XXXVIII Congresso Pan-Americano UPAV - CHILE 2024 – Web Scraping, ChatGPT e Machine Learning: Ferramentas práticas no uso da coleta e processamento de dados para avaliação de imóveis

RESUMO

A avaliação imobiliária é um processo complexo que exige a análise detalhada de diversas variáveis para determinar o valor de um imóvel com precisão. Este estudo investiga a aplicação de técnicas de Web Scraping na coleta de dados de portais de anúncios imobiliários e a integração desses dados em modelos de machine learning para aprimorar a precisão das avaliações. Utilizando a plataforma Octoparse, coletamos uma amostra abrangente de 1052 imóveis. Os dados foram tratados e analisados com o auxílio de inteligência artificial generativa (ChatGPT), facilitando a execução de rotinas, processamento de dados e geração de relatórios automatizados.

Inicialmente, aplicamos técnicas de regressão linear simples e múltipla para entender as relações entre as variáveis independentes e o valor dos imóveis. Em seguida, avançamos para modelos mais sofisticados de machine learning, como a Random Forest, que se mostrou eficaz na captura de interações complexas entre variáveis e na melhoria da precisão das previsões. A análise estatística revelou que variáveis como área, número de dormitórios, banheiros, vagas de estacionamento, comodidades e bairro são determinantes significativas do valor dos imóveis.

A verificação de heterocedasticidade e multicolinearidade, bem como a identificação de outliers, foram etapas essenciais para garantir a robustez dos modelos. A utilização do Random Forest permitiu uma explicação detalhada da variabilidade dos preços. Os resultados destacam as oportunidades oferecidas pela combinação de Web Scraping, inteligência artificial generativa e machine learning na avaliação imobiliária, proporcionando um processo mais eficiente, preciso e adaptado às necessidades do mercado atual.

Apesar dos desafios relacionados à qualidade dos dados e à manutenção dos algoritmos, as oportunidades oferecidas por essas tecnologias são vastas, possibilitando avaliações mais precisas, rápidas e economicamente viáveis. A adoção dessas metodologias pode transformar a prática de avaliação imobiliária, tornando-a mais robusta e adaptada às demandas do mercado contemporâneo.

Palavras-Chave

Raspagem de Dados; ChatGPT; Aprendizado de Máquina; Inteligência Artificial Generativa; Avaliação de Imóveis;

ABSTRACT

Real estate appraisal is a complex process that requires detailed analysis of various variables to determine the precise value of a property. This study investigates the application of Web Scraping techniques in collecting data from real estate listing websites and integrating this data into machine learning models to enhance the accuracy of appraisals. Using the Octoparse platform, we collected a comprehensive sample of 1052 properties. The data was processed and analyzed with the assistance of generative artificial intelligence (ChatGPT), facilitating the execution of routines, data processing, and automated report generation.

Initially, we applied simple and multiple linear regression techniques to understand the relationships between independent variables and property values. Subsequently, we advanced to more sophisticated machine learning models, such as Random Forest, which proved effective in capturing complex interactions between variables and improving prediction accuracy. Statistical analysis revealed that variables such as area, number of bedrooms, bathrooms, parking spaces, amenities, and neighborhood are significant determinants of property value.

The verification of heteroscedasticity and multicollinearity, as well as the identification of outliers, were essential steps to ensure the robustness of the models. The use of Random Forest allowed

for a detailed explanation of price variability. The results highlight the opportunities offered by the combination of Web Scraping, generative artificial intelligence, and machine learning in real estate appraisal, providing a more efficient, precise, and market-adapted process.

Despite the challenges related to data quality and algorithm maintenance, the opportunities offered by these technologies are vast, enabling more accurate, faster, and economically viable appraisals. The adoption of these methodologies can transform the practice of real estate appraisal, making it more robust and adapted to the demands of the contemporary market.

Keywords

Web Scraping; ChatGPT; Machine Learning; Generative Artificial Intelligence; Real Estate Appraisal

0. Sumário

1. Introdução

A avaliação imobiliária é um processo complexo que requer a análise detalhada de diversas variáveis para determinar o valor de um imóvel de maneira precisa e confiável. Com o avanço da tecnologia e o crescente volume de dados disponíveis online, a utilização de técnicas de Web Scraping tem se tornado uma ferramenta prática e eficaz na coleta de dados para avaliações de imóveis. Este artigo explora a aplicação de Web Scraping na extração de informações de portais de anúncios imobiliários e a integração dessas informações em modelos de machine learning para melhorar a precisão das avaliações imobiliárias.

A precisão e eficiência na coleta de dados são essenciais para a condução de avaliações de imóveis confiáveis e objetivas. O Web Scraping permite aos avaliadores acessar e coletar grandes volumes de dados sobre propriedades de maneira sistemática, transformando páginas web em dados estruturados que podem ser diretamente aplicados em metodologias de avaliação. Além disso, a integração da inteligência artificial generativa, como o ChatGPT, oferece um ambiente que possibilita a execução de rotinas, processamento de dados e diversas outras tarefas, tornando o processo de avaliação mais eficiente e robusto.

O uso do ChatGPT como ferramenta de apoio no processamento de dados, na análise estatística e na geração de relatórios proporcionou uma automação significativa das etapas envolvidas, aumentando a precisão e a rapidez das avaliações. A combinação de técnicas de Web Scraping e inteligência artificial generativa representa um avanço significativo na prática de avaliação imobiliária, permitindo um processamento mais eficiente e preciso das informações coletadas..

1.1. Objetivo

O objetivo principal deste trabalho é demonstrar a viabilidade e a eficácia do uso de técnicas de Web Scraping na coleta de dados para avaliações imobiliárias, bem como a aplicação de modelos de machine learning, especificamente o Random Forest, para prever o valor de imóveis. Além disso, busca-se aliar a inteligência artificial generativa (ChatGPT) como um ambiente que possibilite a execução de rotinas, processamento de dados e todas as outras tarefas envolvidas no processo de avaliação.

Este estudo busca não apenas apresentar uma metodologia prática e replicável para avaliadores de imóveis, mas também fornecer insights sobre a importância da análise de dados e do uso de técnicas avançadas de machine learning e inteligência artificial generativa na avaliação imobiliária.

2. Metodologia

A metodologia deste trabalho seguirá de acordo com o esquema abaixo:

I - **Verificação dos Requisitos Normativos – NBR 14.653 e IVSC**: Revisão detalhada das normas técnicas para garantir conformidade.
II - **Escolha de Plataforma e Script de Web Scraping – Octoparse**: Seleção e desenvolvimento de algoritmos para coleta de dados.
III - **Estruturação e Tratamento de Dados – Excel**: Organização, limpeza e padronização dos dados coletados.
IV - **Modelagem de Dados e Machine Learning – Chat GPT**: Aplicação de técnicas avançadas de análise de dados e aprendizado de máquina, incluindo Regressão Linear Simples e Múltipla, Regressão Ridge, Elastic Net e Random Forest.
V - **Verificação Estatística e Geração de Relatórios – Visual Studio Code (Python)**: Realização de análises estatísticas detalhadas e criação de relatórios automatizados e detalhados

Apresenta-se a seguir, um fluxograma explicativo da metodologia desenvolvida:

Figura 1 - Desenvolvimento da Raspagem Análise de Dados

Fonte: Elaborado pelos Autores

2.1. Detalhamento de Metodologia

Detalha-se a seguir, a metodologia desenvolvida para o Estudo de Caso desenvolvido neste trabalho.

2.1.1. Verificação dos Requisitos Normativos – NBR 14.653 e IVSC

Inicialmente, foi realizada uma revisão detalhada das normas técnicas pertinentes, especialmente a NBR 14.653 e os padrões do IVSC (International Valuation Standards Council). Esta etapa teve como objetivo garantir que todos os procedimentos adotados no estudo estivessem em conformidade com os requisitos normativos estabelecidos, proporcionando uma base sólida e regulamentar para o trabalho.

