Uma análise comparativa entre regressão linear e *random forest* para predizer valores de aluguéis na Índia

1st Mateus Pereira da Silva *Faculdade de Computação (FACOM)*
Universidade Federal de Uberlândia (UFU)
Uberlândia, Brasil
mateus.silva23@ufu.br

***Resumo*—Este trabalho tem como objetivo prever os valores de aluguel em cidades indianas e suas subdivisões regionais, utilizando técnicas de aprendizado de máquina. Foram comparados dois modelos: regressão linear e *Random Forest*. O conjunto de dados passou por etapas de pré-processamento, como remoção de *outliers* e codificação *one hot encoding* de variáveis categóricas. Os resultados obtidos mostram que o modelo *Random Forest* obteve melhor desempenho, com $R^2 = 0,72$, enquanto a regressão linear atingiu $R^2 = 0.59$.**

Palavras-chave*—Predição de aluguel, Regressão linear, *Random forest

I. Introdução

A Índia, atualmente o país mais populoso do mundo, apresenta uma tendência contínua de crescimento demográfico, o que intensifica a demanda por habitação nas grandes cidades e regiões metropolitanas. Nesse contexto, compreender a dinâmica dos preços de aluguel se torna essencial tanto para tomadores de decisão quanto para o desenvolvimento de soluções inteligentes no setor imobiliário [1].

Este trabalho tem como objetivo analisar os valores de aluguel em diferentes cidades indianas, assim como em suas subdivisões regionais (denominadas áreas), buscando desenvolver modelos preditivos capazes de estimar o valor ideal para cada imóvel. Inspirado em pesquisas anteriores que empregaram técnicas de aprendizado de máquina na estimativa de preços imobiliários [2]–[4], este estudo se concentra na comparação entre dois métodos: a regressão linear e o algoritmo de *Random Forest*.

Para realizar uma análise mais precisa, foi necessário realizar etapas de pré-processamento na base de dados original. Entre essas etapas, se destacam a remoção de *outliers*, que contribuiu para a melhoria do desempenho dos modelos, e a transformação de variáveis categóricas em atributos binários por meio de codificação do tipo *one hot encoding*.

Os resultados obtidos demonstram coerência com os achados em experimentos anteriores reportados na plataforma Kaggle [5], confirmando a eficácia das técnicas avaliadas. Considerando o coeficiente de determinação R^2 como principal métrica de avaliação, os melhores desempenhos observados foram de 0,72 para o modelo *Random Forest* e 0,59 para a regressão linear, indicando um bom nível de explicação da variabilidade dos dados por parte dos modelos ajustados.

A. Organização do trabalho

Este trabalho está organizado da seguinte forma: a Seção II apresenta os trabalhos relacionados, os quais serviram de base teórica e auxiliaram no desenvolvimento das ideias propostas neste estudo. A Seção III descreve o método adotado, abordando aspectos como a base de dados utilizada, as etapas de pré-processamento, os modelos aplicados e as métricas utilizadas para avaliação. Por fim, a Seção IV apresenta os resultados obtidos, acompanhados de uma análise sobre o desempenho dos modelos.

II. Trabalhos correlatos

Esta seção apresenta trabalhos que analisam regressão linear e *random forest*, focando em informações como método e resultados.

Uma análise multifatorial da detecção por fluorescência de hidrocarbonetos totais de petróleo (HTP) no solo, integrando os métodos *Random Forest* (RF) e Regressão Linear Múltipla (MLR). A metodologia utilizou uma tecnologia de imagem de fluorescência auto-desenvolvida para analisar amostras de três tipos distintos de solo, sob duas concentrações de HTP, com diferentes combinações de umidade, matéria orgânica e minerais. Os resultados revelaram que a umidade teve

influência dominante sobre os sinais de fluorescência, seguida por minerais específicos, como a caulinita, que intensificaram os sinais, e a argila, que os enfraqueceu. A MLR apresentou desempenho estatístico superior em termos de R^2, *mean absolute error* (MAE) e *root mean squared error* (RMSE) na maioria das condições experimentais, enquanto o RF foi sensível à concentração de HTP, perdendo robustez em altas concentrações. A avaliação dos métodos também destacou aspectos de sustentabilidade científica, com a MLR se destacando em transparência e interpretabilidade e, por fim, a tecnologia proposta demonstrando baixo impacto ambiental e alta eficiência computacional [6]

