de dados. Em contrapartida, o submodelo de texto seguiu integralmente a implementação de Bittencourt et al. (2022), ilustrada na Figura 4.5. Cada descrição de anúncio passou por uma etapa relativamente simples de pré-processamento, consistindo na aplicação de caixa baixa a todas as palavras e na remoção de diacríticos e das *stopwords* (termos irrelevantes como “ao”, “do”, “e” etc) fornecidas pela biblioteca NLTK[6].

No entanto, mesmo pré-processadas, as descrições ainda eram informações textuais desestruturadas, inviáveis para aplicação direta em algoritmos de aprendizado de máquina. Nesse contexto, foi necessário convertê-las em vetores numéricos que codificassem as informações contidas nas descrições. Para isso, adotou-se a técnica de TF-IDF (Spärck Jones, 1972), que reduz a importância de um termo em um documento caso ele apareça frequentemente nos demais documentos.

[6]Versão 3.7. Documentação disponível em https://www.nltk.org/.

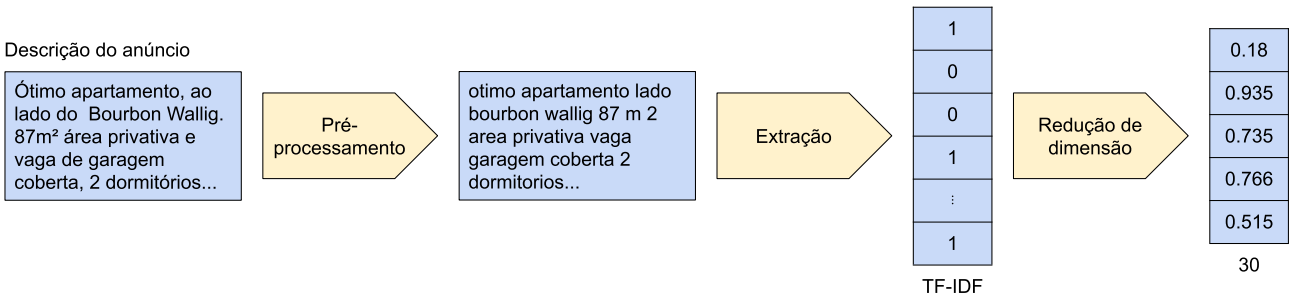


Figura 4.5: Etapas do submodelo de texto

Foram descartados termos presentes em menos de 0,1% das amostras e adotou-se uma estratégia binária de contagem de termos. Nessa abordagem, todas as contagens maiores do que 0 são definidas simplesmente como 1, indicando a presença do termo na descrição textual da amostra. Por fim, utilizou-se o algoritmo Truncated SVD (Halko et al., 2011) para reduzir a dimensão dos vetores numéricos a 30 elementos, um passo essencial para evitar que o grande número de variáveis derivadas da descrição diminuísse a importância das demais variáveis do conjunto de dados.

4.5 Definição de preditores

Para este estudo de caso, foram selecionados três algoritmos de aprendizado de máquina amplamente reconhecidos na literatura sobre MAAs. Os dois primeiros, Random Forest (Breiman, 2001) e Extremely Randomized Trees (Geurts et al., 2006), adotam a abordagem de *ensemble learning*, construindo múltiplos modelos baseados em árvore e combinando suas predições para obter resultados mais robustos. A principal diferença entre eles está na forma como determinam as divisões nos nós durante a construção das árvores. Ao passo em que o Random Forest faz uma seleção aleatória de características para determinar a melhor divisão, o ERT adiciona ainda pontos de divisão aleatórios para cada nó, justificando assim seu nome, “árvores *extremamente* aleatorizadas”.

O terceiro algoritmo abordado são as redes neurais. Conhecidas por sua capacidade de aprender padrões complexos e representações hierárquicas dos dados, essa classe de algoritmos é apropriada a contextos nos quais a relação entre os preditores e a variável de resposta é intrincada (Dayhoff, 1990). Pelli Neto e Zárate (2003) destacam que boa parte das relações entre as características físicas, os valores de utilidade de um imóvel e seu preço de mercado exibem comportamentos não lineares, fazendo com que as redes neurais sejam uma escolha bastante comum na avaliação imobiliária, sobretudo quando são empregadas modalidades complexas, como imagens.

Random Forest e ERT foram implementados utilizando as classes *RandomForestRegressor* e *ExtraTreesRegressor* da biblioteca scikit-learn (Pedregosa et al., 2011)[7], uma opção que aproveitou a tecnologia-base da arquitetura proposta. Ambos os algoritmos mantiveram os hiperparâmetros com seus valores padrão, incluindo *ensembles* de 100 árvores e configurações que não impõem nenhum limite ao seu crescimento.

A rede neural, por sua vez, foi desenvolvida utilizando a biblioteca PyTorch (Paszke et al., 2019)[8], o otimizador Adam (Kingma e Ba, 2017) e uma taxa de aprendizado de 0,001. Sua arquitetura incluiu duas camadas ocultas com 128 e 64 neurônios, respectivamente. Já seu treinamento foi configurado para ocorrer ao longo de 100 épocas, com parada antecipada caso o erro de validação não reduzisse ao menos 0,1% em um período de 10 épocas. Em caso de interrupção precoce, a rede foi programada para restaurar os pesos da época com menor erro registrado. A função de perda adotada foi o MSE, avaliado sobre os 10% das amostras de treino reservados à validação interna da rede.

4.6 Definição de experimentos

As quatro modalidades deste trabalho podem ser combinadas de 15 maneiras distintas. Para o estudo de caso, foram escolhidas 11 dessas combinações, selecionadas estrategicamente para investigar questões relevantes da pesquisa:

- **Quatro modelos unimodais.** Esses modelos individuais permitiram medir até que ponto os dados de uma única modalidade compreendem a formação dos preços dos imóveis. É importante ressaltar que o modelo unimodal de características estruturais também foi uma das linhas de base com as quais foi mensurada a contribuição relativa das modalidades de imagem e texto;
- **Segunda linha de base com localização geográfica.** Esse segundo modelo de linha de base incorporou, além das características estruturais, a modalidade de localização geográfica. Seu objetivo foi verificar se o impacto da agregação de modalidades complexas é influenciado pela presença de informações geográficas precisas, conforme discutido na Seção 4.4;
- **Dois modelos com adição de imagem.** Esses modelos foram configurados para avaliar como a inclusão da modalidade de imagem impacta as duas linhas de base;
- **Dois modelos com adição de texto**: Esses modelos buscaram compreender os efeitos da adição da modalidade de texto às duas linhas de base;

[7] Versão 1.2. Documentação disponível em https://scikit-learn.org/1.2/.
[8] Versão 2.0.1. Documentação disponível em https://pytorch.org/docs/2.0/.

- **Dois modelos com adição de imagem e texto**: Esses modelos exploraram como a inclusão simultânea das modalidades complexas influencia as duas linhas de base.

Combinados aos três preditores descritos na Seção 4.5, esses modelos geraram um conjunto de 33 experimentos, estabelecendo a base experimental do estudo de caso.

4.7 Avaliação de desempenho

A avaliação dos experimentos empregou o método de validação cruzada conhecido como *k-fold*. Nesta técnica, o conjunto de dados foi subdividido em 10 subconjuntos mutuamente exclusivos ($k = 10$), todos com o mesmo tamanho. Cada um desses subconjuntos foi então designado como conjunto de validação para um modelo treinado com os 9 ($k-1$) subconjuntos restantes. Esse processo foi repetido 10 vezes, resultando em 10 avaliações distintas. Por fim, as métricas de cada experimento foram calculadas pela média dos resultados obtidos em cada uma dessas iterações. Essa abordagem foi escolhida por proporcionar uma avaliação de desempenho robusta e abrangente, considerando diversas divisões dos dados e promovendo maior confiabilidade nos resultados finais.