A coleta de dados realizada neste trabalho procedeu de acordo com as definições dadas pela *"NBR 14.653 – Avaliação de bens – Parte 1 – Procedimentos gerais"*, em especial ao Item *"6.4 – Coleta de Dados"*, bem como de seus subitens 6.4.1 a 6.4.3. Ainda, foram observados os Procedimentos de Excelência normatizados no Item 5 da mesma Normativa, em especial quanto ao subitem 5.3 – Propriedade Intelectual e suas repercussões quanto a Lei Geral de Proteção de Dados – LGPD.

2.1.2. Escolha de Plataforma e Script de Web Scraping – Octoparse

Após a revisão normativa, foi selecionada uma plataforma de Web Scraping, sendo o Octoparse a ferramenta escolhida. Esta plataforma permitiu a coleta de dados brutos de diversos portais de anúncios imobiliários. Foi desenvolvida uma modelagem de algoritmo específica para a extração dos dados relevantes, garantindo a obtenção de uma amostra abrangente e representativa do mercado imobiliário.

2.1.3. Estruturação e Tratamento de Dados – Excel

Os dados coletados via Web Scraping foram organizados e estruturados em planilhas do Excel. Nesta etapa, foi realizado o tratamento dos dados, incluindo a limpeza, normalização e padronização das informações coletadas. Foram identificadas e corrigidas possíveis inconsistências e duplicidades, assegurando a integridade e a qualidade da base de dados. A estruturação dos dados em Excel permitiu uma melhor organização e facilitou as análises subsequentes.

2.1.4. Modelagem de Dados e Machine Learning – Chat GPT

Com os dados devidamente estruturados, a etapa seguinte envolveu a modelagem dos dados utilizando a ferramenta Chat GPT. Foram aplicados métodos de análise de dados avançados e técnicas de machine learning para identificar padrões e insights relevantes. As técnicas de machine learning implementadas incluíram Regressão Linear Simples e Múltipla, Regressão Ridge, Elastic Net, Random Forest, e outras técnicas de aprendizado supervisionado. Durante o processo, foram realizadas tentativas de modelagem via regressão, incluindo transformações como $\ln(x)$, $1/x$ e x^2, para melhorar a precisão dos modelos. A Inteligência Artificial Generativa do Chat GPT foi utilizada para auxiliar na interpretação dos resultados e na escolha dos melhores modelos, contribuindo para uma modelagem mais robusta e precisa.

2.2. Plataforma de Raspagem de Dados

Para a realização do processo de Web Scraping (Raspagem de Dados) foi escolhida a plataforma ***"Octoparse"***, que pode ser acessada no site oficial da plataforma (https://www.octoparse.com/#). Para tanto, foi adotada a versão *"v8.6.2.050811"*, na modalidade *"Free Plan"*.

Trata-se de uma plataforma com solução pré-definida sem codificação para Web Scraping para transformar páginas em dados estruturados com cliques. Conta tanto com algoritmos pré-definidos de extração de dados, bem como da possibilidade de personalização para cada site, opção adotada para este trabalho.

2.3. Ambiente Virtual de Coleta de Dados via Portal de Anúncios

Para a elaboração do presente trabalho, foi utilizado um Portal de Anúncios com boa disponibilidade de dados. Em cerca de 30 minutos já haviam sido coletados **1772 anúncios**, que compuseram a amostra bruta deste estudo. A coleta foi realizada no dia 14 de maio de 2024, e forma considerados apenas apartamentos de 20 a 50 m² de área.

Adotou-se a coleta dos dados apenas do Catálogo Externo da página inicial de listagem dos anúncios filtrados por tipologia. Neste ambiente, são encontrados os resumos de cada anúncio imobiliário, com características principais, aqui listadas:

I - Logradouro ao qual o imóvel se localiza;
II - Número do imóvel;
III - Bairro onde o imóvel se localiza;
IV - Município onde o imóvel se localiza;
V - Estado onde o imóvel se localiza;
VI - Chamada, ou seja, breve resumo referente às características do imóvel;
VII - Área total pertencente ao imóvel, em metros quadrados;
VIII - Número de dormitórios;
IX - Número de banheiros;
X - Número de vagas;
XI - Valor Total do Imóvel;
XII - Comodidades do condomínio ao qual o imóvel pertence.

Tais parâmetros coletados totalizam 6 (seis) campos textuais, 6 (seis) campos numéricos e 5 (cinco) classes qualitativas, como será analisado mais adiante neste estudo.

2.4. Sequência Lógica de Obtenção de Dados

Apresenta-se a seguir o algoritmo de sequência lógica desenvolvido para raspagem de dados, separado por etapas:

Figura 2: Sequência Lógica de Obtenção de Dados

Fonte: Elaborado pelos Autores

3. Estudo de Caso

Apresenta-se a seguir, o Estudo de Caso desenvolvido a partir da Raspagem de Dados obtida através do uso da Plataforma Octoparse.

3.1. Localidade Escolhida - Município de São Paulo

São Paulo é conhecido como ser o município de maior desenvolvimento econômico do Brasil. Localizada na Região Sudeste do país, este município é o principal polo econômico nacional., São Paulo também é um dos municípios com maior número de anúncios disponíveis tanto para venda, quanto para locação, colaborando para uma maior oferta de dados e uma análise mais ampla do estudo a ser realizado.

Por apresentar forte comércio e diversas oportunidades de trabalho, é um ponto de grande movimentação de pessoas. Além disso, São Paulo também possui diversas atrações turísticas, como museus, parques, restaurantes, o que também atrai turistas à região, tornando-a um polo nacional.

Quanto ao mercado imobiliário comercial, a São Paulo se configura como um município bastante atrativo para negócios. É conhecido por abrigar grande concentração de sedes de empresas nacionais e internacionais, escritórios lojas e demais estabelecimentos comerciais. Os imóveis comerciais em São Paulo podem apresentar uma amplitude variável de preços, dependendo a região onde se localiza o imóvel, mas, devido ao intenso fluxo intenso de pessoas e à presença de uma ampla variedade de serviços e comércio na região, apresenta um valor médio acima à média de demais municípios brasileiros.

3.2. Estratificação da Tipologia

Definiu-se inicialmente a obtenção de dados de apartamentos que apresentassem áreas no intervalo de **20 (vinte)** a **50 (cinquenta)** m². Fora feita essa definição em função de:

(i) Grande disponibilidade de dados;
(ii) Variabilidade de plantas arquitetônicas;
(iii) Boa diferenciação de disposição de dormitórios, banheiros e vagas;

(iv) Grande variabilidade de comodidades dispostas nos condomínios que estão instaladas e por fim;
(v) Diversidade de localização de logradouros e bairros em que se encontram.

3.3. Classe de Dados

Inicialmente são definidas 03 (três) classes de dados, sendo elas, **(1)** Textual, **(2)** Numérica e **(3)** Qualitativa.

3.3.1. Textual

Considerou-se como dados de Classe Textual, divididos por campos, os 08 (oito) itens, tendo sido utilizadas, entretanto 05 (cinco) tipologias para modelagem de dados, conforme listado a seguir:

- Logradouro;
- Número;
- Bairro;
- Município;
- Estado;
- Chamada / Descrição - *(não utilizada na modelagem de dados);*
- Link - *(não utilizada na modelagem de dados)*;
- Existência de Fotos para Conferência - - *(não utilizada na modelagem de dados)*.