A eficiência de Unidades de Terapia Intensiva (UTIs) através de uma comparação entre Causal *Random Forest* (CRF) e Modelagem de Regressão Linear (LRM), com foco em efeitos heterogêneos de fatores organizacionais sobre a razão de eficiência padronizada média (ASER). Ambos os métodos identificaram variáveis significativas, como a proporção de enfermeiros por leito. No entanto, o CRF foi capaz de revelar associações não captadas pelo LRM, como o impacto positivo de programas de treinamento em terapia intensiva. A principal vantagem metodológica do CRF reside em sua habilidade de estimar efeitos condicionais heterogêneos (CATE) sem necessidade de pré-especificação de interações, possibilitando a identificação de subgrupos de UTIs com benefícios específicos, o que torna o modelo especialmente útil em contextos de gestão em saúde pública [7]

Uma abordagem híbrida que integra *Random Forests* com Modelos Lineares Generalizados (GLMs), denominada $RF+$ e $MDI+$. O modelo RF+ substitui as árvores puramente decisórias por ensembles de GLMs regularizados, incorporando conhecimento prévio e adaptando-se ao tipo de tarefa (regressão ou classificação). O $MDI+$ reinterpreta a métrica tradicional de importância de variáveis (*Mean Decrease in Impurity*) como um R^2 de modelos OLS parciais, corrigindo vieses associados à correlação entre variáveis e à baixa entropia. Os experimentos demonstraram que o $RF+$ supera os RFs clássicos em termos de R^2 e $F_1 - score$, enquanto o $MDI+$ produziu classificações mais estáveis e precisas, com ganhos médios superiores a 10% no *AUROC*. Estudos de caso, como a subtipagem de câncer de mama, evidenciaram que o $MDI+$ identificou preditores consistentes e biologicamente relevantes [8].

O *Random Forest Featuring Linear Extensions* (RaFFLE), um *framework* que utiliza *PILOT trees*— árvores de decisão com modelos lineares nas folhas — como base para *ensembles* de *Random Forest*. A metodologia seleciona dinamicamente entre cinco formas funcionais de regressão em cada nó com base no critério de informação bayesiano, promovendo a adaptação entre relações lineares e não lineares. Testado em 136 conjuntos de dados de regressão, o RaFFLE apresentou desempenho preditivo superior, alcançando uma média de R^2 relativa de 0.99, superando métodos tradicionais como CART, XGBoost e regressões penalizadas. Além disso, o modelo mantém complexidade computacional comparável à de RFs clássicos, apresentando-se como uma alternativa eficiente e precisa para problemas com estruturas mistas de relação entre variáveis [9].

Por fim, uma abordagem que estima os modelos lineares condicionais através de uma adaptação do *Random Forest*. A proposta define o resultado Y como uma função linear de X, condicionada a Z, como apresentado na equação 1

$$E[Y|X, Z] = X^T\beta(Z) \tag{1}$$

O modelo emprega árvores com estimação local via Mínimos Quadrados Ordinários (OLS) e utiliza reamostragem para inferência estatística. As simulações de Monte Carlo evidenciaram que o estimador é acurado para interceptos variáveis e apresenta distribuição residual aproximadamente normal, embora a cobertura para parâmetros de inclinação seja subestimada, principalmente em regiões periféricas da distribuição de Z. Aplicações empíricas revelaram padrões espaciais heterogêneos de crescimento econômico entre municípios brasileiros, destacando a relevância da abordagem para investigação de efeitos parciais heterogêneos em contextos geograficamente estruturados [10].

III. Método

Nesta seção, são descritos a base de dados empregada no estudo, os procedimentos de pré-processamento aplicados e os algoritmos de aprendizado de máquina utilizados na análise. As abordagens modeladas incluem a Regressão Linear Múltipla e o algoritmo *Random Forest*.