Ao todo, foram analisadas seis métricas de avaliação, todas extraídas de trabalhos relacionados: MAPE, MdAPE, R^2, MAE, MSE e RMSE. As três primeiras foram incluídas devido à sua independência de escala, possibilitando comparações com o desempenho de modelos propostos em pesquisas anteriores. Em consonância com outros estudos da literatura, a qualidade dos modelos foi avaliada primordialmente através da RMSE. Essa métrica penaliza erros maiores de forma mais severa, uma característica particularmente relevante na avaliação imobiliária. Além disso, a RMSE é expressa na mesma unidade da variável de interesse (o preço de anúncio dos imóveis), oferecendo uma compreensão intuitiva que facilita a interpretação dos resultados.

4.7.1 Modelo para análise qualitativa

Visto que um dos objetivos deste trabalho é proporcionar uma análise qualitativa dos resultados, especialmente por meio da avaliação individual de predições, tornou-se necessário projetar um protocolo de avaliação mais adequado a esse fim. Primeiramente, concluiu-se que a validação cruzada proporcionada pelo método *k-fold* era dispensável, uma vez que a análise qualitativa baseia-se em um *ranking* de predições *individuais*. Em seu lugar, optou-se por uma abordagem mais simples, dividindo-se o conjunto de dados entre 55.109 amostras de treino (90% do total) e 6.124 amostras de teste (10% do total).

Além disso, percebeu-se que a manutenção da RMSE como métrica principal resultaria na predominância de imóveis mais caros no *ranking* das piores predições, e vice-

versa: como a RMSE é expressa em valores absolutos, propriedades com preços mais elevados tendem naturalmente a apresentar erros mais significativos. Por evitar esse viés indesejado, optou-se pela adoção do MAPE, uma métrica relativa que proporciona uma avaliação mais equitativa, independentemente do preço do imóvel.

Por fim, esse novo protocolo de avaliação foi empregado para analisar o desempenho de um modelo concebido especificamente para a tarefa de análise qualitativa, utilizando todas as modalidades descritas na Seção 4.4 e o preditor de rede neural.

4.8 Definição da camada de fusão

A estratégia adotada para a camada de fusão deste trabalho consiste na concatenação simples dos dados de saída dos submodelos de modalidade, seguindo a abordagem empregada em Bittencourt et al. (2022). Essa escolha foi justificada por dois motivos. Primeiramente, esse método mostrou-se adaptável a todos os experimentos planejados para este estudo de caso (incluindo o modelo de análise qualitativa), independentemente da composição específica de modalidades. Em segundo lugar, a concatenação foi considerada capaz de preservar as características individuais e a riqueza informativa de cada modalidade, capturando eficientemente suas nuances complementares e entregando aos preditores uma representação robusta dos dados.

4.9 Infraestrutura computacional

Os experimentos deste estudo de caso foram executados no serviço SageMaker Studio, da empresa AWS. A instância utilizada possuía 32 vCPUs e 128 GB de RAM.

5. RESULTADOS

Este capítulo apresenta os resultados observados na execução de 33 experimentos, resultantes da combinação dos 11 conjuntos de modalidade detalhados na Seção 4.6 e de três diferentes preditores (Random Forest, ERT e rede neural). A tabela completa com todas as métricas de todos os experimentos pode ser vista no Apêndice B.

5.1 Análise quantitativa

A Tabela 5.1 apresenta os valores de RMSE para todos os 33 experimentos executados. Para facilitar a análise dos resultados, essa tabela está dividida horizontalmente em três seções: 1) modelos unimodais, 2) modelos que adicionam modalidades não estruturadas a uma linha de base de características estruturais e 3) modelos que adicionam modalidades não estruturadas a uma linha de base de características estruturais e localização geográfica. Para os dois preditores baseados em árvores de decisão (Random Forest e ERT), a menor RMSE foi dada pelo modelo trimodal de características estruturais, localização geográfica e texto; para a rede neural, pelo modelo tetramodal.

Tabela 5.1: RMSE observado em todos os experimentos. Melhor valor de cada regressor em negrito. No cabeçalho, as abreviações CE, LG, IM e TX correspondem a características estruturais, localização geográfica, imagem e texto, respectivamente

CE	LG	IM	TX	Random Forest	ERT	Rede Neural
✓				212.753,24	215.946,88	215.470,82
	✓			290.646,00	290.659,03	315.535,80
		✓		307.157,67	306.207,79	332.544,10
			✓	204.304,39	200.284,51	252.024,39
✓		✓		195.396,56	194.053,60	206.892,30
✓			✓	161.182,94	157.369,00	181.886,21
✓		✓	✓	162.116,27	159.852,97	184.063,32
✓	✓			169.613,51	175.879,74	188.313,06
✓	✓	✓		182.301,50	169.558,67	182.972,80
✓	✓		✓	**157.219,13**	**143.814,54**	159.384,05
✓	✓	✓	✓	159.146,46	148.312,54	**156.679,53**

A Tabela 5.2, por sua vez, expande a análise dos modelos unimodais apresentando cinco das seis métricas monitoradas[1]. Nota-se que os melhores resultados de cada métrica, destacados em negrito, estão inteiramente relacionados à modalidade de texto.

[1]Os valores de MSE foram omitidos para melhor legibilidade e por estarem “representados” pela RMSE. Como raiz quadrada é uma função monotônica, as relações de grandeza se mantêm.

Além disso, os melhores valores de cinco das seis métricas analisadas foram obtidos pelo preditor ERT[2], o que guarda grande semelhança com os resultados observados em Bittencourt et al. (2022). Por outro lado, os piores valores para todas as métricas estão relacionados à modalidade de imagem, indicando que ela, isoladamente, não é capaz de explicar a formação dos preços dos imóveis.

Tabela 5.2: Resultados dos modelos unimodais. Melhor valor de cada métrica em negrito

Modalidade	Regressor	MAPE	MdAPE	R^2	MAE	RMSE
Características estruturais	Random Forest	0,3261	0,2466	0,6105	151.427,03	212.753,24
	ERT	0,3274	0,2471	0,5987	152.548,07	215.946,88
	Rede Neural	0,2975	0,2440	0,6005	151.229,41	215.470,82
Localização geográfica	Random Forest	0,5501	0,3620	0,2732	218.118,44	290.646,00
	ERT	0,5501	0,3621	0,2731	218.116,15	290.659,03
	Rede Neural	0,4345	0,3636	0,1433	220.081,41	315.535,80
Imagem	Random Forest	0,6177	0,4126	0,1881	233.216,62	307.157,67
	ERT	0,6001	0,3984	0,1932	228.502,58	306.207,79
	Rede Neural	0,4890	0,4012	0,0486	236.291,40	332.544,10
Texto	Random Forest	0,3140	0,2289	0,6408	142.740,77	204.304,39
	ERT	0,3073	**0,2232**	**0,6548**	**138.860,60**	**200.284,51**
	Rede Neural	**0,2853**	0,2484	0,4539	165.733,92	252.024,39

A Tabela 5.3, por sua vez, mostra a redução de RMSE obtida ao adicionar-se as modalidades de imagem e texto sobre um modelo de linha de base treinado apenas com características estruturais. Os resultados apresentam padrões consistentes. Ao analisar-se a tabela horizontalmente, percebe-se que o preditor ERT foi o que mais se beneficiou da adição de modalidades não estruturadas, enquanto a análise vertical mostra que a adição isolada da modalidade de texto foi a que resultou nas maiores reduções de RMSE.

