Análise preditiva de preços imobiliários utilizando técnicas de aprendizado supervisionado

Lucas Barcelos Mendes[1], Murilo Henrique Tank Fortunato[2]

1 MBA em Data Science e Analytics – Escola Superior de Agricultura Luiz de Queiroz- Universidade de São Paulo (USP)
Piracicaba, SP – Brasil

[2] Professor orientador do MBA em Data Science e Analytics-
Programa de Educação Continuada em Economia e Gestão de Empresas (PECEGE)
Piracicaba, SP - Brasil

barcelosmendes_lucas@hotmail.com, mtank@live.com

Abstract. *This study addressed the application of supervised machine learning techniques to the Ames Housing dataset, exploring the prediction of residential property values in Ames, Iowa. The study emphasized exploratory data analysis (EDA), revealing insights into the distribution of variables, patterns, and anomalies crucial for predictive modeling. Six models were developed using regression techniques, including Linear Regression, "Elastic Net," "Random Forest," and "XGBoost." These techniques were applied sequentially, highlighting the evolution of accuracy and complexity in the predictive models. Data processing, including handling missing data, "outliers," categorical variable encoding, and variable transformation, was crucial for improving the models. The methodology provided insights into the trade-off between simplicity and accuracy, with XGBoost yielding the best results. The study also addressed challenges in real-world data modeling, such as non-normal residual distribution, heteroscedasticity, collinearity, and dimensionality. The results obtained were compared to the literature, confirming the effectiveness of the proposed method for regression prediction with real-world data.*

Resumo. Este trabalho abordou a aplicação de técnicas de aprendizado supervisionado no conjunto de dados Ames Housing, explorando a previsão de valores de propriedades residenciais em Ames, Iowa. O estudo enfatizou a análise exploratória de dados (AED), revelando insights sobre a distribuição de variáveis, padrões e anomalias cruciais para a modelagem preditiva. Foram desenvolvidos 6 modelos empregando técnicas de regressão, incluindo Regressão Linear, "Elastic Net", "Random Forest" e "XGBoost". Estas técnicas foram aplicadas sequencialmente, destacando a evolução da precisão e complexidade dos modelos preditivos. O processamento dos dados, incluindo tratamento de dados ausentes, “outliers”, codificação de variáveis categóricas e transformação de variáveis, foi crucial para o aprimoramento dos modelos. A metodologia proporcionou insights

Cadernos do IME - Série Informática
e-ISSN: 2317-2193 (online)
DOI: 10.12957/cadinf.2026.91017

sobre o “trade-off” entre simplicidade e precisão, com o XGBoost apresentando os melhores resultados. O estudo também abordou desafios da modelagem de dados reais, como distribuição não normal dos resíduos, heterocedasticidade, colinearidade e dimensionalidade. Os resultados obtidos foram comparados à literatura, corroborando a eficácia do método proposto para previsão em regressão com dados reais.

1. INTRODUÇÃO

O aprendizado de máquina representa um amplo campo em tecnologia da informação, estatística, probabilidade, inteligência artificial, psicologia, neurobiologia e muitas outras disciplinas [Nasteski, 2017]. Os algoritmos de "machine learning" utilizam uma variedade de métodos estatísticos, probabilísticos e de otimização para aprender com experiências anteriores e detectar padrões úteis em conjuntos de dados variados (possivelmente grandes, não estruturados e complexos) [Uddin et al., 2019]. Existem duas categorias que permitem uma ampla classificação desse tema em Ciência de

Dados, segundo Hyde et al. [2019]: aprendizado não supervisionado e aprendizado supervisionado. Para problemas de classificação e regressão, de acordo com Uddin et al. [2019], algoritmos de aprendizado de máquina supervisionado são particularmente úteis. Quanto ao objetivo, eles envolvem algoritmos que usam variáveis de entrada (independentes) para prever uma classificação alvo (variável dependente). Entretanto, ao contrário da aprendizagem não supervisionada, a aprendizagem supervisionada envolve conjuntos de dados onde a previsão alvo é conhecida no momento do treinamento. Um modelo é considerado bem-sucedido quando pode prever com precisão o resultado desejado para um conjunto de dados de treinamento com um certo grau de precisão e ser generalizado para novos conjuntos de dados além daqueles usados para treinar o modelo [Hyde et al., 2019].

Alguns conjuntos de dados disponíveis publicamente permitem a aplicação e análise de técnicas de aprendizado de máquina, como o conjunto Ames Housing, criado por Cock [2011]. Este "dataset" contém uma série de características envolvidas na avaliação dos valores de imóveis residenciais em Ames, Iowa, catalogadas entre os anos de 2006 e 2010. Segundo o responsável pela aquisição dos dados, esse conjunto é ideal para aplicação e estudo em projetos de regressão com informações reais, contendo um número significativo de variáveis e exigindo o desenvolvimento de técnicas que vão além dos algoritmos simplificados mais usualmente utilizados durante as etapas de ensino e aprendizagem em Ciência de Dados. Como exemplo, alguns métodos clássicos de aprendizado supervisionado que podem ser aplicados a problemas como os apresentados em Ames Housing são: regressão linear simples, regressão linear múltipla, regressão "Elastic Net", "Random Forest" e "Extreme Gradient Boosting" (XGBoost).

A regressão linear (simples ou múltipla) é um método de avaliação e modelagem de dados que estabelece relações lineares entre variáveis dependentes e independentes, modelando essas relações a partir da aprendizagem até os resultados de treino e validação. Em suma, a regressão linear simples é uma modelagem com única variável independente,

enquanto a regressão linear múltipla apresenta mais de uma variável independente [Maulud e Abdulazeez, 2020].

"Elastic Net" é uma técnica de regressão que combina as propriedades de duas outras técnicas: regressão Ridge e Lasso. Ela oferece uma solução particularmente eficaz na redução da variância e na seleção de variáveis, sendo útil em cenários onde o número de preditores é relativamente grande ou superior ao número de observações. Este método se destaca também por sua capacidade de lidar com multicolinearidade e por realizar seleção de variáveis automaticamente [Zou e Hastie, 2005].

Já o "Random Forest" é um método de "machine learning" baseado em árvores de decisão que utiliza a técnica "ensemble learning", ou seja, combina o resultado de múltiplas árvores de decisão para melhorar a precisão das previsões. Algumas das grandes vantagens do "Random Forest" são sua capacidade de reduzir o risco de "overfitting" e sua alta eficácia em lidar com grandes conjuntos de dados, especialmente com uma alta dimensionalidade de variáveis [Breiman, 2001].