A. Base de dados

A base de dados utilizada neste estudo compreende informações sobre 4746 imóveis residenciais disponíveis para aluguel na Índia, incluindo casas, apartamentos e *flats*. Este *dataset* possui informações como número de cômodos (quarto, sala e cozinha estão incluídas, no *dataset* é denominado de BHK - bedroom, hall e kitchen), valor do aluguel, tamanho do imóvel (em pés quadrado), número de andares, tipo de área (super área, área útil ou construída) localização, cidade, status de mobília (mobilhado, semi-mobiliado ou não mobilhado), perfil do inquilino, número de banheiros e contrato de

negociação [11]. Nos experimentos, a base de dados foi dividida em 90% para treinos e 10% para teste. A Tabela I apresenta os atributos dessa base de dados (com os nomes originais), com informações como números de amostras não nulas e o tipo do atributo.

Tabela I
DESCRIÇÃO DAS COLUNAS DO DATASET DE HABITAÇÃO

Coluna	Tipo de dado	Valores não nulos
Posted On	*object*	4746
BHK	*int64*	4746
Rent	*int64*	4746
Size	*int64*	4746
Floor	*object*	4746
Area Type	*object*	4746
Area Locality	*object*	4746
City	*object*	4746
Furnishing Status	*object*	4746
Tenant Preferred	*object*	4746
Bathroom	int64	*4746*
Point of Contact	*object*	4746

B. Pré-processamento

O *dataset* não apresentou grandes desafios em relação a valores faltantes ou duplicados, por esse motivo não foi necessário utilizar nenhuma técnica para lidar com esses problemas. Em relação aos *outliers*, utilizando como métrica de identificação a média e o desvio padrão, foram encontrados 173 amostras(o que representa 3.65% da quantidade de total de amostras), isso foi aplicado apenas nos atributos aluguel (valor), tamanho do imóvel, quantidade de banheiros e quantidade de cômodos. Essa métrica verifica quais valores estão fora do desvio padrão esperado e os classifica como *outliers* [12]. Dessa forma, foi gerado quatro versões da base de dados: (i) a base de dados original, sem nenhuma alteração; (ii) a base de dados com os atributos originais (ou seja, são utilizados apenas os valores numéricos nas análises), mas sem a presença de *outliers*; (iii) Base de dados com atributos modificados utilizando o *one hot encoder* e com a presença de *outliers*; e (iv) base de dados com atributos modificados e sem a presença de *outliers*. Essas quatro formas da base de dados, auxiliam a compreensão de como os modelos estão se comportando na presença de *outliers* e utilizando apenas as variáveis numéricas (posteriormente, testando também com os atributos categóricos transformados em atributos booleanos).

C. Modelos

Para predição foi selecionado a regressão linear e *random forest*. A escolha se deu por dois motivos. O primeiro, considerando as características do *dataset* que possui apenas onze características. Ou seja, utiliza técnicas como aprendizado profundo (lida bem com grandes volumes de dados com muitas *features*) pode causar *overfitting* e técnicas mais simples como *naives bayes* (lida bem com variáveis independentes) pode não capturar informações oriunda das características. Dessa forma, regressão linear apresenta uma performance em casos com *dataset* com valores bem relacionados entre si e árvores de decisão auxiliam no processo de explicabilidade do modelo [13], [14]. Em segundo plano, baseado nos trabalhos relacionados (veja na Seção II) que intercalam o uso das duas técnicas para realização predições. Nos trabalhos relacionados, as técnicas são utilizadas em conjunto para explorar as qualidades de uns todos métodos, neste trabalho elas são utilizadas como contraste uma da outra. Ou seja, são comparadas para verificar quais se adequam melhor as modificações realizadas na base de dados [14].

1) Regressão linear: É uma técnica estatística que utilizada para modelar relações entre variáveis dependentes Y e uma ou mais variáveis independentes X. Essa técnica busca estimar uma relação entre os dados, permitindo previsões sobre novos valores de X. De certo modo, uma regressão linear funciona como uma equação de primeiro grau. Na qual, dado um valor para Y é possível encontrar os valores para X e baseados nos valores de X também é possível encontrar os valores de Y [15].