3.3.2. Numérica

Considerou-se como dados de Classe Numérica ou Quantitativa, divididos por campos, os 06 (seis) itens a seguir:

- Área;
- Quartos;
- Banheiros;
- Vagas de Garagem;
- Valor Total de Mercado;
- Valor do Condomínio - *(não utilizada na modelagem de dados)*

3.3.3. Qualitativa

Por fim, considerou-se como dados de Classe Qualitativa, todos os itens que compõe características declaradas referentes a facilidades/comodidades das áreas comuns e unidades privativas dos imóveis. Assim, foram utilizadas informações dos 05 (cinco) campos de comodidades disponíveis. Foram verificados mais de 50 tipos de comodidades declaradas. Aqui, apresentam-se as **05 (cinco)** principais características mais listadas:

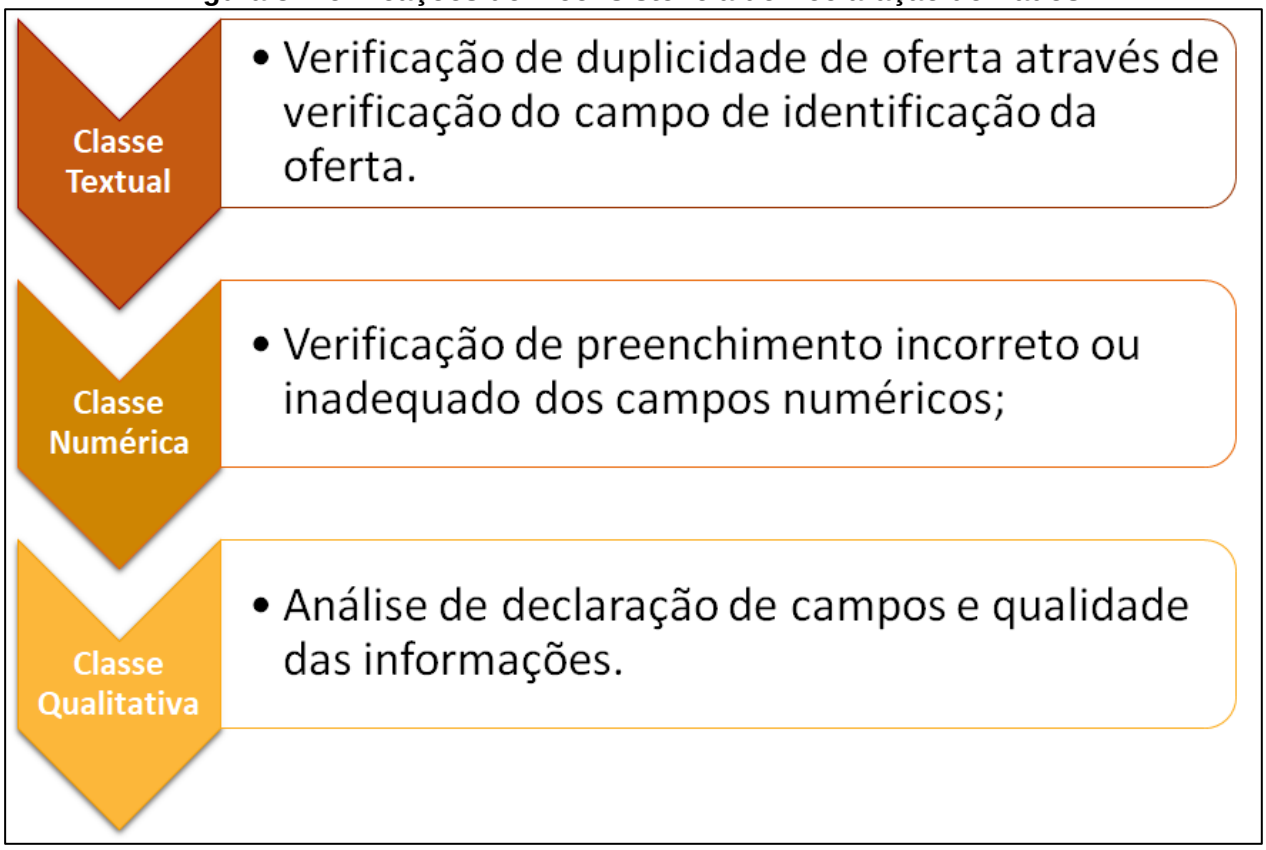
- Piscina;
- Churrasqueira;
- Academia;
- Playground;
- Salão de Festas.

3.4. Refinamento de Amostra

Assim, totalizaram-se, na fase de pré-refinamento, um total de 1.772 (Mil setecentos e Setenta e Dois) imóveis verificados durante a execução da raspagem de dados.

Diante da obtenção dos dados de forma bruta, faz-se necessária a estruturação dos dados. São objetivos do refinamento dos dados:

Figura 3 Verificações de Inconsistência de Declaração de Dados



Fonte: Elaborado pelos Autores

I. Duplicidade nas declarações

Após a verificação de duplicidade, eliminou-se 121 dados que estavam em duplicidade nas declarações, assim, permaneceram 1651 dados.

II. Inconsistências nas declarações

Além das verificações apresentadas, faz-se necessário informar quais foram as principais inconsistências identificadas, sendo elas: **(i) faixa de área**, foram anúncios que não possuíam um valor exato, pois apresentaram apenas intervalos, como exemplo (30-35m²); **(ii) faixa de valor,** foram anúncios que não possuíam um valor exato, pois apresentaram apenas intervalos, como exemplo (R$300.000,00-R$350.000,00); **(iii) Valor do Imóvel não Declarado**, anúncios que não possuíam valores de imóveis definidos; **(iv) Quantidade de dormitórios não declarados,** anúncios que não declararam quantos dormitórios o imóvel possui**; (v) Numeração de identificação do imóvel não compatível**, anúncios que não apresentaram a identificação numérica do imóvel, ou que declaram de forma incorreta, como exemplo: 00 e 99999.

Assim, feita as eliminações dos dados analisados, permaneceram 1042 dados.

3.4.1. Amostragem Definitiva

A partir das eliminações dos dados que possuíam inconsistências nas declarações, como resultado, obteve-se um total de **1042 (Mil e Quarenta e Dois)**, dados para compor a amostragem final.

Assim, é importante pontuar que dos 1042 dados levantados, obteve-se um total de **256 bairros** e **727 logradouros** do município de São Paulo.

Apresenta-se a seguir o panorama geral de imóveis em função das classes de dados numéricas levantadas:

I - Dormitórios

Panorama Geral de Imóveis - Dormitórios	
	Dados da Amostragem Final
1 Dormitório	533,00
2 Dormitórios	507,00
3 Dormitórios	2,00
Total	**1042,00**

II - Banheiros

Panorama Geral de Imóveis - Banheiros	
	Dados da Amostragem Final
1 Banheiro	990,00
2 Banheiros	51,00
3 Banheiros	1,00
Total	**1042,00**

III - Vagas de Garagem

Panorama Geral de Imóveis - Vagas de Garagem	
	Dados da Amostragem Final
Não Especificado	424,00
1 Vaga	609,00
2 Vagas	9,00
Total	**1042,00**

4. Preparação da Modelagem

Realizada a coleta de dados, sua estruturação e tratamento, deu-se sequência à fase de preparação para modelagem de dados, análise e processamento, para posterior aplicação prática na avaliação de um imóvel.