O método XGBoost é um sistema escalonável de aumento de árvore de decisão amplamente utilizado por cientistas de dados, com capacidade de fornecer resultados de alta qualidade para muitos problemas de regressão. Ele pode processar um conjunto de dados muito grande usando uma quantidade mínima de recursos computacionais por meio de padrões específicos, destacando-se, assim, por sua velocidade e desempenho [Chen e Guestrin, 2016].

Este trabalho abordou a aplicação prática de técnicas de aprendizado de máquina supervisionado, analisando e comparando as diferentes técnicas citadas de modelagem preditiva no conjunto de dados Ames Housing. O objetivo é demonstrar a aplicabilidade dessas técnicas em um contexto real, além de fornecer insights sobre como diferentes abordagens podem influenciar principalmente a precisão e a complexidade dos modelos preditivos.

2. MATERIAL E MÉTODOS

Esta seção descreve os materiais e metodologias utilizados nesta pesquisa, com o intuito de garantir clareza, reprodutibilidade e potencial aplicação dos métodos por outros pesquisadores na área de Ciência de Dados.

2.1 Conjunto de Dados Ames Housing

O principal material utilizado neste estudo foi o conjunto de dados "Ames Housing", que contém uma lista abrangente de características associadas à avaliação dos valores de propriedades residenciais em Ames, Iowa. O conjunto de dados possui 2.930 observações e várias variáveis explicativas, sendo 23 nominais, 23 ordinais, 14 discretas e 20 contínuas, todas envolvidas na avaliação dos valores dos imóveis. A base de dados é de domínio público e pode ser obtida em plataformas como o Kaggle, tendo sido publicada inicialmente

no "Journal of Statistics Education" [Cock, 2011].

A primeira etapa deste estudo foi realizar uma análise exploratória de dados (AED) para entender as características do conjunto de dados, suas distribuições e os possíveis relacionamentos entre as variáveis. De acordo com Sahoo et al. [2019], a AED é uma abordagem que visa compreender as informações que podem ser extraídas do conjunto de dados, as hipóteses iniciais para o estudo, e identificar aspectos cruciais como: número de colunas e linhas, tipo dos dados, necessidade de tratamento de dados ausentes, transformação de variáveis, identificação de "outliers", compreensão das distribuições e outras estatísticas das variáveis.

Quando necessário, para codificação de variáveis categóricas, foram utilizadas duas meto-odologias clássicas da literatura, conforme Al-Shehari e Alsowail [2021]: codificação de rótulo para variáveis ordinais e codificação "one-hot" para transformar essas variáveis em numéricas.

Para normalização de variáveis, quando necessário, foi aplicada a transformação logarítmica, que apresentou bons resultados para o conjunto de dados, conforme a literatura [Kaggle, 2021]. Especificamente, foi aplicada a transformação logarítmica natural às variáveis com alta assimetria, visando aproximar sua distribuição a uma normal. No caso do conjunto de dados utilizado, a transformação foi implementada com a biblioteca numpy, utilizando uma função que adiciona 1 a cada valor antes de aplicar o logaritmo, garantindo a consistência da transformação mesmo para valores próximos de zero.

2.2 Métodos de Modelagem Supervisionada

Após a AED, foram empregadas diferentes técnicas de regressão para desenvolver modelos preditivos dos preços dos imóveis no conjunto de dados Ames Housing. Especificamente, utilizaram-se as técnicas de regressão linear simples e múltipla, além dos algoritmos "Elastic Net", "Random Forest" e XGBoost. Cada método foi selecionado pela sua relevância para os objetivos da pesquisa, permitindo uma comparação abrangente da precisão preditiva e da complexidade dos modelos.

A regressão linear simples foi utilizada como ponto de partida para a construção dos modelos devido à sua baixa complexidade, propondo a estimativa da variável de saída com base em apenas uma variável dependente. Modelos de regressão linear múltipla foram implementados em seguida, visando aumentar a precisão dos resultados. Neste contexto, também foi aplicada a técnica "forward stepwise", que seleciona iterativamente as melhores variáveis, reduzindo problemas de colinearidade e "overfitting". A técnica "forward stepwise" é um método iterativo onde o modelo começa sem variáveis e, a cada etapa, adiciona-se a variável que mais melhora o modelo, até que não haja mais melhorias significativas [Emmert-Streib e Dehmer, 2019].

Uma outra abordagem para lidar com os problemas mencionados foi a implementação do modelo "Elastic Net". Esse modelo utiliza técnicas de regularização que ajudam a minimizar

o impacto de características do conjunto de dados, como o número elevado de variáveis preditoras em relação ao número de observações e a presença de variáveis categóricas com distribuições desiguais entre suas categorias [Zou e Hastie, 2005].

Os modelos finais escolhidos, com maior complexidade, foram "Random Forest" e XGBoost. Esses modelos, baseados em árvores de decisão, têm uma grande capacidade de generalização, lidando melhor com categorias desbalanceadas, pois podem aprender a importância das categorias raras sem precisar de reamostragem [Breiman, 2001]. Para a verificação e comparação dos desempenhos de cada modelo, foram utilizadas as métricas R^2 ajustado e "Root Mean Square Error" (RMSE), com o conjunto de dados dividido em subconjuntos de treinamento e teste (80% para treino e 20% para teste) para avaliar a generalização e a precisão preditiva dos modelos. Para validar a significância global dos modelos, foi utilizada a estatística F. A hipótese nula testada foi a de que todos os coeficientes de regressão são iguais a zero, indicando que as variáveis independentes não explicam a variável dependente.

O intervalo de confiança utilizado foi de 95%.
As representações gráficas dos resultados de modelagem neste estudo foram feitas com valores normalizados, a fim de facilitar a comparação entre diferentes gráficos apresentados.

2.3 Kaggle

O Kaggle é uma plataforma voltada para concursos de modelagem preditiva, onde cientistas de dados competem para resolver problemas utilizando dados reais. A plataforma permite que os participantes testem suas habilidades em uma comunidade global e serve como referência para comparar o desempenho de modelos, combinando teoria com aplicação prática [Tauchert et al., 2020].
Neste estudo, o desafio "House Prices: Advanced Regression Techniques" do Kaggle foi utilizado como base comparativa para avaliar as metodologias desenvolvidas, especificamente com relação às previsões de preços de imóveis do conjunto de dados Ames Housing [Kaggle, 2021].

2.4 Ferramentas e Tecnologias

O desenvolvimento deste trabalho foi realizado em Python, uma linguagem amplamente utilizada em aplicações de aprendizado de máquina devido à sua simplicidade, tempo de desenvolvimento reduzido e sintaxe uniforme [Mehare et al., 2023].