2) Random Forest: é um método que utiliza um conjunto de classificadores baseados em árvore de decisão, combinando as previsões de dessas árvores para obter o melhor resultado. Cada árvore é gerada com base nos valores de um conjunto independente de vetores aleatórios. Esses vetores são gerados a partir de uma distribuição fixa, ao contrário da abordagem adaptativa usado no *AdaBoost*, em que a distribuição de probabilidade é variada para lidar com exemplos difíceis de classificar. *Random forest* lida bem com bases de dados com alto nível de características, enquanto árvores de decisão simples (considerando variados métodos de implementação) apresentam dificuldades com alto nível de características [14]. Neste trabalhos foi utilizado 1000 árvores no bosque, foram feitos testes com valores de 20, 50, 100 e 200, no entanto, 1000 árvores apresentou os melhores resultados.

D. Métricas de avaliação

As métricas de avaliação foram baseadas nos trabalhos correlatos encontrados (veja em Seção II).

1) R^2: ou coeficiente de terminação, mede o quanto a variável dependente é explicada pelo modelo de regressão. Em uma regressão linear simples (o caso deste trabalho), R^2 coincide com o quadrado do coeficiente de correlação de Pearson. Geralmente, os valores podem variar entre 0 e 1 (também pode ser expressa de forma percentual), mas modelos não lineares podem assumir valores negativos. Quanto mais próximo de 1, mais capaz o modelo é de "capturar" variações da variável dependente. Porém, R^2 muito alto também pode representar *overfitting*, por isso é importante analisar outras métricas [16], [17]. A Equação 2 apresenta a fórmula para calcular essa métrica.

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

2) Erro percentual absoluto médio: do inglês *mean absolute percentual error* (MAPE) mede a precisão de previsões ao calcular a média dos valores absolutos dos erros percentual entre os valores reais e os valores estimados. O MAPE é utilizado tanto em estudos de regressão porque auxilia na interpretação [16], [17]. A Equação 3 apresenta a fórmula para calcular essa métrica.

$$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)} \tag{3}$$

3) Erro quadrático médio: do inglês *mean squared error* (MSE) pode ser definido como a média aritmética dos quadrados das diferenças entre os valores previstos e os valores observados. Dessa forma, essa métrica penaliza erros grandes por causa da elevação ao quadrado. Assim, o MSE é especialmente sensível a *outliers* [16], [17]. A Equação 4 apresenta a fórmula para calcular essa métrica.

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

4) Raiz do erro quadrático médio: do inglês *root mean square error* (RMSE) pode ser definido como a raiz quadrada do MSE. Ela pode ser interpretada como o desvio padrão da diferença entre os valores observados e os valores previstos. Assim como MSE, é sensível a *outliers* [16], [17]. A Equação 5 apresenta a fórmula para calcular essa métrica.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

IV. Resultados

Conforme pode ser observado nas Tabelas II e III, que apresentam os resultados obtidos tanto para o modelo de regressão linear quanto para o modelo baseado em árvore de decisão (*random forest*), as modificações aplicadas à base de dados tiveram um impacto significativo no desempenho dos modelos. Essas alterações incluírem remoção de *outliers* e transformação dos dados categóricos. Essas etapas de pré-processamento se mostraram essenciais para melhorar a capacidade preditiva dos modelos, refletindo-se em métricas de avaliação mais favoráveis. É importante destacar que determinadas métricas, como o MSE e a RMSE, são sensíveis à presença de *outliers*. A existência deles pode prejudicar a avaliação da performance do modelo, uma vez que esses pontos influenciam o cálculo dessas métricas, elevando seus valores e, por consequência, indicando um desempenho inferior. Portanto, o tratamento e a remoção ou mitigação desses valores atípicos são passos cruciais para garantir a robustez e a confiabilidade dos resultados dos modelos preditivos apresentados.