Tabela 5.3: Redução de RMSE ao adicionar-se modalidades não-estruturadas a um modelo de linha de base de características estruturais

	Random Forest	ERT	Rede Neural
+ Imagem	8,16%	10,14%	3,98%
+ Texto	24,24%	27,13%	15,59%
+ Imagem e Texto	23,80%	25,98%	14,58%

É importante observar que, embora menores, as reduções de RMSE resultantes da adição isolada da modalidade de imagem são significativas, entre aproximadamente 4 e 10%, o que está de acordo com os trabalhos apresentados no Capítulo 3. Por fim, a tabela mostra que, nesse estudo de caso, adicionar as duas modalidades não estruturadas simultaneamente não equivaleu à soma das reduções resultantes das adições isoladas.

[2]Considerando-se novamente que o preditor com menor RMSE terá, por consequência, o menor MSE.

Pelo contrário, a redução de RMSE desses modelos trimodais foi levemente menor do que a resultante da adição isolada da modalidade de texto.

Por fim, a Tabela 5.4 mostra a redução de RMSE obtida ao adicionar-se as modalidades não estruturadas a um modelo treinado com características estruturais e localização geográfica. Em geral, os valores observados são significativos, ainda que mais tímidos em comparação à linha de base anterior. Nota-se que a adição da modalidade de texto combinada ao preditor ERT mais uma vez produziu a maior redução, mas já não há um padrão claro entre as linhas ou colunas: enquanto a adição isolada da modalidade de imagem piorou a RMSE para Random Forest em mais de 7%, o modelo tetramodal apresentou os melhores resultados para a rede neural.

Tabela 5.4: Redução de RMSE ao adicionar-se modalidades não-estruturadas a um modelo de linha de base de características estruturais e localização geográfica

	Random Forest	ERT	Rede Neural
+ Imagem	-7,48%	3,59%	2,84%
+ Texto	7,31%	18,23%	15,36%
+ Imagem e Texto	6,17%	15,67%	16,80%

Por fim, a estratégia que apresentou os melhores resultados entre todos os experimentos avaliados foi adicionar a modalidade de texto a um modelo de linha de base de características estruturais e localização geográfica. Combinado ao preditor ERT, esse modelo trimodal apresentou os melhores MAPE (0,1915), R^2 (0,822), MAE (94.118,81), MSE (20.682.622.299,74) e RMSE (143.814,54), representando incrementos de 13,65%, 12,03%, 12,08%, 33,14% e 18,23%, respectivamente. O melhor MdAPE foi obtido pela própria linha de base combinada ao preditor ERT (0,1342).

5.2 Análise qualitativa

Essa seção detalha as análises qualitativas realizadas a fim de melhorar a interpretabilidade dos resultados observados nesse estudo de caso. Nesse contexto, a Tabela 5.5 apresenta as características das amostras de teste associadas aos 100 menores e 100 maiores valores de MAPE, respectivamente. Para chegar-se a essas amostras, o modelo descrito na Seção 4.7.1 foi treinado e testado 100 vezes a fim de mitigar a instabilidade inerente à natureza aleatória das redes neurais[3]. Em outras palavras, os resultados apresentados abaixo são as saídas médias dessas execuções.

Entre os menores valores, ou seja, menores erros, nota-se um padrão consistente relacionado às modalidades não estruturadas: as amostras com menor MAPE têm 9,36%

[3]Todas as execuções compartilharam os mesmos conjuntos de treino e teste.

Tabela 5.5: Características das amostras de teste com os 100 melhores e os 100 piores valores de MAPE. As colunas representam valores médios das amostras observadas, com a exceção de “Aptos”, que indica o percentual de apartamentos

	Aptos	Área	Nº quartos	Nº banheiros	Imagens	Descrição	Preço (R$)
100 melhores	83%	92,98	2,31	1,79	2,92	573,83 carac.	511.396,27
100 piores	70%	121,54	2,45	1,86	2,67	552,56 carac.	338.689,87

mais imagens e descrições ligeiramente mais longas, outro indicador de que imagem e texto têm impacto positivo na capacidade preditiva dos modelos.

Entre os maiores valores, por outro lado, destaca-se a proporção de apartamentos menor do que a média do conjunto de dados (81,57%). Esse é um resultado esperado, pois o conjunto de dados é desbalanceado quanto à proporção de cada tipo de propriedade. Logo, é natural que o modelo tenha um desempenho inferior ao predizer o preço de casas. Também percebe-se que os valores médios de área, número de quartos e número de banheiros são maiores entre os maiores erros, consoante com a maior representação de casas nesse grupo (ver Tabela 4.3). Com os valores de preço, no entanto, verifica-se o contrário: embora o preço médio de casas para todo o conjunto de dados seja maior do que o de apartamentos, nota-se que as amostras com maior MAPE têm preço médio significativamente mais baixo do que às de menor MAPE. Em análise posterior, constatou-se que o conjunto de treino utilizado continha apenas 181 amostras de casas com preços de até R$ 350.000, o que pode ser uma explicação plausível para o padrão observado.

A Figura 5.1 mostra que quase metade das amostras com menor MAPE possuíam um *grid* completo de imagens, com fotografias de sala, cozinha, quarto e banheiro. A Figura 5.2, por sua vez, compara os termos mais relevantes identificados nas descrições textuais das amostras de teste associadas aos 100 melhores e 100 piores MAPE. Nota-se, em ambos os grupos, termos relacionados entre si, como “porto”/“alegre”, “menino”/“deus” e “bourbon”/“country”, referências à cidade de Porto Alegre, ao bairro Jardim Carvalho e ao *shopping* Bourbon Country, respectivamente. Percebe-se ainda que a nuvem de palavras relacionada aos maiores valores de MAPE contém três vezes mais termos do que a associada aos menores valores, incluindo palavras que não têm nenhuma relação com o imóvel em si, como “agendar”, “agora”, “conhecer”, “estuda” e “informacoes”.

5.2.1 Análise individual

Por fim, as amostras da última execução foram analisadas individualmente visando identificar características das imagens e textos que pudessem explicar a inclusão desses imóveis entre os valores mais baixos ou mais altos de MAPE. Dentre as amostras com os melhores valores, todos próximos de zero, observaram-se descrições textuais que

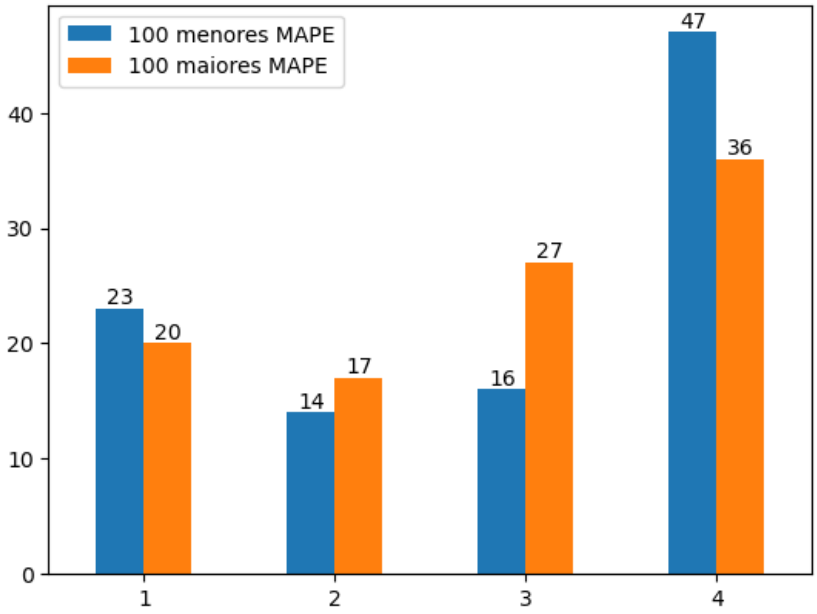


Figura 5.1: Distribuição das amostras de teste associadas aos 100 menores e 100 maiores MAPE em relação ao número de imagens