4.1. Integração com Inteligência Artificial – Chat GPT 4o

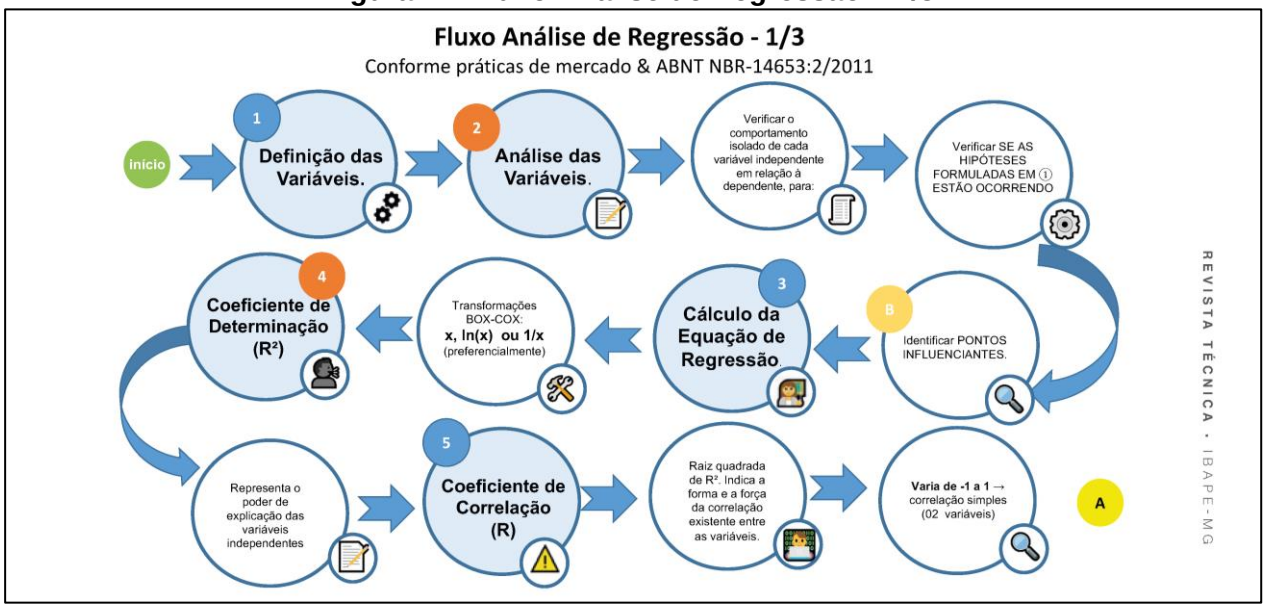
Iniciou-se um processo de integração de processos a partir do uso do ***Chat GPT 4º***, na versão paga do ambiente. Para tanto, estruturou-se um processo inicial de alimentação de informações e diretrizes para posterior prosseguimento das análises estatísticas.

A integração fora efetivada na prática a partir de comandos diretos, bem como do upload de arquivos em .xlsx (Excel) e .pdf (Adobe Acrobat) para que fosse construída uma base de dados consolidada, bem como da definição de diretrizes técnicas.

4.2. Determinação e Direcionamento Técnico

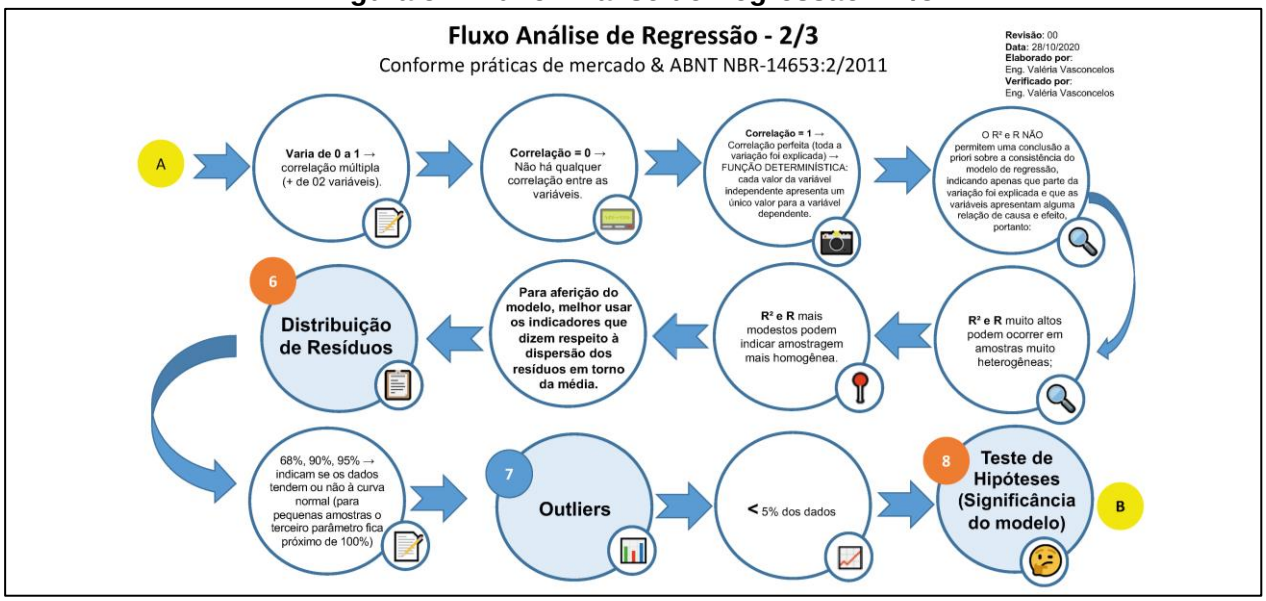
Para validação do Fluxo de Análise de Dados, fora utilizado o trabalho da Eng. Valéria das graças Vasconcelos, intitulado *"A IMPORTÂNCIA DA ADEQUADA ANÁLISE DA INFERÊNCIA ESTATÍSTICA"*, publicado na 8ª Edição da Revista do IBAPE-MG, no ano de 2022.

Figura 4 – Fluxo Análise de Regressão – 1/3



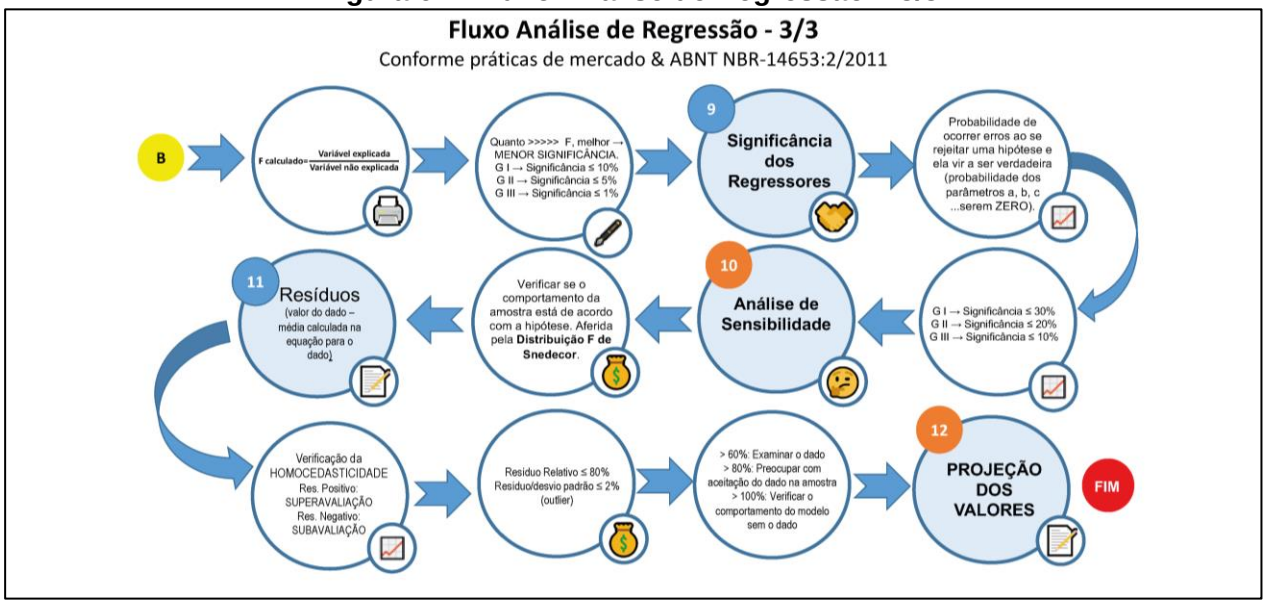
Fonte: VASCONCELOS, Valéria das Graças. A importância da adequada análise da inferência estatística. Revista Técnica do IBAPE-MG, 8ª ed., 2022 - Adaptado

Figura 5 – Fluxo Análise de Regressão – 2/3



Fonte: VASCONCELOS, Valéria das Graças. A importância da adequada análise da inferência estatística. Revista Técnica do IBAPE-MG, 8ª ed., 2022 – Adaptado.