Além disso, foram utilizadas algumas bibliotecas essenciais para projetos de Ciência de Dados e aprendizado de máquina. A seguir, são apresentadas as principais bibliotecas utilizadas para modelagem preditiva e elaboração de gráficos neste estudo:

A biblioteca **Statsmodels**, voltada para análise estatística e econométrica em Python, foi utilizada para os modelos de regressão linear simples e múltipla [Seabold e Perktold, 2010].

A biblioteca **Scikit-Learn**, considerada uma ferramenta referência para aprendizado de máquina e com uma ampla gama de algoritmos, foi essencial para a criação dos modelos "Elastic Net" e "Random Forest" [Abraham et al., 2014].

Para o modelo XGBoost, foi utilizada a biblioteca homônima, conhecida por sua flexibilidade e eficiência em resolver problemas relacionados à Ciência de Dados [Vaudevan et al., 2022].

A visualização gráfica dos resultados foi realizada utilizando a biblioteca **Matplotlib**, que facilitou a compreensão das distribuições e estatísticas dos dados e do desempenho dos modelos. Essa ferramenta de visualização é caracterizada por uma linguagem concisa, alta precisão e código simples de fácil entendimento [Cao et al., 2021].

2.5 Análise Exploratória dos Dados

A AED forneceu informações valiosas sobre a estrutura do conjunto de dados, revelando padrões, anomalias e distribuições das variáveis, aspectos fundamentais para a fase de modelagem. O conjunto Ames Housing é composto por 2.930 observações, 79 variáveis preditoras e 1 variável resposta ("Sale Price"). A Tabela 1 apresenta as estatísticas descritivas da variável dependente.

Tabela 1. Estatísticas descritivas de "Sale Price"

Estatística	Valor (em dólares americanos)
Média	180.411,58
Desvio Padrão	78.554,86
Mínimo	12.789,00
Primeiro Quartil	129.500,00
Segundo Quartil	160.000,00
Terceiro Quartil	213.500,00
Máximo	625.000,00

Fonte: Resultados originais da pesquisa

Os valores obtidos indicam uma possível violação da normalidade na distribuição de "Sale Price", especialmente devido ao valor máximo ser significativamente superior à média. A Figura 1 ilustra visualmente uma assimetria com inclinação à direita. Os valores de curtose e assimetria para a variável são, respectivamente, 1,74 e 5,11. De acordo com Demir [2022], um valor de assimetria superior a 0 indica uma distribuição com uma cauda mais longa à direita, enquanto valores positivos de curtose sugerem uma distribuição com caudas mais pesadas e um pico mais acentuado em comparação à distribuição normal. Em outras palavras, esses resultados indicam a presença de "outliers" ou valores muito elevados de "Sale Price", que distorcem sua distribuição e podem representar um desafio para as previsões.

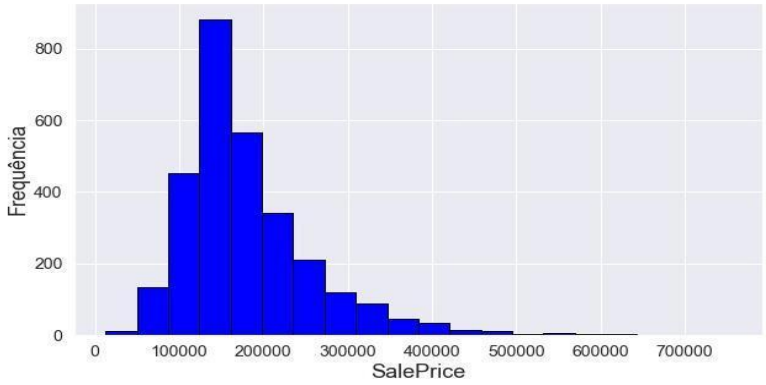

Figura 1. Histograma de "Sale Price"(Fonte: Resultados originais da pesquisa)

A matriz de correlação de Pearson é uma ferramenta utilizada para identificar a colinearidade entre as variáveis do problema, conforme destacado por Dufera et al. [2023]. Segundo o estudo, essa matriz estatística mede a força e a direção das relações lineares entre pares de variáveis. Cada célula da matriz contém um coeficiente de correlação, que varia de -1 a 1, refletindo a intensidade e o tipo da relação (positiva ou negativa) entre as variáveis. Valores próximos a 1 ou -1 indicam uma forte correlação, enquanto valores próximos a 0 sugerem uma correlação fraca ou inexistente.

A Figura 2 apresenta essa matriz para as variáveis numéricas do conjunto Ames Housing. Cores mais frias indicam uma forte correlação inversa, o que significa que, quando os valores de uma variável aumentam, os valores da variável correlacionada diminuem. Por outro lado, cores mais quentes representam uma forte correlação direta, ou seja, o aumento em uma variável está associado ao aumento na variável correlacionada. É importante destacar que a presença de variáveis colineares pode levar ao "overfitting" nos modelos preditivos [Emmert-Streib e Dehmer, 2019].