(a) 100 menores MAPE

(b) 100 maiores MAPE

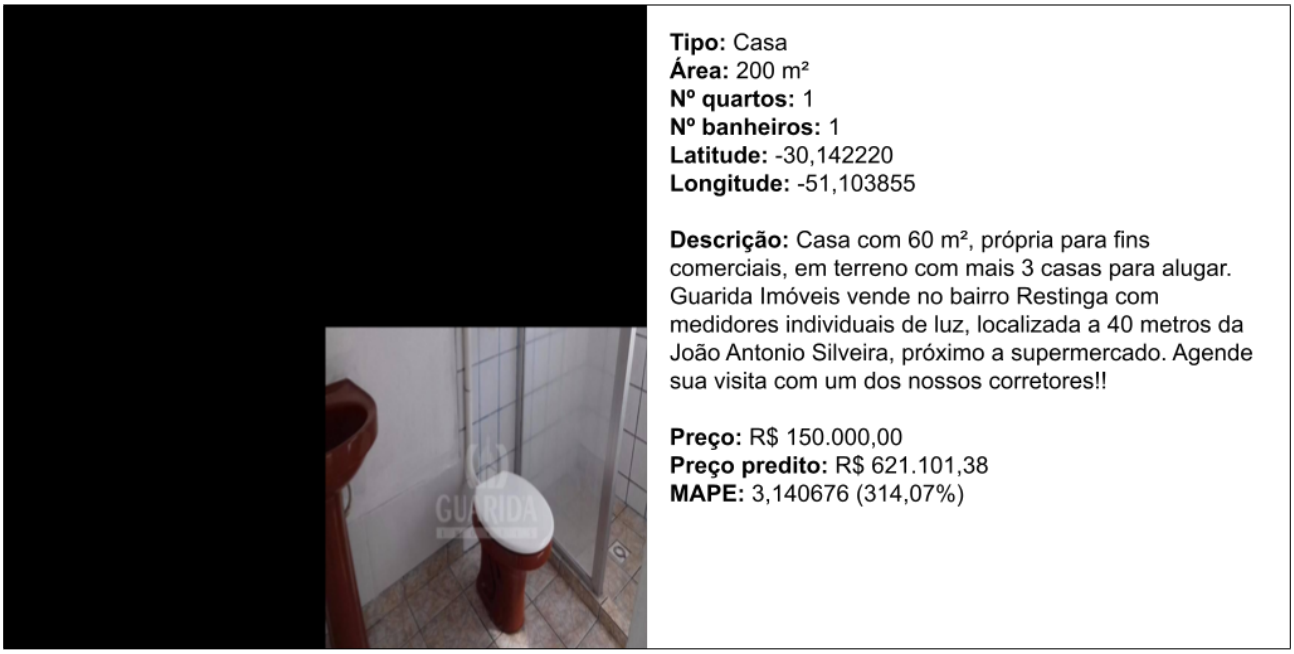
Figura 5.2: Nuvens de palavras com os termos mais relevantes das descrições textuais das amostras associadas aos 100 menores e 100 maiores MAPE

enfatizam aspectos que podem desempenhar um papel importante na decisão de compra de um imóvel. Destacam-se estabelecimentos próximos e amenidades como churrasqueira, incidência de sol, mobília, móveis sob medida, portaria, reforma recente, salão de festas, vagas de garagem e vista. Em relação às imagens, evidenciou-se o já mencionado aumento no número de *grids* completos e a qualidade das fotografias, capturadas em sua maioria a partir de ângulos que permitem uma visão abrangente de cada peça, como exemplificado na Figura 5.3.

Do outro lado do espectro de erros, observaram-se valores de MAPE entre 0,6743 e 3,1407. Evidenciaram-se diferenças na quantidade e qualidade das imagens, como ilustrado na Figura 5.4, e indícios nas descrições textuais que podem explicar, ao menos em parte, a baixa qualidade das predições. Por exemplo:

Figura 5.3: Exemplo de amostra bem avaliada pelo modelo, com erro notadamente baixo

- Um anúncio descrevia duas propriedades distintas. Ao analisar-se variáveis objetivas como área e número de quartos, concluiu-se que a segunda parte da descrição, relacionada a uma sala comercial, foi agregada erroneamente, possivelmente por falha humana durante a criação do anúncio;
- O anúncio apresentado na Figura 5.4 descrevia uma casa com área de 60 m², enquanto a variável objetiva *Área* foi erroneamente especificada como 200 m², indicando outro possível erro humano na elaboração do anúncio;
- Outro anúncio descrevia uma propriedade em processo de inventário, o que explica a discrepância entre o preço predito pelo modelo e o preço anunciado, já que a compra de um imóvel nessas condições traz riscos ao comprador. Em análise posterior, constatou-se que o termo “inventario” foi descartado pelo algoritmo de TF-IDF;
- Diversos imóveis foram anunciados com preço abaixo de mercado. Não foi possível identificar as prováveis causas dessa disparidade com base nos dados disponíveis. O apartamento ilustrado na Figura 5.5, por exemplo, é bem localizado e possui comodidades desejadas pelos compradores, como banheira de hidromassagem e churrasqueira. Mesmo com essas características, a propriedade foi anunciada com um preço notavelmente baixo, fazendo com que esse imóvel registrasse o sétimo maior MAPE dentre as amostras analisadas.



Tipo: Casa
Área: 200 m²
Nº quartos: 1
Nº banheiros: 1
Latitude: -30,142220
Longitude: -51,103855

Descrição: Casa com 60 m², própria para fins comerciais, em terreno com mais 3 casas para alugar. Guarida Imóveis vende no bairro Restinga com medidores individuais de luz, localizada a 40 metros da João Antonio Silveira, próximo a supermercado. Agende sua visita com um dos nossos corretores!!

Preço: R$ 150.000,00
Preço predito: R$ 621.101,38
MAPE: 3,140676 (314,07%)

Figura 5.4: Exemplo de amostra mal avaliada pelo modelo, com erro notadamente alto

Tipo: Apartamento
Área: 64 m²
Nº quartos: 2
Nº banheiros: 1
Latitude: -30,061941
Longitude: -51,218128

Descrição: Apartamento no Menino Deus com Garagem, dois dormitórios transformado em tres dormitórios, banheiro com hidro, amplo living, cozinha americana, terraço com churrasqueira.

Preço: R$ 185.250,00
Preço predito: R$ 496.447,38
MAPE: 1,679878 (167,99%)

Figura 5.5: Exemplo de amostra mal avaliada sem causa provável identificada

6. DISCUSSÃO

A revisão sistemática dos resultados foi fundamental para uma compreensão mais profunda do comportamento dos modelos. A análise qualitativa, por exemplo, deixou claro que a escolha dos submodelos de modalidade e da camada de fusão guarda uma grande relação com o desempenho dos MAA.

A análise qualitativa dos maiores erros, por sua vez, mostrou que algumas predições poderiam ser realizadas de forma mais precisa por seres humanos, capazes de identificar incoerências nas informações dos anúncios. Desprovido dessa capacidade, o MAA prediz o preço do imóvel supondo que todos os dados foram preenchidos corretamente. Em outros casos, no entanto, presume-se que uma pessoa avaliadora, tendo acesso às mesmas informações fornecidas ao modelo, chegaria a uma avaliação (e erro) semelhante, já que nenhum dado do anúncio oferece justificativa plausível para o imóvel estar sendo vendido a um valor significativamente distinto dos observados para propriedades semelhantes em localizações equivalentes.

Como muitos dos erros do modelo se justificam, sua verdadeira capacidade de compreender a formação dos preços dos imóveis é maior do que as métricas sugerem. Em alguns casos, ele fez predições acertadas com base nos dados disponíveis, isto é, previu um preço que corresponde, de fato, *ao que foi anunciado*.

