Autor: Bruno Henrique Gazzinelli
Engenheiro Civil - CREA-MG 235.299/D – IBAPE 1.113 - Belo Horizonte/MG
bruno@bhgengenharia.com
Co-Autor: Luiz Flávio de Moares Tamietti
Engenharia Civil – Belo Horizonte/MG

WEBSCRAPPING, CHATGPT E MACHINE LEARNING: UMA BREVE ABORDAGEM

PALAVRAS-CHAVE: RASPAGEM DE DADOS. CHATGPT. APRENDIZADO DE MÁQUINA. INTELIGÊNCIA ARTIFICIAL GENERATIVA. AVALIAÇÃO DE IMÓVEIS.

INTRODUÇÃO

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.

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.

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

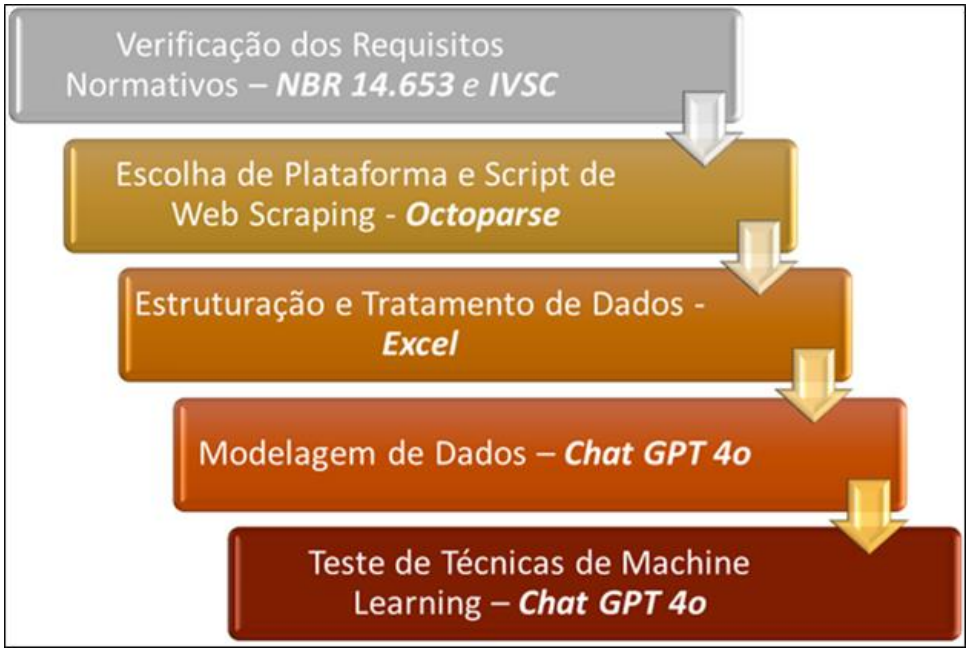

Figura 1 – Desenvolvimento da Raspagem Análise de Dados - Fonte: Autores

SEQUÊNCIA LÓGICA DE OBTENÇÃO DE DADOS

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

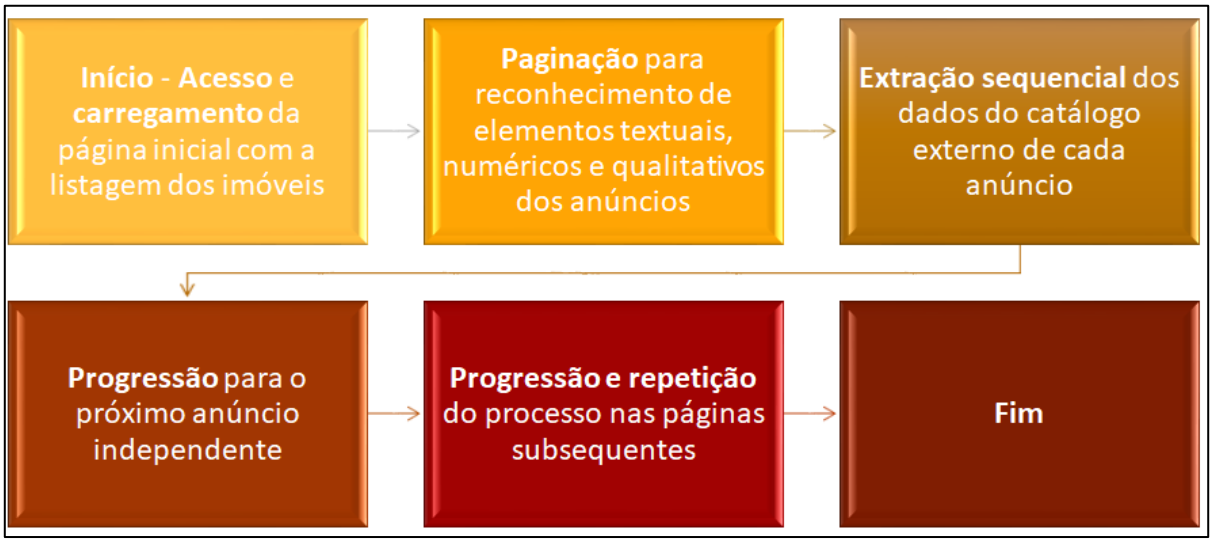

Figura 1: Sequência Lógica de Obtenção de Dados - Fonte: Autores

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.

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.

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.

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 Artificial**.

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).

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

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.
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.

Figura 3: Fluxo de Análise de Dados de Regressão - Fonte: Autores

ANÁLISE TÉCNICA

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.

BREVE RESUMO DE CONCLUSÕES

1 - Automatização da Coleta: A implementação de Web Scraping permitiu a coleta eficiente de grandes volumes de dados imobiliários, reduzindo significativamente o tempo e os recursos que seriam necessários em um processo manual; **2 - Precisão das Avaliações:** A integração com técnicas de machine learning, como o Random Forest, mostrou-se eficaz ao capturar interações complexas entre variáveis, resultando em previsões mais precisas dos valores dos imóveis**. 3 - Análise Personalizada:** O uso de inteligência artificial generativa, como o ChatGPT, proporcionou um ambiente robusto para o processamento de dados e a geração de relatórios personalizados, atendendo às necessidades específicas de cada avaliação. **4 - Eficiência e Custos:** A automação dos processos não apenas aumentou a eficiência das operações, mas também reduziu os custos, permitindo que os avaliadores concentrem seus esforços em aspectos mais estratégicos e complexos das avaliações imobiliárias.

REFERÊNCIAS

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

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.