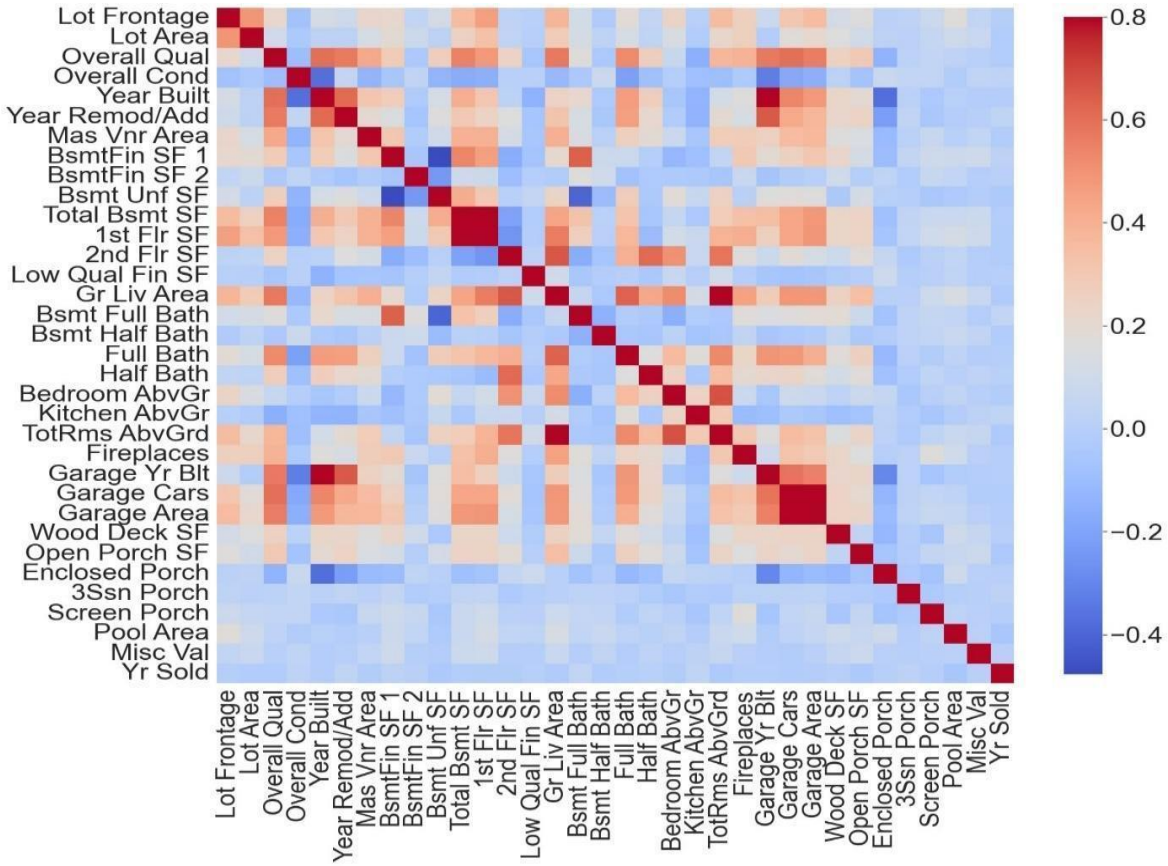

Figura 2. Matriz de Correlação de Pearson
Fonte: Resultados originais da pesquisa

Sob outra ótica, os coeficientes de correlação de Pearson entre as variáveis numéricas preditoras e a variável de resposta oferecem uma indicação de quais atributos do conjunto de dados possuem maior capacidade de prever o comportamento dos preços dos imóveis. A Tabela 2 apresenta os coeficientes correspondentes. Observa-se que a variável "Overall Qual" se destaca em sua correlação com "Sale Price", o que está em consonância com a ideia de que os preços dos imóveis são definidos, em grande parte, pelas suas condições gerais (avaliadas pela variável "Overall Qual").

Outro aspecto importante capturado durante a análise exploratória dos dados foi a ausência de dados no conjunto. A Figura 3 ilustra a quantidade de valores nulos em Ames Housing para cada uma das variáveis que apresentam ao menos 1 valor nulo.

Tabela 2. Coeficientes de Correlação de Pearson dos preditores numéricos com "Sale Price"

Variável Independente	Correlação com "Sale Price"
Overall Qual	0,80
Gr Liv Area	0,71
Garage Cars	0,65
Garage Area	0,64
Total Bsmt SF	0,63
1st Flr SF	0,62
Year Built	0,56
Full Bath	0,55
Year Remod/Add	0,53
Garage Yr Blt	0,53
Mas Vnr Area	0,51
TotRms AbvGrd	0,50
Fireplaces	0,47
BsmtFin SF 1	0,43
Lot Frontage	0,36
Wood Deck SF	0,33
Open Porch SF	0,31
Half Bath	0,29
Bsmt Full Bath	0,28
2nd Flr SF	0,27
Lot Area	0,27
Bsmt Unf SF	0,18
Bedroom AbvGr	0,14
Enclosed Porch	0,13
Kitchen AbvGr	0,12
Screen Porch	0,11
Overall Cond	0,10
Pool Area	0,07
Low Qual Fin SF	0,04
Bsmt Half Bath	0,04
3Ssn Porch	0,03
Yr Sold	0,03
Misc Val	0,02
BsmtFin SF 2	0,01

Fonte: Resultados originais da pesquisa

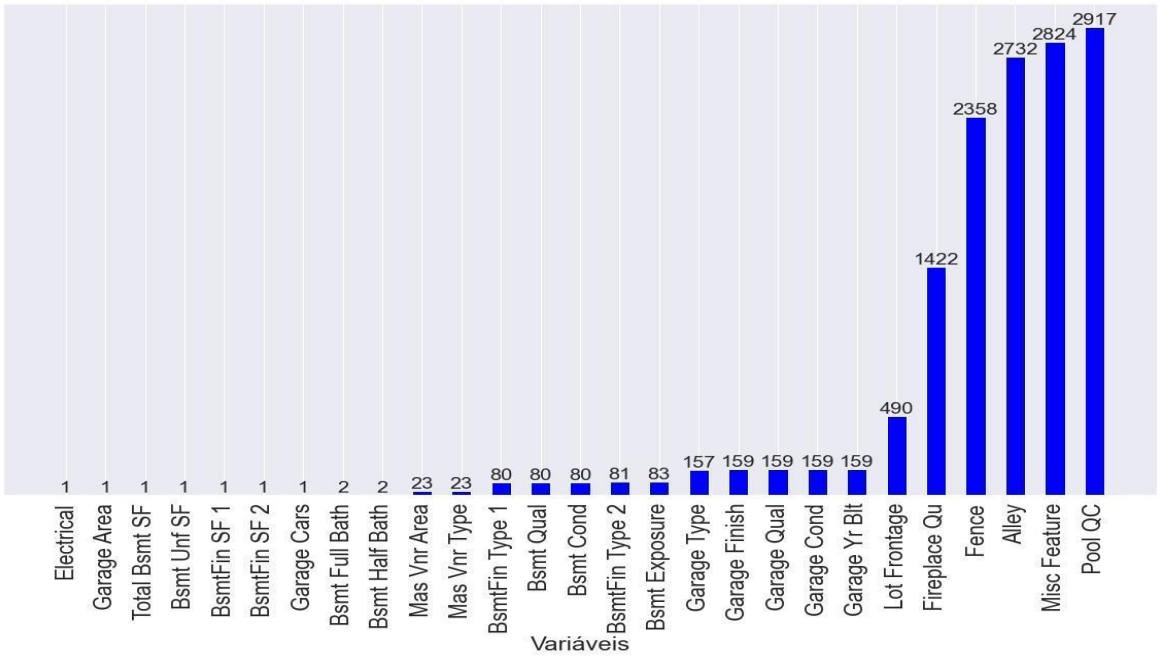

Figura 3. Dados faltantes em "Ames Housing"
Fonte: Resultados originais da pesquisa

Seguindo as orientações fornecidas por Cock [2011] em sua descrição do conjunto de dados, foram adotadas as seguintes estratégias para lidar com valores nulos, visto que estes impactam diretamente a construção dos modelos: para variáveis numéricas em que a descrição sugere que valores nulos representam a ausência da característica, os valores nulos foram substituídos por zero. Para variáveis categóricas em que os valores nulos indicam a ausência da característica, foi atribuída a categoria mais relevante para essa classificação a cada variável, substituindo o valor nulo. Nos outros casos, foi utilizada a média (para variáveis numéricas) ou a moda (para variáveis categóricas) para preencher os dados faltantes.