As próximas seções detalham outros aspectos específicos dos resultados.

6.1 Impacto das modalidades não estruturadas

Um dos principais objetivos desse trabalho é analisar o impacto de modalidades de informação não estruturadas no desempenho de MAAs. Mais especificamente, buscou-se 1) verificar se esse estudo de caso está de acordo com a literatura em relação à contribuição positiva da modalidade de imagem e 2) analisar em profundidade os efeitos da adição da modalidade de texto, um tipo de informação amplamente disponível mas ainda pouco explorado em trabalhos semelhantes. Os resultados apresentados no Capítulo 5 mostram que ambas as modalidades cumprem o propósito de otimizar MAAs, evidenciado pelas reduções de RMSE, mas a intensidade dessa otimização varia de acordo com o conjunto de todas as modalidades incluídas e o preditor adotado.

Enquanto a modalidade de imagem, isoladamente, não foi capaz de explicar a formação dos preços dos imóveis, os modelos unimodais baseados em texto apresentaram os melhores resultados para todas as métricas analisadas, superando o desempenho de modalidades que contêm informações básicas como tipo do imóvel, área e localização. Esse resultado surpreendente demonstra que a formação do preço de um imóvel

engloba uma gama muito maior de fatores que, dentre os modelos unimodais, é melhor compreendida pela modalidade de texto.

Além disso, destaca-se que o impacto positivo da utilização da descrição textual não se limita aos modelos unimodais. Como demonstrado na Tabela 6.1, a adição dessa modalidade a diferentes modelos de linha de base, em conjunto com todos os preditores analisados, melhorou os valores de praticamente todas as métricas avaliadas.

Tabela 6.1: Melhoria nas métricas ao comparar-se modelos que incluem a modalidade de texto com seus equivalentes que não a possuem. As abreviações CE, LG e IM correspondem a características estruturais, localização geográfica e imagem, respectivamente

Linha de base	Regressor	MAPE	MdAPE	R^2	MAE	RMSE
CE	Random Forest	30,39%	32,08%	27,18%	27,84%	24,24%
	ERT	33,07%	35,12%	31,43%	30,93%	27,13%
	Rede Neural	25,36%	22,92%	19,11%	19,02%	15,59%
CE + LG	Random Forest	1,74%	-16,85%	4,64%	0,83%	7,31%
	ERT	13,65%	-2,89%	12,03%	12,08%	18,23%
	Rede Neural	20,85%	15,84%	12,46%	17,22%	15,36%
CE + IM	Random Forest	21,90%	20,28%	15,25%	19,17%	17,03%
	ERT	21,46%	19,97%	15,41%	19,21%	17,62%
	Rede Neural	24,28%	19,76%	12,14%	14,96%	11,03%
CE + LG + IM	Random Forest	15,05%	12,64%	9,53%	13,67%	12,70%
	ERT	12,38%	6,67%	7,72%	11,85%	12,53%
	Rede Neural	18,40%	16,10%	10,78%	15,89%	14,37%

Por fim, a análise qualitativa apresentada na Seção 5.2 demonstrou que anúncios que são fáceis de avaliar por seres humanos também o são por MAAs. Por consequência, elaborá-los com fotos de qualidade e descrições textuais relevantes melhora não apenas a capacidade de um modelo de predizer seu preço, mas também sua própria utilidade como ferramenta de apoio à tomada de decisão de compra de um imóvel.

6.2 Impacto da modalidade de localização geográfica

Um subconjunto dos experimentos foi pensado para avaliar se o impacto das modalidades não estruturadas é alterado pela presença de informações geográficas precisas, como latitude e longitude. Os resultados quantitativos apresentados na Seção 5.1 mostram que imagem e texto continuam a contribuir positivamente na redução de erro quando o modelo de linha de base inclui a modalidade de localização. Contudo, essa contribuição é mais tímida e passa a depender mais das modalidades e preditor utilizados.

Essa análise evidencia a importância da modalidade de localização geográfica e mostra que ela desempenha um papel significativo nos resultados[1]. É válido ressaltar, no entanto, que as modalidades não estruturadas continuam a contribuir de maneira positiva, reforçando sua relevância no contexto dos MAA. Por fim, comprova-se a hipótese levantada em Bittencourt et al. (2022) de que a modalidade de texto pode ser uma substituta razoável na ausência de informações geográficas precisas.

6.3 Custo computacional

É relevante destacar que o modelo de características estruturais, localização geográfica e texto, em conjunto com o preditor ERT, teve desempenho superior ao de modelos tetramodais que incluíam imagem, pois esse resultado tem duas implicações antagônicas em relação ao custo computacional. Por um lado, o submodelo de texto introduzido em Bittencourt et al. (2022) e adotado nesse estudo de caso faz uso de algoritmos simples, há longo estabelecidos (Spärck Jones, 1972) e com baixo custo computacional.

Por outro lado, observou-se que preditores baseados em árvores de decisão (ERT e Random Forest) conseguem aproximar os preços de maneira consistente. Essa tendência decorre do fato de a biblioteca utilizada para implementar esses algoritmos (scikit-learn) não impor, por padrão, nenhum limite ao tamanho e à complexidade das árvores, o que otimiza os resultados a um custo computacional que pode ser proibitivo. Foi devido a esse custo, por exemplo, que os experimentos desse estudo de caso precisaram ser executados em uma instância mais robusta (e cara) do serviço do SageMaker Studio, conforme descrito na Seção 4.9.

6.4 Limitações da arquitetura

Ainda que a arquitetura tenha se mostrado um bom ponto de partida para explorar o potencial da multimodalidade na resolução de problemas de aprendizado de máquina, é importante ressaltar uma limitação significativa: seu fluxo de dados unidirecional, que vai da camada de transformação de dados até o preditor uma única vez. Na prática, essa característica impede a implementação de redes neurais completas, pois não é possível fazer uma retropropagação dos erros de ponta a ponta. Em outras palavras, as redes neurais da arquitetura estão limitadas a atuar isoladamente como submodelos de modalidade, camadas de fusão ou preditores. Ainda que a arquitetura suporte submodelos que

[1] A relevância da modalidade de localização geográfica fica evidente nos melhores valores (destacados em negrito) da Tabela 5.1, localizados exclusivamente na terceira seção.

utilizem dados de múltiplas modalidades, o fluxo unidirecional impede que o aprendizado de uma modalidade seja aplicado às demais em sucessivas interações.

Mesmo que esse estudo de caso não almeje desenvolver um modelo com as melhores predições possíveis, é relevante ter em conta essa limitação. Isso evita interpretações equivocadas, deixando claro que os resultados observados não são os melhores alcançáveis, mas sim uma base sólida para o desenvolvimento de modelos produtivos.

6.5 Limitações do estudo de caso

Além das limitações inerentes à arquitetura proposta, a forma com que ela foi aplicada a esse estudo de caso também apresenta algumas deficiências. O conjunto de dados utilizado, por exemplo, é bastante desbalanceado em relação aos tipos de imóvel, já que mais de 80% das amostras correspondem a apartamentos. Esse viés fez com que os modelos testados tivessem um desempenho inferior ao predizer o preço de casas, o que foi evidenciado na análise qualitativa apresentada na Seção 5.2.

Adicionalmente, as ResNet-50 empregadas no pré-processamento da modalidade de imagem não foram treinadas com as fotos coletadas para esse estudo de caso. Como provável consequência, o algoritmo de classificação atribuiu diferentes classes a cada um dos cômodos escolhidos para compor os *grids*. Mesmo após a adoção dessas classes adicionais como *aliases* equivalentes, muitas imagens não foram classificadas como algum dos cômodos desejados, gerando *grids* incompletos ou mesmo vazios para diversas amostras. Isso ocorreu apesar de ser bastante improvável que os anúncios originais não contivessem imagens de todos os cômodos.