Tabela II
Resultados Regressão linear

	Métricas			
dataset	MSE	$MAPE$	R^2	$RMSE$
(i)	$4.5X10^9$	$1.4X10^0$	$2.9X10^{-1}$	$6.7X10^{-1}$
(ii)	$7.8X10^8$	$7.5X10^{-1}$	$3.1X10^{-1}$	$2.8X10^4$
(iii)	$3.0X10^9$	$1.0X10^0$	$4.4X10^{-1}$	$5.5X10^4$
(iv)	$4.0X10^8$	$7.0X10^{-1}$	$5.9X10^{-1}$	$2.0X10^4$

Tabela III
Resultados random forest

	Métricas			
dataset	MSE	$MAPE$	R^2	$RMSE$
(i)	$4.0X10^8$	$7.0X10^{-1}$	$5.9X10^{-1}$	$2.0X10^4$
(ii)	$8.0X10^8$	$7.0X10^{-1}$	$3.0X10^{-1}$	$2.8X10^4$
(iii)	$1.9X10^9$	$4.1X10^{-1}$	$6.7X10^{-1}$	$4.3X10^4$
(iv)	$2.8X10^8$	$3.7X10^{-1}$	$7.2X10^{-1}$	$1.7X10^4$

Os resultados obtidos nas métricas de desempenho dos modelos são consistentes e corroboram os resultados apresentados em outros experimentos na plataforma Kaggle [5]. Isso sugere que os valores encontrados estão de acordo com a complexidade da base de dados e estão alinhados com outras análises realizadas. Dessa forma, apesar das limitações observadas, os resultados obtidos possuem validade e contribuem para a compreensão do comportamento dos modelos preditivos aplicados a este conjunto de dados.

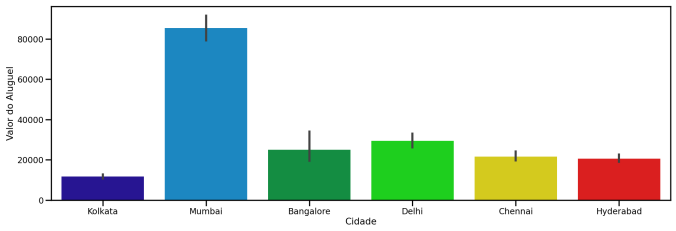

Figura 1. Média de Valor do Aluguel por Cidade

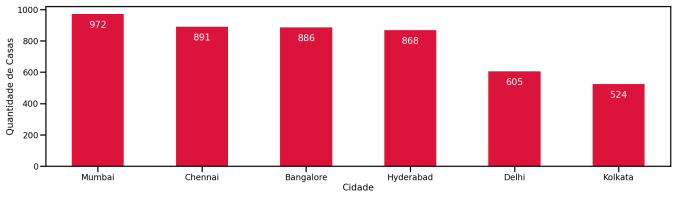

Figura 2. Número de casas disponíveis para aluguel por cidade

A base de dados utilizada neste estudo apresenta complexidades para a análise e o desempenho dos modelos preditivos. Um dos principais desafios é o baixo número de amostras disponíveis, principalmente quando o recorte é feito por cidade. Além disso, a granularidade dos dados também se estende aos setores dentro das cidades, que no *dataset* são referidos como "áreas". Cada cidade é composta por múltiplas áreas, e a quantidade reduzida de amostras em cada uma dessas áreas aumenta a complexidade da análise. Dentro de uma mesma cidade, os valores por área podem apresentar uma grande volatilidade. Dessa forma, dependendo da área considerada, os valores dos aluguéis podem variar significativamente, o que reforça a complexidade na modelagem e na previsão precisa dos preços dentro dos diferentes contextos urbanos. A Figura 1 apresenta essa relação, em que *Mumbai* possui um valor médio de para aluguel muito superior aos das outras cidades e mesmo considerando apenas a cidade de *Mumbai*, existe uma notável variação dos preços dos alugueis.