3. Resultados e Discussão

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da aplicação de seis modelos distintos de previsão supervisionada. A discussão a seguir integra esses resultados com as informações conhecidas do desafio "Kaggle" até a data de conclusão deste estudo. Essa comparação oferece uma visão mais aprofundada sobre o desempenho dos modelos desenvolvidos, no contexto dos avanços atuais e das contribuições de outros pesquisadores na área.

3.1 Modelagem

Seis modelos distintos foram construídos em sequência para previsão dos valores dos imóveis em Ames Housing, sendo referenciados de "Modelo 1" a "Modelo 6". Como ponto de partida, escolheu-se um modelo de menor complexidade: a regressão linear simples.

Dessa forma, o Modelo 1 serve como uma base interessante para comparação com os demais modelos. Visando melhorar os resultados desse modelo - considerando que deveria ser escolhida uma única variável preditora - foi selecionado o atributo "Overall Qual", pois apresentou forte correlação com a variável de resposta, conforme mostrado na Tabela 2. O valor de p para a estatística F obtido para o Modelo 1 foi $p < 0{,}001$, indicando que o modelo possui significância global na estimativa da variável resposta, com base na avaliação da estatística F apresentada em Wang e Cui [2013]. Os resultados de R^2 ajustado e RMSE para o primeiro modelo estão disponíveis na Tabela 3.

Tabela 3. Métricas de desempenho do Modelo 1

Métrica	Conjunto de Treino	Conjunto de Teste
R^2 ajustado	0,708	0,665
RMSE	0,057	0,055

Fonte: Resultados originais da pesquisa

A Figura 4 ilustra para dados de treino e teste (A) a dispersão dos resíduos das predições do modelo e (B) a curva de valores preditos comparados aos valores reais. A avaliação visual dessas dispersões pode ser usada comparativamente aos demais modelos, de maior complexidade, desenvolvidos na sequência. Não obstante, pode-se observar uma grande amplitude dos resíduos e grande descasamento entre a curva de referência e os resultados obtidos na comparação “Valores Preditos vs Valores Reais”, como era de se esperar para valores altos de RMSE.

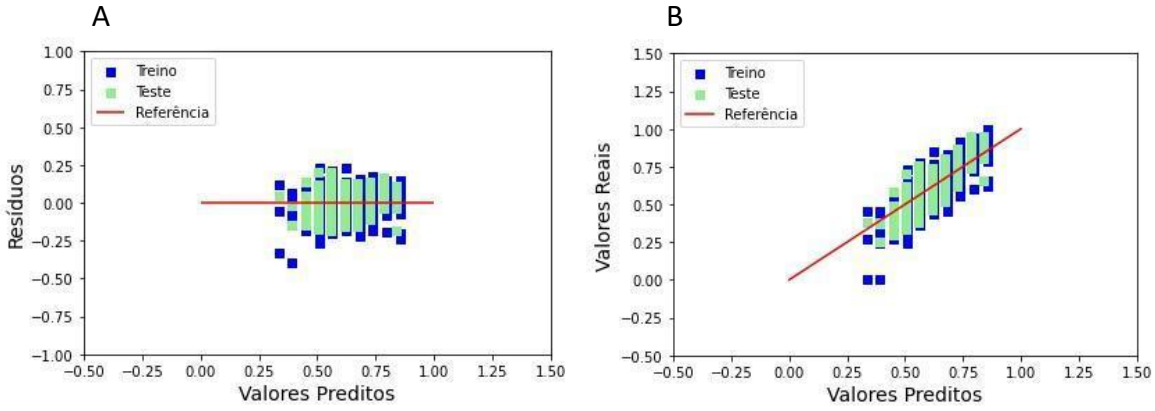

Figura 4. A) Dispersão dos resíduos do Modelo 1. B) Comparativo de valores preditos e reais para o Modelo 1
Fonte: Resultados originais da pesquisa

O Modelo 2 foi desenvolvido com o objetivo de aprimorar a acurácia obtida no Modelo 1. Para isso, aumentou-se a complexidade do algoritmo, passando da Regressão Linear Simples para a Regressão Linear Múltipla. Utilizaram-se todas as 79 variáveis

preditoras presentes no conjunto de dados, sendo que as variáveis categóricas foram codificadas utilizando a técnica "one-hot". O valor da probabilidade da estatística F obtido para o Modelo 2 foi $p < 0,001$, indicando que o modelo também apresenta significância global na estimação da variável resposta. Os resultados de R^2 ajustado e RMSE para o segundo modelo estão apresentados na Tabela 4. Observa-se uma melhoria significativa nesses parâmetros para os dados de treino, quando comparados ao Modelo 1. Contudo, essa melhora não é igualmente observada nos dados de teste, sugerindo que o Modelo 2 apresenta problemas de "overfitting".

Tabela 4. Métricas de desempenho do Modelo 2

Métrica	Conjunto de Treino	Conjunto de Teste
R^2 ajustado	0,937	0,648
RMSE	0,023	0,048

Para a construção do Modelo 3 objetivou-se a redução do "overfitting" encontrado no modelo anterior. Para tal foram propostas as seguintes modificações: variáveis categóricas ordinais foram codificadas por rótulos em detrimento à técnica "one-hot", permitindo-se assim a manutenção do sentido de ordem dessas variáveis; reduziu-se a cardinalidade dos dados aplicando-se o agrupamento às variáveis nominais com dispersão desigual de categorias – ou seja, categorias mais raras (que representem menos de 2% do conjunto de forma isolada) foram agrupadas em uma única categoria "outros"; aplicou-se transformação logarítmica às variáveis numéricas com valor de assimetria superior a 0,5 – incluindo-se a variável resposta – com intuito de minimizar o impacto de "outliers" nas observações; como metodologia de seleção de atributos usou-se o procedimento "forward stepwise", para mitigação de problemas de colinearidade; e eliminou-se manualmente 2 observações com valores residenciais superiores a 4.000 dólares americanos, uma vez que os autores do "dataset" apontam estes casos como problemas de coleta de dados (Cock, 2011).