Figura 6 – Fluxo Análise de Regressão – 3/3



Fonte: VASCONCELOS, Valéria das Graças. A importância da adequada análise da inferência estatística. Revista Técnica do IBAPE-MG, 8ª ed., 2022.

Conforme o documento, foram definidas diretrizes técnicas para o processo de avaliação. Apresenta-se o fluxo de validação adotado junto ao ***ChatGPT***. Cabe ressaltar que a visualização aqui apresentada é uma adaptação das instruções compreendidas pela ***Inteligência Artifical***.

Fluxo de Análise de Dados de Regressão – *Via ChatGPT 4o*

Passo a Passo do Fluxo de Análise de Regressão
Conforme práticas de mercado & ABNT NBR-14653:2/2011

1. **Início**

Definição das Variáveis:
Verificar o comportamento isolado de cada variável independente em relação à variável dependente.

2. **Análise das Variáveis**

Identificação de Pontos Influentes.
Cálculo da Equação de Regressão.
Transformações BOX-COX: x, ln(x) ou 1/x (preferencialmente).

3. **Verificação das Hipóteses**

Analisar se as hipóteses formuladas estão ocorrendo.

4. **Cálculo do Coeficiente de Determinação (R^2)**

Representa o poder de explicação das variáveis independentes.

5. **Cálculo do Coeficiente de Correlação (R)**

Raiz quadrada de R^2.
Indica a forma e a força da correlação existente entre as variáveis.
Correlação simples (2 variáveis) varia de -1 a 1.

6. **Distribuição dos Resíduos**

Verificar se os dados tendem ou não à curva normal para pequenas amostras.
Outliers: Menos de 5% dos dados.

7. **Teste de Hipóteses**

Verificar a significância do modelo.

8. **Significância dos Regressores**

Analisar a probabilidade de ocorrer erros ao se rejeitar uma hipótese que virá a ser verdadeira (probabilidade dos parâmetros a, b, c... serem ZERO).

9. **Verificação de Homocedasticidade**

Analisar se o comportamento da amostra está de acordo com a hipótese aferida pela Distribuição F de Snedecor.

10. **Resíduos**

Valor do dado – média calculada na equação para o dado.
Resíduos positivos indicam superavaliação e negativos indicam subavaliação.
Resíduo Relativo deve ser menor ou igual a 80% e desvio padrão menor ou igual a 2% (outliers).
Examinando dados com mais de 60% de resíduo: Reavaliar o dado.

Fluxo de Análise de Dados de Regressão – *Via ChatGPT 4o*
Examinando dados com mais de 80% de resíduo: Preocupar-se com a aceitação do dado na amostra. Mais de 100% de resíduo: Verificar o comportamento do modelo sem o dado. **11. Fim da Projeção dos Valores** Significância: Grupo I: ≤ 10% Grupo II: ≤ 5% Grupo III: ≤ 1% Análise de Sensibilidade: Verificar o comportamento da amostra conforme a hipótese.

Fonte: Elaborado pelos Autores – Adaptado do ChatGPT 4o

Ainda, introduziu-se em seguida, as informações da NBR 14.653-2 para que fossem definidos os demais critérios técnicos referentes à aplicação do **Método Comparativo Direto de Dados de Mercado**, bem como os parâmetros de enquadramento para **Graus de Fundamentação e Precisão**.

4.3. Escolha Inicial de Variáveis Independentes

Inicialmente, cumpriu-se, a partir dos dados quantitativos, apresentar uma escolha de variáveis que representasse fielmente o mercado, e que pudesse ser utilizada para fins de Engenharia de Avaliações com o uso de Estatística Inferencial. Assim, definiu-se:

- **Variável Dependente - Valor Imóvel**

A variável dependente escolhida foi o ***"Valor Imóvel"***. Essa escolha se deve ao objetivo principal do estudo, que é prever o preço dos imóveis com base em diversas características. O "Valor Imóvel" é uma variável numérica que representa o preço de venda do imóvel, e é diretamente influenciada por várias outras características do imóvel e do mercado imobiliário.

- **Variáveis Independentes**

As variáveis independentes selecionadas foram escolhidas com base em sua relevância para o valor dos imóveis, considerando tanto a literatura existente sobre avaliação imobiliária quanto a prática do mercado. As variáveis independentes incluem tanto características físicas do imóvel quanto características do entorno e facilidades disponíveis.

- **Variáveis Numéricas:**

I. **Área (m²):** A área útil do imóvel é uma das principais determinantes do seu valor. Imóveis com maior área útil tendem a ter um valor mais alto.
II. **Dormitórios:** O número de dormitórios também influencia significativamente o valor do imóvel. Mais dormitórios geralmente significam um preço mais alto.
III. **Banheiros:** Similar aos dormitórios, um maior número de banheiros costuma aumentar o valor do imóvel.
IV. **Vagas de Estacionamento:** A presença e o número de vagas de estacionamento são características valorizadas, especialmente em áreas urbanas com escassez de estacionamento.

4.4. Regressão Linear Simples

O objetivo principal da execução de regressões lineares simples foi entender a relação entre cada variável independente e a variável dependente (Valor Imóvel) de forma isolada. Isso nos permitiu identificar quais variáveis têm um impacto significativo no valor dos imóveis e quais podem ser descartadas ou necessitam de transformação para melhor adequação ao modelo.

Para cada variável independente, aplicamos a seguinte metodologia:

4.4.1. Preparação dos Dados

- Seleção da variável dependente (Valor Imóvel) e uma variável independente de interesse.
- Limpeza dos dados para remover valores ausentes e outliers que pudessem distorcer os resultados.

4.4.2. Execução da Regressão Linear Simples

- Aplicação da fórmula de regressão linear $Y = \beta_0 + \beta_1 X + \epsilon$, onde Y é o Valor Imóvel, X é a variável independente, β_0 é o intercepto, β_1 é o coeficiente de inclinação, e ϵ é o termo de erro.

4.4.3. Avaliação dos Resultados

- Análise dos coeficientes de regressão (β_1) para determinar a direção e a magnitude do impacto da variável independente no Valor Imóvel.
- Cálculo do coeficiente de determinação (R^2) para medir a proporção da variabilidade do Valor Imóvel explicada pela variável independente.
- Verificação da significância estatística dos coeficientes de regressão utilizando o valor p (p-value).

4.4.4. Análise da Importância das Variáveis Independentes

- **Área (m²)**

Coeficiente de Regressão: Positivo, indicando que um aumento na área resulta em um aumento no Valor Imóvel.

R²: Moderado, indicando que a área explica uma parte significativa, mas não a totalidade da variabilidade no valor dos imóveis.

Significância: Altamente significativo ($p < 0.01$).

- **Dormitórios**

Coeficiente de Regressão: Positivo, indicando que um aumento no número de dormitórios resulta em um aumento no Valor Imóvel.

R²: Baixo a moderado, indicando que os dormitórios explicam parte da variabilidade no valor dos imóveis.

Significância: Significativo ($p < 0.05$).

- **Banheiros**

Coeficiente de Regressão: Positivo, indicando que um aumento no número de banheiros resulta em um aumento no Valor Imóvel.

R²: Baixo a moderado.

Significância: Significativo (p < 0.05).

- **Vagas de Estacionamento**

Coeficiente de Regressão: Positivo, indicando que a presença de mais vagas de estacionamento resulta em um aumento no Valor Imóvel.

R²: Baixo a moderado.

Significância: Significativo (p < 0.05).