Ambas as tarefas de classificação de imagens e extração de características visuais poderiam ter sido beneficiadas pelo “descongelamento” das camadas finais da ResNet-50. Esse ajuste permitiria que as camadas iniciais retivessem características gerais e abstratas, aprendidas no pré-treinamento com os bancos de dados Places365 e ImageNet. Ao mesmo tempo, possibilitaria que as camadas finais fossem retreinadas com as fotos dos imóveis de Porto Alegre, adaptando o modelo ao contexto brasileiro ao incorporar variações culturais, arquitetônicas e de *design*. Como resultado, essa mudança no pré-processamento das imagens poderia ter aprimorado a eficácia da utilização de características visuais nesse estudo de caso.

6.6 Outros experimentos

Por fim, cabe destacar que os 33 experimentos apresentados ao longo desse trabalho representam apenas uma pequena parcela do conjunto total de ensaios realiza-

dos. Na modalidade de características estruturais, por exemplo, foram testados diferentes métodos de escalonamento. Na modalidade de localização geográfica, explorou-se a utilização de latitude e longitude de forma separada. A rede neural do submodelo de imagem, por sua vez, foi extensamente avaliada com diversas configurações, incluindo variações com e sem *dropout*, vários níveis de profundidade e diferentes funções de perda.

Além disso, foram explorados outros métodos para truncar e unir modalidades, bem como diversas outras configurações para a camada de fusão, incorporando modelos de *boosting* como XGBoost (Chen e Guestrin, 2016), LightGBM (Ke et al., 2017) e CatBoost (Prokhorenkova et al., 2018). Como esses experimentos não se mostraram competitivos, optou-se por manter apenas as combinações apresentadas na Seção 4.6.

7. CONCLUSÃO

A multimodalidade desempenha um papel fundamental na avaliação imobiliária, pois aproxima os MAA da experiência humana de avaliação. Ao incorporar modalidades não estruturadas como imagens e texto, esses modelos passam a considerar uma gama mais ampla de fatores, mimetizando a forma como pessoas especialistas estimam o preço de imóveis e resultando, por fim, em predições mais precisas. Até o momento, no entanto, a pesquisa nesse campo concentra-se apenas na modalidade de imagens e em uma paisagem urbana bastante específica — casas de subúrbio na América do Norte.

Nesse contexto, o presente trabalho contribui para preencher essas duas lacunas na literatura, apresentando um estudo de caso focado em uma paisagem diversa e explorando, pela primeira vez, a utilização conjunta das modalidades de imagem e texto. Baseado em um extenso conjunto de dados coletado a partir de anúncios de apartamentos e casas a venda em Porto Alegre, Brasil, o estudo primou por uma análise detalhada dos resultados, tanto em termos quantitativos quanto qualitativos. Empregando um protocolo experimental sistemático, incluindo validação cruzada por *10-fold*, a pesquisa proporcionou uma análise aprofundada das contribuições individuais e conjuntas das diferentes modalidades na predição de preços de imóveis.

Como uma contribuição adicional, foi introduzida uma arquitetura multimodal de componentes intercambiáveis. Essa arquitetura foi desenvolvida para permitir a rápida prototipagem de modelos multimodais, agilizando a compreensão de como diferentes modalidades influenciam nas mais diversas tarefas de aprendizado de máquina.

Os resultados observados evidenciam que as modalidades de imagem e texto desempenharam um papel significativo na otimização dos modelos de avaliação imobiliária, o que foi atestado nas reduções de RMSE. No entanto, a magnitude dessa otimização varia consideravelmente de acordo com a combinação de modalidades utilizada e o preditor adotado. Destacam-se os resultados contundentes alcançados pela modalidade de texto, capaz de reduzir a RMSE de 7 a 27% em comparação com dois modelos de linha de base distintos. O experimento mais bem-sucedido entre os 33 avaliados foi o modelo trimodal de características estruturais, localização geográfica e texto. Combinado ao preditor ERT, essa configuração obteve um MAPE de 0,1915 e um R^2 de 0,822.

Por outro lado, é importante ressaltar as limitações constatadas tanto na arquitetura multimodal proposta quanto na sua aplicação ao estudo de caso. Em especial, identificou-se que a abordagem de transferência de conhecimento aplicada ao pré-processamento de imagens pode ter produzido resultados insatisfatórios, o que pode explicar, em parte, por que a utilização de características visuais não alcançou resultados tão expressivos quanto a modalidade de texto.

Por fim, pode-se responder às questões de pesquisa que norteiam esse trabalho:

- QP1 - De maneira geral, a adição isolada da modalidade de imagem resultou na redução da RMSE em ambas as linhas de base analisadas, variando de 3 a 10%, dependendo do experimento, o que está de acordo com a literatura. A única exceção ocorreu ao aplicar o preditor Random Forest à linha de base de características estruturais e localização geográfica, onde a RMSE apresentou um aumento de 7,48%;

- QP2 - A utilização de descrições textuais revelou-se bastante promissora na otimização de MAAs, resultando em melhorias significativas em praticamente todas as métricas avaliadas. Essa modalidade apresentou o melhor desempenho entre os experimentos unimodais, evidenciando sua capacidade superior de compreender a formação de preços em comparação com modalidades “tradicionais”, que abrangem informações básicas como tipo do imóvel, área e localização;

- QP3 - O método mais promissor para extrair informação relevante de descrições textuais consistiu de uma combinação de TF-IDF com contagem binária e Truncated SVD para reduzir a dimensão dos vetores gerados a 30 elementos;

- QP4 - Identificaram-se características comuns nas imagens e textos que impactaram positivamente a capacidade preditiva dos MAA multimodais. As imagens, em geral, foram capturadas de ângulos que proporcionam uma visão completa dos cômodos. As descrições textuais, por sua vez, destacam aspectos relevantes na decisão de compra de um imóvel, como a proximidade de pontos de interesse e amenidades como churrasqueira e piscina.

Em conclusão, esse trabalho demonstrou a relevância de fazer com que os MAA levem em conta todo o espectro de informações que uma pessoa avaliadora reconheceria. Diante da crescente disponibilidade de dados, nos mais variados formatos, o estudo de caso apresentado fornece evidências convincentes de que é viável explorar modalidades não estruturadas para aprimorar o desempenho de modelos de avaliação imobiliária, lançando nova luz sobre um problema clássico de pesquisa.

7.1 Trabalhos futuros

Alguns tópicos que não foram cobertos pelo escopo desse trabalho podem ser explorados em pesquisas futuras. Além de buscar soluções para as já mencionadas limitações de arquitetura e implementação, algumas dessas perspectivas incluem:

- Estender a metodologia empregada nesse estudo de caso a um conjunto de dados proveniente de um mercado imobiliário diferente do Brasil. Dessa forma, será possível avaliar sua capacidade de generalização;

- Com base nas observações realizadas durante a análise qualitativa, explorar modelos capazes de mensurar o nível de qualidade de um anúncio imobiliário, avaliando a presença de dados úteis e precisos e identificando possíveis falhas humanas no preenchimento das informações;
- Desenvolver um submodelo de imagem que avalie individualmente a foto de cada cômodo, calculando a avaliação da amostra a partir da média das avaliações individuais. Essa estratégia viabilizaria a experimentação com qualquer número de classes, já que não seria necessário organizá-las em um *grid*;
- Aprimorar o submodelo de texto incluindo sequências de dois termos (bigramas) e três termos (trigramas) como parte de seu pré-processamento;
- Testar diferentes estratégias para a camada de fusão.