O valor da probabilidade da estatística F obtida para o Modelo 3 foi $p < 0,001$, evidenciando que o modelo também apresenta significância global para estimação da variável resposta. Os resultados de R^2 ajustado e RMSE para o segundo modelo podem ser visualizados na Tabela 5. Apesar dos resultados no conjunto de treino serem ligeiramente inferiores aos obtidos para o Modelo 2, houve melhora significativa para os resultados no conjunto de teste, mostrando que as técnicas aplicadas para redução de colinearidade, impacto de "outliers" e cardinalidade dos dados foram bem-sucedidas.

Tabela 5. Métricas de desempenho do modelo 3

Métrica	Conjunto de Treino	Conjunto de Teste
R^2 ajustado	0,904	0,889
RMSE	0,034	0,037

Fonte: Resultados originais da pesquisa

A Figura 6 ilustra para dados de treino e teste (A) a dispersão dos resíduos das predições do modelo e (B) a curva de valores preditos comparados aos valores reais. Como observação mais evidente destas ilustrações, nota-se menores diferenças entre os resultados obtidos para os conjuntos de treino e teste – indicativo de redução de problemas de "overfitting".

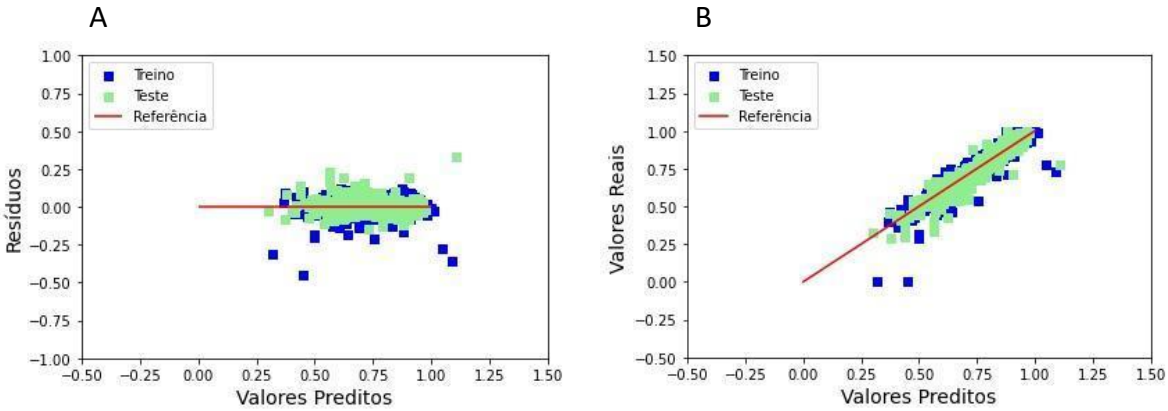

Figura 6. A) Dispersão dos resíduos do Modelo 3. B) Comparativo de valores preditos e reais para o Modelo 3
Fonte: Resultados originais da pesquisa

Os Modelos 4, 5 e 6 foram construídos considerando-se as mesmas técnicas de tratamento de dados aplicadas ao Modelo 3, porém com metodologias de regressão mais avançadas. Primeiramente, no Modelo 4, utilizou-se a modelagem "Elastic Net", principalmente como alternativa à técnica de seleção de atributos "forward stepwise" usada anteriormente, além de ser uma técnica indicada para conjunto de dados que apresentam problemas de colinearidade.

Os resultados de R^2 ajustado e RMSE para o quarto modelo podem ser visualizados na Tabela 6. Em comparação com o Modelo 3, houve ligeira redução desses parâmetros (à exceção do R^2 ajustado para o conjunto de treino, que permaneceu o mesmo). Portanto, a aplicação das metodologias de tratamento de dados aliadas ao procedimento "forward stepwise" se mostraram praticamente tão eficientes quanto a aplicação do algoritmo "Elastic Net" para o problema de regressão em Ames Housing neste estudo.

Tabela 6. Métricas de desempenho do Modelo 4

Métrica	Conjunto de Treino	Conjunto de Teste
R^2 ajustado	0,904	0,875
RMSE	0,033	0,036

Fonte: Resultados originais da pesquisa

A Figura 7 ilustra para dados de treino e teste (A) a dispersão dos resíduos das predições do modelo e (B) a curva de valores preditos comparados aos valores reais. Também uma pequena diferença entre os Modelos 3 e 4 pode ser apontada nesta etapa, tendo o Modelo 4 aprimorado ligeiramente a dispersão dos resíduos, configurando um modelo com menor heterocedasticidade, com base na análise visual dos resíduos.

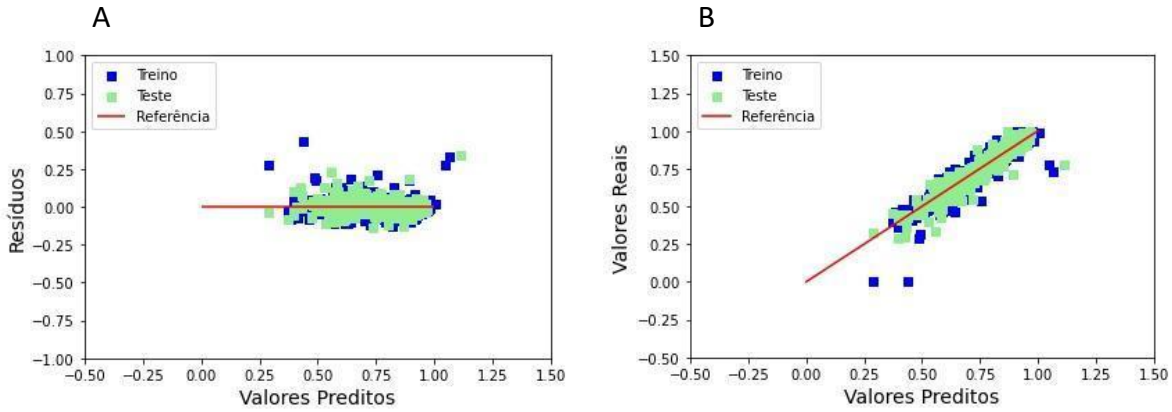

Figura 7. A) Dispersão dos resíduos do Modelo 4. B) Comparativo de valores preditos e reais para o Modelo 4
Fonte: Resultados originais da pesquisa

A técnica "Random Forest" foi aplicada para o quinto modelo desenvolvido neste trabalho. Os resultados obtidos de R^2 ajustado e RMSE – disponíveis na Tabela 7 – demonstram que este modelo mais complexo, baseado em árvores de decisão, possui uma alta capacidade de predição, atingindo valores elevados de R^2 ajustado e valores baixos de RMSE para o conjunto de treino. Em contraste ao resultado obtido para o Modelo 2, os problemas de "overfitting" para este modelo foram suavizados, com resultados para os dados de teste próximos aos obtidos para os Modelos 3 e 4.