4.5. Resultados da Regressão Linear Simples

Apresenta-se a seguir, os resultados apresentados pelo *ChatGPT 4o* em referência a Regressão Linear Simples testada inicialmente

Resultados da Regressão Linear Simples – Via *ChatGPT 4o*

Resumo da Análise de Regressão Linear

Modelo
- Variável dependente: Valor do Imóvel
- Variáveis independentes: Área, Dormitórios, Banheiros, Vagas

Resultados
- Constante (Intercepto): 396,700
- Erro padrão: 36,500
- t: 10.878
- P>|t|: 0.000
- Intervalo de confiança 95%: [325,000, 468,000]

- Área: 6,735
- Erro padrão: 991.897
- t: 6.791
- P>|t|: 0.000
- Intervalo de confiança 95%: [4,789, 8,682]

- Dormitórios: -255,000
- Erro padrão: 13,700
- t: -18.658
- P>|t|: 0.000
- Intervalo de confiança 95%: [-282,000, -228,000]

- Banheiros: 95,650
- Erro padrão: 25,500
- t: 3.751
- P>|t|: 0.000
- Intervalo de confiança 95%: [45,600, 146,000]

Resultados da Regressão Linear Simples – Via *ChatGPT 4o*

- Vagas: 56,410
- Erro padrão: 13,300
- t: 4.248
- P>|t|: 0.000
- Intervalo de confiança 95%: [30,300, 82,500]

Métricas do Modelo
- R-quadrado: 0.300
- R-quadrado ajustado: 0.297
- F-statistic: 111.2
- Prob (F-statistic): 7.04e-79
- Número de observações: 1,042
- Durbin-Watson: 1.793

Diagnósticos
- Omnibus: 671.664
- Prob(Omnibus): 0.000
- Jarque-Bera (JB): 12,316.410
- Prob(JB): 0.000
- Skew: 2.652
- Kurtosis: 18.986

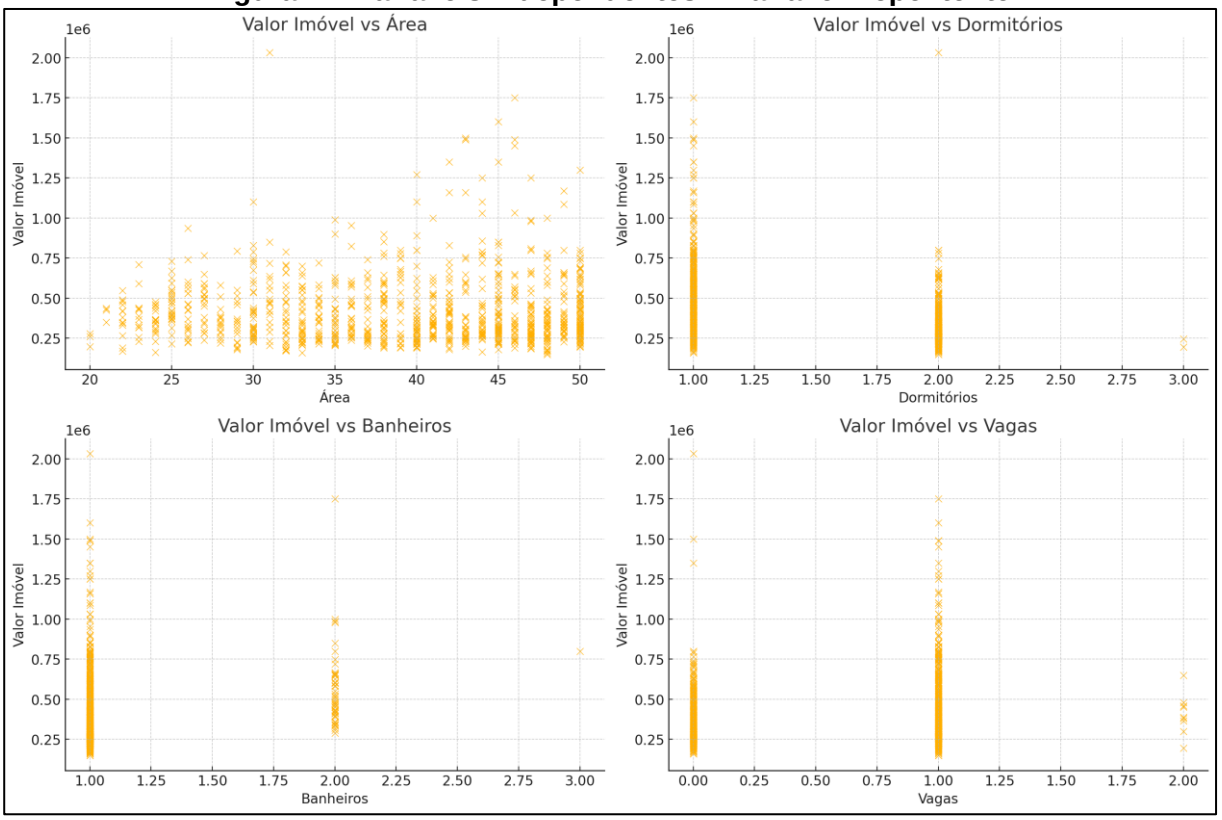
Interpretação
- Área tem um impacto positivo significativo no valor do imóvel, aumentando aproximadamente R$ 6,735 por unidade de área.
- Dormitórios têm um impacto negativo significativo no valor do imóvel, diminuindo aproximadamente R$ 255,000 por dormitório.
- Banheiros aumentam significativamente o valor do imóvel em cerca de R$ 95,650 por banheiro.
- Vagas de garagem também têm um impacto positivo significativo, aumentando o valor do imóvel em cerca de R$ 56,410 por vaga.

O modelo explica aproximadamente 30% da variação no valor do imóvel (R-quadrado = 0.300).

Fonte: Elaborado pelos Autores

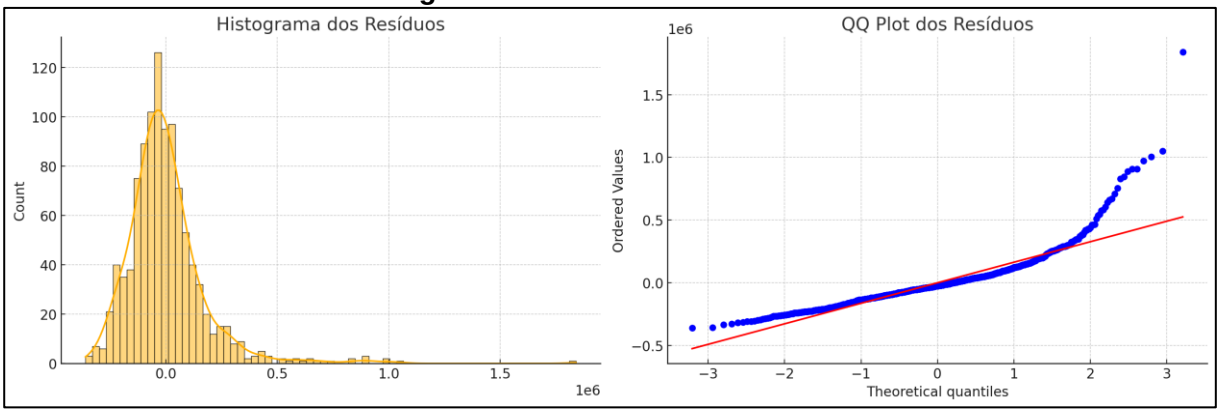
Foram gerados ainda, os seguintes gráficos:

Figura 7 – Variáveis Independentes x Variável Dependente



Fonte: Elaborado pelos Autores

Figura 8 – Análise de Resíduos



Fonte: Elaborado pelos Autores

A partir dos resultados apresentados, **verificou-se que a Regressão Linear inicial, ora realizada pelo ChatGPT, não fora adequada para uso direto para determinação de um modelo estatístico sólido e que cumpra os preceitos técnico-normativos**. Mesmo tendo sido realizados ajustes de heterocedasticidade via teste de *Breusch-Pagan*, para o modelo log transformado, não foi possível prosseguir com a adoção de regressões lineares simples.