Esta base de dados possui seis cidades, sendo elas *Mumbai* que possui 972 amostras, *Chennai* que conta com 891 amostras, *Bangalore* com 886 amostras, *Hyderabad* com 868 amostras, *Delhi* com 605 amostras e *Kolkata* com 524 amostras, conforma apresentado na Figura 2. Outro fator que influenciou o desempenho dos modelos é que a maior parte das variáveis presentes no conjunto de dados é de caráter categórico, e isso influencia na quantidade de variáveis que podem ser utilizadas nos dois modelos. A predição com base de dados original (sem remoção de *outliers* e sem transformações) foi efetuada considerando apenas três variáveis numéricas contínuas, que são: o tamanho do imóvel, a quantidade de cômodos e a quantidade de banheiros. Mesmo considerando a aplicação do *one hot encoder*, técnica que transforma variáveis categóricas em colunas binárias para possibilitar sua utilização em modelos de aprendizado de máquina, foi observado que essas variáveis codificadas apresentam baixa correlação com o valor do aluguel, embora isso tenha melhorado os resultados. Vale ressaltar que a técnica de *one hot encoder* utilizada, também aumenta a dimensionalidade da base de dados, logo precisa ser usa com cautela.

Referências

[1] A. Gandhi, "Economic implications of population growth in india," *Innovative Research Thoughts*, vol. 10, pp. 80–91, 06 2024.

[2] Z. Cai and Y. Zhao, "House rent analysis with linear regression model—— a case study of six cities in india," *Highlights in Science, Engineering and Technology*, vol. 38, pp. 576–582, 03 2023.

[3] P. Waddell and A. Besharati-Zadeh, "A comparison of statistical and machine learning algorithms for predicting rents in the san francisco bay area," 2020. [Online]. Available: https://arxiv.org/abs/2011.14924

[4] T. Yoshida and H. Seya, "Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset," 2021. [Online]. Available: https://arxiv.org/abs/2107.12539

[5] S. Banerjee and K. D. Contributors, "House rent prediction dataset," https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset, 2025, acesso em: 4 de julho de 2025.

[6] G. Shi, R. Yang, N. Zhao, G. Yin, and W. Liu, "Multifactorial analysis of fluorescence detection for soil total petroleum hydrocarbons using random forest and multiple linear regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 264, p. 105444, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743925001297

[7] L. S. Bastos, S. A. Wortel, F. Bakhshi-Raiez, A. Abu-Hanna, D. A. Dongelmans, J. I. Salluh, F. G. Zampieri, G. Burghi, S. Hamacher, F. A. Bozza, N. F. de Keizer, and M. Soares, "Comparing causal random forest and linear regression to estimate the independent association of organisational factors with icu efficiency," *International Journal of Medical Informatics*, vol. 191, p. 105568, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505624002314

[8] A. Agarwal, A. M. Kenney, Y. S. Tan, T. M. Tang, and B. Yu, "Integrating random forests and generalized linear models for improved accuracy and interpretability," 2025. [Online]. Available: https://arxiv.org/abs/2307.01932

[9] J. Raymaekers, P. J. Rousseeuw, T. Servotte, T. Verdonck, and R. Yao, "A powerful random forest featuring linear extensions (raffle)," 2025. [Online]. Available: https://arxiv.org/abs/2502.10185

[10] R. Masini and M. Medeiros, "Balancing flexibility and interpretability: A conditional linear model estimation via random forest," 2025. [Online]. Available: https://arxiv.org/abs/2502.13438

[11] S. Banerjee, "House rent prediction dataset," https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset, 2022, accessed: 2025-07-02. [Online]. Available: https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset

[12] A. Seheult, P. Green, P. Rousseeuw, and A. Leroy, “Robust regression and outlier detection.” *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 152, p. 133, 01 1989.

[13] N. Nurhachita and E. S. Negara, “A comparison between deep learning, naïve bayes and random forest for the application of data mining on the admission of new students,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, p. 324, 06 2021.

[14] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining: Pearson new international edition*, ser. Pearson custom library. London, England: Pearson Education, Jul. 2013.

[15] B. Deaner and S. Kwon, “Extrapolation in regression discontinuity design using comonotonicity,” 2025. [Online]. Available: https://arxiv.org/abs/2507.00289

[16] P. Bruce and A. Bruce, *Estatística prática para cientistas de dados*. Alta Books, 2019.

[17] I. D. Dinov, *Data science and predictive analytics*, 1st ed. Cham, Switzerland: Springer International Publishing, Sep. 2018.