Tabela 7. Métricas de desempenho do Modelo 5

Métrica	Conjunto de Treino	Conjunto de Teste
R^2 ajustado	0,973	0,842
RMSE	0,017	0,040

Fonte: Resultados originais da pesquisa

A Figura 8 ilustra para dados de treino e teste (A) a dispersão dos resíduos das predições do modelo e (B) a curva de valores preditos comparados aos valores reais.

Novamente nota-se uma distribuição mais adequada dos resíduos ao longo dos valores preditos, indicando redução de heterocedasticidade do modelo. O comparativo "Valores Preditos vs Valores Reais" ilustra a acurácia elevada obtida especialmente para os dados de treino quando da aplicação da técnica "Random Forest".

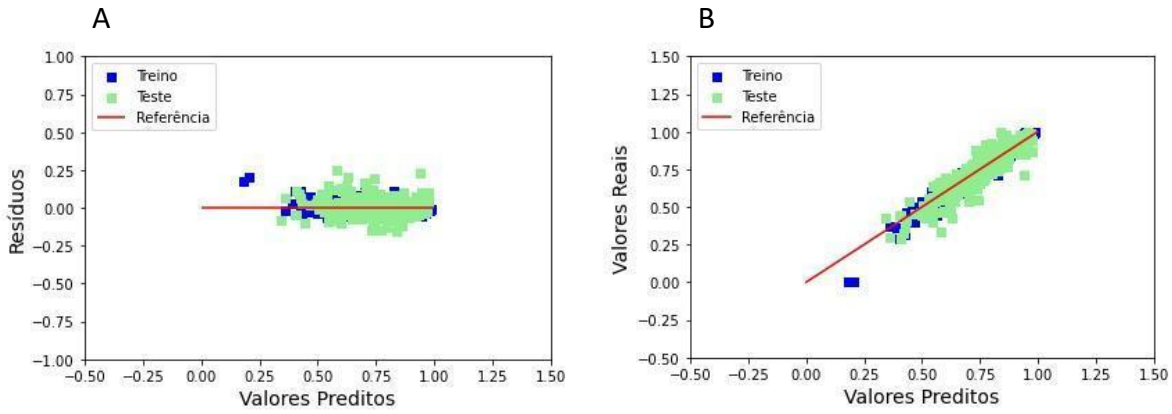

Figura 8. A) Dispersão dos resíduos do Modelo 5. B) Comparativo de valores preditos e reais para o Modelo 5
Fonte: Resultados originais da pesquisa

Por fim, o sexto modelo consistiu na aplicação do método XGBoost. Os resultados para R^2 ajustado e RMSE podem ser observados na Tabela 8. Tanto para o conjunto de treino quanto para o de teste, o Modelo 6 apresentou os melhores resultados deste estudo, demonstrando-se (dentre os demais métodos testados) como o mais adequado para lidar com a complexidade dos dados existentes no conjunto Ames Housing. Comparando-se os valores de RMSE obtidos para o conjunto de teste com os melhores resultados apresentados na literatura, obtém-se que, se ranqueado junto à competição Kaggle, o modelo alcançaria a posição 57 dentre todas as submissões já recebidas até a data de elaboração deste estudo (Kaggle, 2021).

Tabela 8. Métricas de desempenho do Modelo 6

Métrica	Conjunto de Treino	Conjunto de Teste
R^2 ajustado	0,991	0,905
RMSE	0,010	0,034

Fonte: Resultados originais da pesquisa

A Figura 9 ilustra para dados de treino e teste (A) a dispersão dos resíduos das predições do modelo e (B) a curva de valores preditos comparados aos valores reais. Percebe-se com relação aos demais modelos uma distribuição mais homogênea dos resíduos – redução de heterocedasticidade – além da evidente melhor acurácia do último modelo (tanto para dados de treino quanto para dados de teste). Em contrapartida, mesmo neste caso, quando aplicados os testes de Shapiro-Wilk (para verificação da normalidade dos resíduos)

e de Breusch-Pagan (para verificação de homoscedasticidade), os resíduos do modelo ainda se mostram estatisticamente não aderentes à normalidade e com presença de heterocedasticidade. Isto demonstra uma dificuldade de aplicação das metodologias de aprendizado de máquina a dados reais, usualmente complexos para modelagem. Os resultados obtidos destes testes sugerem que as formas funcionais utilizadas para as variáveis preditoras e variável resposta potencialmente não são as mais adequadas para aplicação no modelo, bem como podem existir variáveis não incluídas na predição que são necessárias para a adequada estimação de "Sale Price".

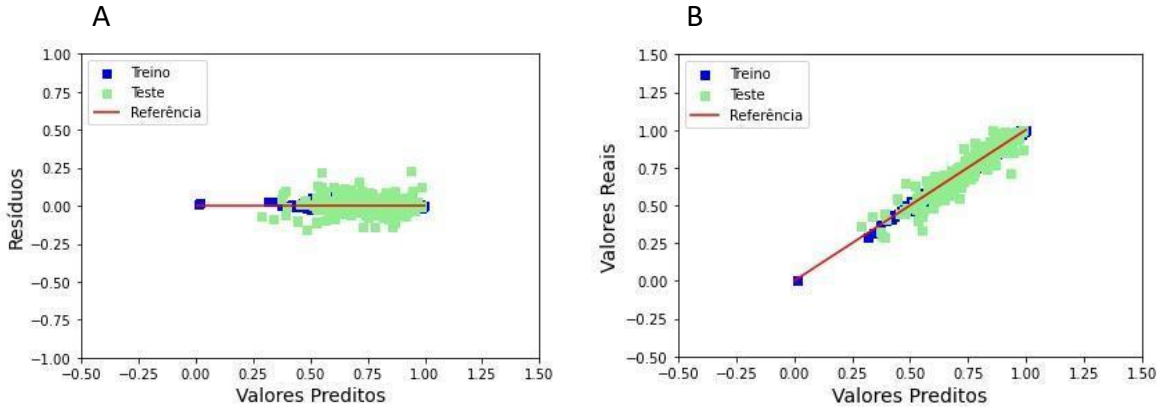

Figura 9. A) Dispersão dos resíduos do Modelo 6. B) Comparativo de valores preditos e reais para o Modelo 6
Fonte: Resultados originais da pesquisa

Os resultados obtidos são corroborados por outros trabalhos, como em Han (2023), que obteve resultados melhores na regressão de "Ames Housing" quando aplicadas técnicas "Gradient Boosting", tendo-as comparado com outros modelos preditivos como: "Elastic Net", LASSO, LightGBM e "Stacked Model". Já Xu (2023) destaca a colinearidade entre as variáveis preditoras como principal desafio na proposição de uma solução para o problema.