REFERÊNCIAS BIBLIOGRÁFICAS

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 2481–2495.

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, pp. 423–443.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). SURF: Speeded up robust features. In: *Computer Vision – ECCV 2006*, pp. 404–417. Springer Berlin Heidelberg, Berlin, Heidelberg, Alemanha.

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., and Manjunath, B. S. (2017). Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 320–329. IEEE.

Bessinger, Z. and Jacobs, N. (2016). Quantifying curb appeal. In: *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 4388–4392. IEEE.

Bin, J., Gardiner, B., Li, E., and Liu, Z. (2020). Multi-source urban data fusion for property value assessment: A case study in philadelphia. *Neurocomputing*, vol. 404, pp. 70–83.

Bin, J., Gardiner, B., Liu, Z., and Li, E. (2019). Attention-based multi-modal fusion for improved real estate appraisal: a case study in los angeles. *Multimed. Tools Appl.*, vol. 78, pp. 31163–31184.

Bittencourt, L. F., Parraga, O., Ruiz, D. D., Manssour, I. H., Musse, S. R., and Barros, R. C. (2022). Leveraging textual descriptions for house price valuation. In: Junior, J. C. X. and Rios, R. A., editores, *Intelligent Systems - 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 - December 1, 2022, Proceedings, Part I*, vol. 13653 de *Lecture Notes in Computer Science*, pp. 355–369. Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, vol. 24, pp. 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, vol. 45, pp. 5–32.

Brostow, G. J., Shotton, J., Fauqueur, J., and Cipolla, R. (2008). Segmentation and recognition using structure from motion point clouds. In: *Lecture Notes in Computer Science*, pp. 44–57. Springer Berlin Heidelberg, Berlin, Heidelberg, Alemanha.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and

Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, Praga, República Tcheca. Springer.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pp. 89–96, New York, NY, EUA. ACM Press.

Chen, L., Chen, J., Hajimirsadeghi, H., and Mori, G. (2020). Adapting grad-CAM for embedding networks. In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2783–2792. IEEE.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, San Francisco, CA, EUA. ACM.

Dayhoff, J. E. (1990). *Neural network architectures: an introduction*. Van Nostrand Reinhold Co., New York, NY, EUA.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In: *Proceedings of the Thirteenth International Conference on Machine Learning*, vol. 96, pp. 148–156, San Francisco, CA, EUA. Morgan Kaufmann Publishers Inc.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232.

G, H. G., Walvekar, G., and Kakka, V. (2020). Private real estate: Valuation and sale price comparison 2020. Disponível em: https://www.msci.com/www/research-paper/private-real-estate-valuation/02648015587/. Acesso em: 9 jan. 2022.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, vol. 63, pp. 3–42.

H. Ahmed, E. and Moustafa, M. (2016). House price estimation from visual and textual features. In: *Proceedings of the 8th International Joint Conference on Computational Intelligence*, pp. 62–68, Porto, Portugal. SciTePress.

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, vol. 53, pp. 217–288.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, vol. 51, pp. 1–36.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv Preprint 1704.04861, 2017. 9 p. Disponível em: https://arxiv.org/abs/1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Los Alamitos, CA, EUA. IEEE.

Instituto Brasileiro de Geografia e Estatística (2020). Domicílios brasileiros. Disponível em: https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/21130-domicilios-brasileiros.html. Acesso em: 31 dez. 2021.

International Association of Assessing Officers (2018). Standard on automated valuation models (AVMs). *Journal of Property Tax Assessment & Administration*, vol. 15, pp. 67–101.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, vol. 30, pp. 3146–3154.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization. Published as a conference paper at the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015. ArXiv Preprint 1412.6980, 2017. 15 p. Disponível em: https://arxiv.org/abs/1412.6980.

Kok, N., Koponen, E.-L., and Martínez-Barbosa, C. A. (2017). Big data in real estate? from manual appraisal to automated valuation. *The Journal of Portfolio Management*, vol. 43, pp. 202–211.

Kostic, Z. and Jevremovic, A. (2020). What image features boost housing market predictions? *IEEE Trans. Multimedia*, vol. 22, pp. 1904–1916.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, vol. 25, pp. 1097–1105.

Law, S., Paige, B., and Russell, C. (2019). Take a look around: Using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology*, vol. 10, pp. 1–19.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, vol. 1, pp. 541–551.

Lee, C. and Park, K.-H. (2020). Using photographs and metadata to estimate house prices in south korea. *Data Technol. Appl.*, vol. 55, pp. 280–292.

Liao, W.-C. and Wang, X. (2012). Hedonic house prices and spatial quantile regression. *J. Hous. Econ.*, vol. 21, pp. 16–27.

Liu, S. and Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734, Kuala Lumpur, Malásia. IEEE.

Liu, X., Xu, Q., Yang, J., Thalman, J., Yan, S., and Luo, J. (2018). Learning multi-instance deep ranking and regression network for visual house appraisal. *IEEE Trans. Knowl. Data Eng.*, vol. 30, pp. 1496–1506.

Loft (2020). Diferença de preços de anúncio e venda de imóveis pode chegar a 30%. Disponível em: https://blog.loft.com.br/loft-exame-valor-anunciado-vs-venda/. Acesso em: 2 jan. 2022.

Muhr, V., Despotovic, M., Koch, D., Döller, M., and Zeppelzauer, M. (2017). Towards automated real estate assessment from satellite images with cnns. In: Aigner, W., Moser, T., Blumenstein, K., Zeppelzauer, M., Iber, M., and Schmiedl, G., editores, *Proceedings of the 10th Forum Media Technology and 3rd All Around Audio Symposium, St. Pölten, Austria, November 29-30, 2017*, vol. 2009 de *CEUR Workshop Proceedings*, pp. 14–23. CEUR-WS.org.

Murray, N., Marchesotti, L., and Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, Providence, RI, EUA. IEEE.

Nouriani, A. and Lemke, L. (2022). Vision-based housing price estimation using interior, exterior & satellite images. *Intelligent Systems with Applications*, vol. 14, n. 200081. 8 p.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., and French, N. (2003). Real estate appraisal: a review of valuation methods. *J. Prop. Invest. Fin.*, vol. 21, pp. 383–401.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). *PyTorch: an imperative style, high-performance deep learning library*, pp. 8026–8037. Curran Associates Inc., Red Hook, NY, EUA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.

Pelli Neto, A. and Zárate, L. E. (2003). Avaliação de imóveis urbanos com a utilização de redes neurais artificiais. In: *Anais do IBAPE - XII COBREAP*, Belo Horizonte, MG, Brasil. 14 p.

Peng, N., Li, K., and Qin, Y. (2020). Leveraging multi-modality data to airbnb price prediction. In: *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, pp. 1066–1071. IEEE.

Poursaeed, O., Matera, T., and Belongie, S. (2018). Vision-based real estate price estimation. *Mach. Vis. Appl.*, vol. 29, pp. 667–676.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6639–6649, Red Hook, NY, EUA. Curran Associates Inc.

Rosebrock, A. (2019). Keras, Regression, and CNNs. Disponível em: <https://pyimagesearch.com/2019/01/28/keras-regression-and-cnns/>. Acesso em: 3 de dez. de 2023.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, vol. 27, pp. 379–423.

Snook, W. D. (1998). Advisory Opinion 18. Relatório Técnico, The Appraisal Foundation, Appraiser Standards Board.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, vol. 28, pp. 11–21.

Srirutchataboon, G., Prasertthum, S., Chuangsuwanich, E., Pratanwanich, P. N., and Ratanamahatana, C. (2021). Stacking ensemble learning for housing price prediction: A case study in thailand. In: *2021 13th International Conference on Knowledge and Smart Technology (KST)*, pp. 73–77, Bangsaen, Chonburi, Tailândia. IEEE.