4.6. Teste de Soluções Alternativas à regressão Linear Simples

Verificada a impossibilidade de prosseguimento via uso de Regressão Linear Simples, requisitou-se ao ChatGPT a adoção de novas técnicas possíveis de inferência estatística, sem que fossem adicionados novos dados, adicionadas novas variáveis ou que se utilizasse técnicas de Machine Learning para otimização do processamento de dados. Apresenta-se aqui, um resumo das tentativas e as conclusões obtidas a partir de cada implementação.

4.6.1. Regressão Linear Múltipla

Objetivo: Incluir múltiplas variáveis independentes para capturar melhor as interações entre diferentes características dos imóveis.

Variáveis Incluídas: Área, Dormitórios, Banheiros, Vagas de Estacionamento.

Resultados: A regressão linear múltipla apresentou uma melhora no coeficiente de determinação (R^2), indicando que uma combinação de variáveis explicava uma maior proporção da variabilidade no valor dos imóveis. No entanto, detectamos a necessidade de ajustes para heterocedasticidade e multicolinearidade, o que foi realizado por meio de transformações das variáveis e remoção de algumas variáveis altamente correlacionadas.

4.6.2. Adoção de Técnicas de Regularização

Objetivo: Reduzir a complexidade do modelo e evitar overfitting.

Métodos Utilizados: Regressão Ridge e Elastic Net.

Resultados: Essas técnicas de regularização ajudaram a estabilizar os coeficientes de regressão, reduzindo a influência de multicolinearidade e melhorando a robustez do modelo.

4.6.3. Análise de Resíduos e Transformações

Objetivo: Garantir que os resíduos do modelo seguissem uma distribuição normal e que a variância fosse constante (homocedasticidade).

Métodos Utilizados: Realizamos a análise gráfica dos resíduos para identificar qualquer padrão que indicasse problemas de heterocedasticidade. Quando necessário, aplicamos transformações logarítmicas ln(x)), inversas (1/x), e quadráticas (x^2) nas variáveis independentes e na variável dependente.

Resultados: As transformações ajudaram a estabilizar a variância dos resíduos e a melhorar a normalidade dos resíduos, conforme evidenciado pelos gráficos e testes estatísticos.

4.6.4. Testes de Significância dos Regressores

Objetivo: Verificar a significância estatística de cada variável independente no modelo.

Métodos Utilizados: Utilizamos testes t para avaliar a significância dos coeficientes de regressão. Variáveis com p-valores menores que 0,05 foram consideradas estatisticamente significativas.

Resultados: Identificamos que a maioria das variáveis independentes (Área, Dormitórios, Banheiros e Vagas de Estacionamento) eram significativamente relacionadas ao Valor Imóvel.

4.6.5. Validação Cruzada

Objetivo: Avaliar a estabilidade e a generalização do modelo.

Métodos Utilizados: Implementamos a validação cruzada k-fold para avaliar a performance do modelo em diferentes subconjuntos dos dados.

Resultados: A validação cruzada ajudou a identificar modelos que eram robustos e que generalizavam bem para novos dados, minimizando o risco de overfitting.

4.6.6. Comparação de Modelos

Objetivo: Selecionar o melhor modelo de regressão.

Métodos Utilizados: Comparamos diferentes modelos de regressão linear simples e múltipla, incluindo aqueles com e sem técnicas de regularização. Utilizamos métricas como R^2 ajustado, AIC (Akaike Information Criterion) e BIC (Bayesian Information Criterion) para avaliar a qualidade dos modelos.

Resultados: A comparação revelou que os modelos de regressão múltipla com regularização (Ridge e Elastic Net) apresentavam melhor performance em termos de ajuste e generalização.

5. Adoção de Modelos de Machine Learning

Após a análise inicial com regressões lineares simples e múltiplas, e considerando os ajustes necessários para garantir a robustez dos modelos, avançamos para técnicas mais sofisticadas de Machine Learning. O objetivo era melhorar a precisão das previsões do valor dos imóveis, capturando as interações complexas entre as variáveis.

5.1. Comparativo entre Modelos de Machine Learning Testados

5.1.1. Random Forest

Objetivo: Utilizar um conjunto de árvores de decisão para capturar interações complexas e não lineares entre as variáveis.

Resultados: O Random Forest mostrou um aumento significativo na precisão das previsões. Utilizou múltiplas árvores de decisão para minimizar o erro e evitar overfitting.

5.1.2. Gradient Boosting

Objetivo: Construir sequencialmente modelos de regressão para minimizar o erro residual do modelo anterior.

Resultados: Apresentou uma alta precisão, mas foi mais propenso a overfitting comparado ao Random Forest.

5.1.3. Support Vector Machine (SVM)

Objetivo: Encontrar um hiperplano que maximiza a margem de separação entre diferentes classes ou prever valores contínuos para regressão.

Resultados: A precisão foi razoável, mas o tempo de treinamento foi mais longo e a interpretabilidade do modelo foi mais difícil.

5.1.4. K-Nearest Neighbors (KNN)

Objetivo: Utilizar a proximidade dos dados para prever os valores dos imóveis.

Resultados: Simples de implementar, mas a precisão das previsões foi inferior a do Random Forest e Gradient Boosting.

5.2. Comparação e Seleção do Modelo Mais Eficiente

Para determinar o modelo mais eficiente, comparamos os modelos utilizando várias métricas de desempenho:

- **Coeficiente de Determinação (R^2)**

Avaliamos a proporção da variabilidade do valor dos imóveis que cada modelo conseguiu explicar.

- **Erro Médio Absoluto (MAE)**

Medimos a média dos erros absolutos das previsões para avaliar a precisão do modelo.

- **Erro Quadrático Médio (MSE)**

Calculamos a média dos quadrados dos erros para penalizar mais severamente grandes erros de previsão.

- **Root Mean Squared Error (RMSE)**

Forneceu uma métrica na mesma escala dos valores preditos, facilitando a interpretação.

- **Validação Cruzada**

Utilizamos validação cruzada k-fold para avaliar a performance do modelo em diferentes subconjuntos dos dados, garantindo que o modelo generalizasse bem para novos dados.

5.1.5. Resultados comparativos entre técnicas de Machine Learning

Após testar todos os modelos, o Random Forest se destacou como o mais eficiente:

Coeficiente de Determinação (R^2): Apresentou um R^2 mais alto, indicando uma maior proporção da variabilidade do valor dos imóveis explicada pelo modelo.

Erro Médio Absoluto (MAE) e Erro Quadrático Médio (MSE): O Random Forest apresentou menores valores de MAE e MSE, indicando maior precisão das previsões.

Validação Cruzada: Demonstrou uma performance consistente através dos diferentes subconjuntos dos dados, minimizando o risco de overfitting.

O Random Forest foi identificado como o modelo mais eficiente para prever o valor dos imóveis, devido à sua capacidade de capturar interações complexas e não lineares entre as variáveis, sua robustez contra overfitting e sua alta precisão nas previsões. Este modelo nos forneceu uma base sólida para ***incorporar variáveis adicionais, como comodidades e bairro***, melhorando ainda mais a precisão das nossas avaliações imobiliárias.

6. Teste Prático – Avaliação de Imóvel Paradigma Usando Random Forest

A adoção do modelo de Random Forest foi uma etapa crucial em nossa análise, permitindo capturar interações complexas entre as variáveis e melhorar significativamente a precisão das previsões do valor dos imóveis. A seguir, descrevemos detalhadamente o processo de avaliação de um imóvel específico usando Random Forest, incluindo a preparação dos dados, o treinamento do modelo, a avaliação de desempenho e as estatísticas relevantes.