A tabela 9 sumariza os resultados obtidos para a métrica R^2 Ajustado em cada um dos modelos descritos.

Tabela 9. Resumo de R^2 Ajustado para os 6 modelos aplicados

Conjunto	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6
Treino	0,708	0,937	0,904	0,904	0,973	0,991
Teste	0,665	0,648	0,889	0,875	0,842	0,905

Fonte: Resultados originais da pesquisa

A tabela 10 sumariza os resultados obtidos para a métrica RMSE em cada um dos modelos descritos.

Tabela 10. Resumo de RMSE para os 6 modelos aplicados

Conjunto	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6
Treino	0,057	0,023	0,034	0,033	0,017	0,010
Teste	0,055	0,048	0,037	0,036	0,040	0,034

Fonte: Resultados originais da pesquisa

4. CONCLUSÃO

Este estudo investigou diversas técnicas de aprendizado de máquina supervisionado, aplicando-as ao conjunto de dados Ames Housing com o objetivo de prever os valores das propriedades residenciais em Ames, Iowa. As metodologias empregadas variaram desde a Regressão Linear Simples e Múltipla até técnicas mais avançadas, como "Elastic Net", "Random Forest" e "Extreme Gradient Boosting".

A abordagem sequencial adotada no estudo, desde a análise exploratória dos dados até a aplicação de diferentes técnicas de regressão, destacou a importância de avaliar cuidadosamente cada etapa do processo de modelagem. O tratamento de dados, que incluiu o tratamento de valores ausentes, a identificação de "outliers", a codificação de variáveis categóricas e a aplicação de técnicas de transformação de variáveis, como a transformação logarítmica, foi crucial para o aprimoramento dos modelos preditivos. A transição de modelos mais simples para mais complexos evidenciou as compensações entre a simplicidade do modelo e sua capacidade preditiva.

Os resultados obtidos, especialmente com o modelo XGBoost, mostraram que a modelagem conseguiu alcançar boa eficácia na estimação da variável resposta, comparando-se favoravelmente aos resultados obtidos na competição Kaggle, que utiliza o mesmo conjunto de dados.

Contudo, o estudo também ressaltou os desafios e considerações ao trabalhar com dados reais. Um desafio significativo encontrado durante a análise foi a distribuição não normal dos resíduos em todos os modelos, juntamente com testes de heterocedasticidade insatisfatórios, o que indicou possíveis problemas com o ajuste do modelo e as suposições da modelagem.

Além disso, é importante observar que algumas limitações da base de dados podem ter impactado os resultados obtidos. A abordagem adotada para lidar com os dados faltantes, com o preenchimento de valores ausentes por médias globais, pode ter introduzido viés nos modelos, o que pode ter causado vazamento de dados. A análise, que focou principalmente em relações lineares com base na correlação de Pearson, pode não ter capturado relações não lineares ou sazonais que poderiam ser detectadas por transformações de variáveis ou modelos mais avançados. Esses fatores devem ser considerados ao interpretar os resultados e podem ser investigados em pesquisas futuras para aperfeiçoar as técnicas aplicadas.

Por fim, pesquisas futuras podem buscar otimizar ainda mais esses modelos, integrando fontes adicionais de dados para aprimorar as previsões, contribuindo assim para o avanço da Ciência de Dados e suas aplicações práticas em dados reais.

Referencias

Abraham, A.; Pedregosa, F.; Eickenberg, M.; Gervais, P.; Mueller, A.; Kossaifi, J.; Gramfort, A.; Thirion, B.; Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8: 14-24.

Al-shehari, T.; Alsowail, R. A. (2021). An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, 23(10): 1258-1282.

Breiman, L. (2001). Random forests. *Machine Learning*, 45: 5-32.

Cao, S.; Zeng, Y.; Yang, S.; Cao, S. (2021). Research on Python Data Visualization Technology. *Journal of Physics: Conference Series*, 1757(1): 12122-12130.

Chen, T.; Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13: 785-794.

Cock, D. (2011). Ames, Iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19: 3-18.

Demir, S. (2022). Comparison of Normality Tests in Terms of Sample Sizes under Different Skewness and Kurtosis Coefficients. *International Journal of Assessment Tools in Education*, 9(2): 397-409.

Dufera, A. G.; Liu, T.; Xu, J. (2023). Regression models of Pearson correlation coefficient. *Statistical Theory and Related Fields*, 7(2): 97-106.

Emmert-Streib, F.; Dehmer, M. (2019). Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error. *Machine Learning and Knowledge Extraction*, 1: 521-551.

Han, Y. (2023). Price Prediction of Ames Housing Through Advanced Regression Techniques. *BCP Business & Management*, 38: 1965-1974.

Hyde, K. K.; Novack, M. N.; LaHaye, N.; Parlett-Pelleriti, C.; Anden, R.; Dixon, D. R.; Linstead, E. (2019). Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review. *Journal of Autism and Developmental Disorders*, 6(2): 128-146.

Kaggle. (2021). House Prices: Advanced Regression Techniques. Disponível em: https://www.kaggle.com/c/house-prices-advanced-regression-techniques. Acesso em: 06 mar. 2024.

Maulud, D.; Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(2): 140-147.

Mehare, H. Bin; Anilkumar, J. P.; Usmani, N. A. (2023). The Python Programming Language. A Guide to Applied Machine Learning for Biologists. 1ed. Mohammad Sufian Badar. Riverside, CA, USA.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons*, 4: 51-62.

Sahoo, K.; Samal, A. K.; Pramanik, J.; Pani, S. K. (2019). Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, 8(12): 4727-4735.

Seabold, S.; Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*, 1:92-96.

Tauchert, C.; Buxmann, P.; Lambinus, J. (2020). Crowdsourcing data science: A qualitative analysis of organizations' usage of kaggle competitions. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1: 393-398.

Uddin, S.; Khan, A.; Hossain, M. E.; Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1): 281-297.

Vaudevan, M.; Narayanan, R. S.; Nakeeb, S. F.; Abhishek. (2022). Customer churn analysis using XGBoosted decision trees. *Indonesian Journal of Electrical Engineering and Computer Science*, 25(1): 488-495.

Wang, S.; Cui, H. (2013). Generalized F test for high dimensional linear regression coefficients. *Journal of Multivariate Analysis*, 117: 134-149.

Xu, X. (2023). The real estate price prediction of US prediction based on multi-factorial linear regression models. *BCP Business & Management*, 1: 1-6.

Zou, H.; Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2): 301-320.