Stivanello, M. E. and Brignoli, R. (2023). An approach based on cnn to residential environment classification focused on real estate business. *Revista de Sistemas e Computação - RSC*, vol. 13. Disponível em: https://revistas.unifacs.br/index.php/rsc/article/view/8251. Acesso em: 9 dez. 2023. 6 p.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Los Alamitos, CA, USA. IEEE Computer Society.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826.

Tan, K.-H. and Lim, B. P. (2018). The artificial intelligence renaissance: deep learning and the road to human-level machine intelligence. *APSIPA Transactions on Signal and Information Processing*, vol. 7, n. e6. 19 p.

Tietz, M., Fan, T. J., Nouri, D., Bossan, B., and skorch Developers (2017). skorch: A scikit-learn compatible neural network library that wraps pytorch. Disponível em: https://skorch.readthedocs.io/en/stable/. Acesso em: 9 dez. 2023.

Whaley III, D. L. (2005). The interquartile range: Theory and estimation. Dissertação (mestrado em ciências matemáticas), East Tennessee State University, Johnson City, TN, EUA. 2005.

Wu, Y. and Zhang, Y. (2021). Mixing deep visual and textual features for image regression. In: *Advances in Intelligent Systems and Computing*, pp. 747–760. Springer International Publishing, Cham, Suíça.

Yao, Y., Zhang, J., Hong, Y., Liang, H., and He, J. (2018). Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Trans. GIS*, vol. 22, pp. 561–581.

You, Q., Pang, R., Cao, L., and Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Trans. Multimedia*, vol. 19, pp. 2751–2759.

Zhang, Y. and Dong, R. (2018). Impacts of street-visible greenery on housing prices: Evidence from a hedonic price model and a massive street view image dataset in beijing. *ISPRS International Journal of Geo-Information*, vol. 7(3), n. 104. 19 p.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, Los Alamitos, CA, EUA. IEEE Computer Society.

Zhao, Y., Chetty, G., and Tran, D. (2019). Deep learning with xgboost for real estate appraisal. In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1396–1401, Xiamen, China. IEEE.

Zhou, B., Lapedriza, A., Torralba, A., and Oliva, A. (2017). Places: An image database for deep scene understanding. *Journal of Vision*, vol. 17, pp. 296–296.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, pp. 487–495, Cambridge, MA, EUA. MIT Press.

APÊNDICE A – DEDUPLICAÇÃO DE AMOSTRAS

Proprietárias ou proprietários buscando vender um imóvel frequentemente o anunciam em mais de uma imobiliária. Essas imobiliárias, por sua vez, ofertam o imóvel não apenas em seu próprio *site*, mas também em agregadores como o que foi utilizado para coletar o conjunto de dados desse estudo de caso. Nessas plataformas, é comum encontrar vários anúncios redundantes para uma mesma propriedade ao realizar uma busca, como por uma rua específica.

Para mitigar o impacto dessas informações duplicadas no desempenho dos modelos, foram aplicadas as seguintes heurísticas para a detecção de amostras duplicadas, preservando-se apenas a amostra com descrição textual mais longa:

1. Anúncios com a mesma URL (variável que não foi mantida no restante do trabalho);
2. Anúncios com o mesmo tipo, área, número de quartos, número de banheiros, conjunto de imagens e preço;
3. Anúncios com o mesmo tipo, conjunto de imagens e preço.

APÊNDICE B – RESULTADOS COMPLETOS DOS EXPERIMENTOS

A Tabela B.1 apresenta os resultados completos para todos os experimentos analisados no estudo de caso. Nota-se que os melhores valores de cada métrica, destacados em negrito, estão inteiramente associados ao preditor ERT.

Tabela B.1: Resultados completos dos 33 experimentos. Melhor valor de cada métrica em negrito. No cabeçalho, as abreviações CE, LG, IM e TX correspondem a características estruturais, localização geográfica, imagem e texto, respectivamente

CE	LG	IM	TX	Preditor	MAPE	MdAPE	R^2	MAE	MSE	RMSE
				Random Forest	0,3261	0,2466	0,6105	151.427,03	45.263.940.342,31	212.753,24
✓				ERT	0,3274	0,2471	0,5987	152.548,07	46.633.053.965,96	215.946,88
				Rede Neural	0,2975	0,2440	0,6005	151.229,41	46.427.676.206,66	215.470,82
				Random Forest	0,5501	0,3620	0,2732	218.118,44	84.475.094.780,69	290.646,00
	✓			ERT	0,5501	0,3621	0,2731	218.116,15	84.482.672.808,40	290.659,03
				Rede Neural	0,4345	0,3636	0,1433	220.081,41	99.562.843.123,64	315.535,80
				Random Forest	0,6177	0,4126	0,1881	233.216,62	94.345.836.707,40	307.157,67
		✓		ERT	0,6001	0,3984	0,1932	228.502,58	93.763.212.461,09	306.207,79
				Rede Neural	0,4890	0,4012	0,0486	236.291,40	110.585.579.857,05	332.544,10
				Random Forest	0,3140	0,2289	0,6408	142.740,77	41.740.283.986,07	204.304,39
			✓	ERT	0,3073	0,2232	0,6548	138.860,60	40.113.883.491,41	200.284,51
				Rede Neural	0,2853	0,2484	0,4539	165.733,92	63.516.291.343,70	252.024,39
				Random Forest	0,2928	0,2152	0,6714	136.475,63	38.179.817.427,63	195.396,56
✓		✓		ERT	0,2840	0,2062	0,6759	133.244,87	37.656.800.482,95	194.053,60
				Rede Neural	0,3011	0,2391	0,6317	146.379,79	42.804.423.563,06	206.892,30
				Random Forest	0,2270	0,1675	0,7764	109.268,35	25.979.939.345,73	161.182,94
✓			✓	ERT	0,2191	0,1603	0,7868	105.363,20	24.765.002.570,93	157.369,00
				Rede Neural	0,2221	0,1881	0,7153	122.467,36	33.082.591.882,63	181.886,21
				Random Forest	0,2287	0,1716	0,7738	110.313,82	26.281.684.397,41	162.116,27
✓		✓	✓	ERT	0,2231	0,1650	0,7801	107.651,44	25.552.972.655,08	159.852,97
				Rede Neural	0,2280	0,1918	0,7084	124.483,68	33.879.304.855,36	184.063,32
				Random Forest	0,2221	0,1387	0,7523	106.828,80	28.768.742.050,46	169.613,51
✓	✓			ERT	0,2217	**0,1342**	0,7337	107.049,49	30.933.683.677,88	175.879,74
				Rede Neural	0,2433	0,1907	0,6948	127.831,28	35.461.807.169,16	188.313,06
				Random Forest	0,2625	0,1900	0,7140	125.002,20	33.233.838.011,16	182.301,50
✓	✓	✓		ERT	0,2298	0,1589	0,7526	111.643,32	28.750.143.229,41	169.558,67
				Rede Neural	0,2392	0,1879	0,7119	124.352,15	33.479.044.482,46	182.972,80
				Random Forest	0,2183	0,1621	0,7873	105.944,17	24.717.855.724,83	157.219,13
✓	✓		✓	ERT	**0,1915**	0,1381	**0,8220**	**94.118,81**	**20.682.622.299,74**	**143.814,54**
				Rede Neural	0,1925	0,1605	0,7814	105.812,51	25.403.275.270,59	159.384,05
				Random Forest	0,2230	0,1660	0,7820	107.913,66	25.327.596.839,18	159.146,46
✓	✓	✓	✓	ERT	0,2014	0,1483	0,8107	98.413,63	21.996.609.969,60	148.312,54
				Rede Neural	0,1952	0,1577	0,7887	104.589,29	24.548.476.575,51	156.679,53

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br