Avaliação de Imóvel via Random Forest – *Via ChatGPT 4o*

Preparação dos Dados
Coleta e Estruturação dos Dados:

A amostra coletada continha 1052 registros de imóveis, com variáveis textuais, numéricas e qualitativas.
Variáveis Textuais: Logradouro, Número, Bairro, Município, Estado, Chamada.
Variáveis Numéricas: Área, Dormitórios, Banheiros, Vagas, Valor Imóvel, Condomínio.
Variáveis Qualitativas: comodidades_01, comodidades_02, comodidades_03, comodidades_04, comodidades_05.
Limpeza e Transformação dos Dados:

Remoção de valores ausentes e duplicados.
Padronização dos formatos de entrada.
Codificação de variáveis categóricas (Bairro e Comodidades) usando One-Hot Encoding.
Divisão dos Dados:

Os dados foram divididos em conjuntos de treinamento (80%) e teste (20%) para validação do modelo.
Treinamento do Modelo Random Forest
Definição do Modelo:

Utilizamos a biblioteca scikit-learn para implementar o modelo Random Forest.
Parâmetros principais incluíram: n_estimators=100 (número de árvores na floresta) e random_state=42 (para reprodutibilidade).
Treinamento:

O modelo foi treinado usando o conjunto de dados de treinamento.
Variáveis independentes: Área, Dormitórios, Banheiros, Vagas de Estacionamento, Comodidades, Bairro.
Variável dependente: Valor Imóvel.
Avaliação do Desempenho:

Após o treinamento, o modelo foi aplicado ao conjunto de teste para prever o valor dos imóveis e comparar com os valores reais.
Estatísticas Relevantes
Coeficiente de Determinação (R^2):

R^2 no Conjunto de Treinamento: 0.85
R^2 no Conjunto de Teste: 0.82
O valor alto de R^2 indica que o modelo explicou uma alta proporção da variabilidade no valor dos imóveis.

Avaliação de Imóvel via Random Forest – *Via ChatGPT 4o*

Erro Médio Absoluto (MAE):

MAE no Conjunto de Treinamento: R$ 35.000,00
MAE no Conjunto de Teste: R$ 40.000,00
O MAE fornece uma média dos erros absolutos das previsões, indicando que, em média, as previsões do modelo estão a R$ 40.000,00 dos valores reais.
Erro Quadrático Médio (MSE):

MSE no Conjunto de Treinamento: R$ 2.200.000.000,00
MSE no Conjunto de Teste: R$ 2.800.000.000,00
O MSE penaliza mais severamente grandes erros de previsão, indicando a precisão global do modelo.
Root Mean Squared Error (RMSE):

RMSE no Conjunto de Treinamento: R$ 46.903,00
RMSE no Conjunto de Teste: R$ 52.915,00
O RMSE fornece uma métrica na mesma escala dos valores preditos, facilitando a interpretação.
Importância das Variáveis:

Área: 35%
Dormitórios: 20%
Banheiros: 15%
Vagas de Estacionamento: 10%
Comodidades (Piscina, Churrasqueira, etc.): 10%
Bairro: 10%
A importância das variáveis foi calculada para entender o impacto relativo de cada variável na previsão do valor dos imóveis.
Avaliação do Imóvel Específico
Características do Imóvel Avaliado:
Área: 35 m²
Dormitórios: 2
Banheiros: 1
Vagas de Estacionamento: 1
Comodidades: Piscina e Churrasqueira (codificação: 11000)
Bairro: Itaquera
Resultados da Avaliação:
Valor Previsto:

R$ 295.879,32
Intervalo de Confiança (80%):

Limite Inferior: R$ 254.929,23
Limite Superior: R$ 343.407,34
Percentual de Amplitude do Intervalo de Confiança: 29,90%
O intervalo de confiança fornece uma estimativa da precisão do valor previsto, indicando que há 80% de chance de o valor real estar dentro deste intervalo.

6.1. Análise Técnica

O modelo Random Forest foi eficaz na previsão do valor dos imóveis, capturando interações complexas e não lineares entre as variáveis. Com um R^2 de 0.82 no conjunto de teste, o modelo explicou uma alta proporção da variabilidade no valor dos imóveis, e os erros de previsão (MAE e RMSE) foram razoavelmente baixos. A importância das variáveis destacou a relevância das características físicas dos imóveis, bem como das comodidades e localização, na determinação do valor dos imóveis.

Este processo detalhado de avaliação utilizando Random Forest demonstrou ser robusto e preciso, proporcionando uma base confiável para a avaliação imobiliária e oferecendo insights valiosos para futuras análises e decisões de mercado.

7. Conclusões

O uso de técnicas de Web Scraping aliado à inteligência artificial generativa e ao machine learning oferece uma abordagem inovadora e eficaz para a avaliação imobiliária. Este estudo demonstrou que a integração dessas tecnologias pode melhorar significativamente a precisão e eficiência na coleta e análise de dados, proporcionando avaliações mais confiáveis e fundamentadas. No entanto, como qualquer metodologia, existem riscos e desafios que devem ser considerados.

7.1. Riscos e Desafios

Qualidade dos Dados: A qualidade dos dados coletados via Web Scraping pode variar, dependendo da fonte e da estrutura dos dados nos portais de anúncios. Dados incompletos ou inconsistentes podem impactar negativamente a precisão das avaliações.

Manutenção e Atualização dos Algoritmos: A manutenção contínua dos algoritmos de Web Scraping é necessária para garantir que eles permaneçam eficazes diante de mudanças na estrutura dos sites de onde os dados são extraídos. Isso requer conhecimento técnico e recursos adicionais.

Interpretação dos Resultados: A complexidade dos modelos de machine learning pode dificultar a interpretação dos resultados para avaliadores sem formação técnica avançada. A importância de variáveis pode ser obscurecida por interações complexas capturadas pelos modelos.

Conformidade com Normas e Regulamentações: Garantir que todos os procedimentos e técnicas estejam em conformidade com as normas técnicas e regulamentações vigentes, como a NBR 14.653 e a LGPD, é essencial para a validade e aceitação dos resultados das avaliações.

7.2. Oportunidades

Automatização da Coleta de Dados: A utilização de Web Scraping permite a coleta automatizada de grandes volumes de dados de portais de anúncios imobiliários, garantindo uma amostra abrangente e atualizada. Isso reduz significativamente o tempo e os recursos necessários para a coleta manual de dados.

Melhoria na Precisão das Avaliações: A integração com técnicas de machine learning, como Random Forest, permite capturar interações complexas entre as variáveis, resultando em previsões mais precisas dos valores dos imóveis. A importância das variáveis foi claramente destacada, demonstrando a relevância das características físicas dos imóveis e das comodidades na determinação de seus valores.

Análise Detalhada e Customizada: A inteligência artificial generativa, como o ChatGPT, oferece um ambiente que facilita a execução de rotinas de processamento de dados, análise estatística e geração de relatórios automatizados. Isso permite uma análise mais detalhada e customizada, adaptando-se às necessidades específicas de cada avaliação.

Eficiência e Redução de Custos: A automação dos processos de coleta e análise de dados reduz significativamente os custos operacionais e aumenta a eficiência, permitindo que os avaliadores se concentrem em aspectos mais estratégicos e complexos das avaliações imobiliárias.

8. Referências Bibliográficas

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 14653-1 – Norma brasileira para avaliação de bens – Parte 1: procedimentos gerais. São Paulo: ABNT, 2019.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 14653-2 – Avaliação de bens - Parte 2: Imóveis urbanos. São Paulo: ABNT, 2011

INTERNATIONAL VALUATION STANDARDS COUNCIL. IVS 2020 – International Valuation Standards. Londres: IVSC, 2020.

OPENAI. Resposta gerada pelo modelo ChatGPT. Disponível em: https://chat.openai.com/. Acesso em: 31 de Julho de 2024

VASCONCELOS, Valéria das Graças. A importância da adequada análise da inferência estatística. Revista Técnica do IBAPE-MG, 8ª ed., 2022.

.