

MESA REDONDA – INTELIGÊNCIA ARTIFICIAL NA ENGENHARIA DE AVALIAÇÕES

Eng. Bruno Henrique Gazzinelli

00 – Apresentação



- Inicialmente, apresenta-se a seguir os **tópicos** que serão abordados neste breve momento:

1. Introdução e Objetivo do Estudo
2. Metodologia;
3. Estudo de Caso
4. Preparação da Modelagem;
5. Machine Learning;
6. Teste Prático – Random Forest;
7. Conclusões.

1 - Introdução e Objetivo do Estudo

Como e porque combinar métodos de *Web Scraping*,
Inteligência Artificial Descritiva e *Machine Learning*?

01 - Contextualização e Relevância do Estudo



- **Complexidade da Avaliação Imobiliária**

- A avaliação imobiliária é um processo que exige a análise detalhada de diversas variáveis para alcançar precisão e confiabilidade. Esse cenário destaca a importância da **coleta de dados** robusta e estruturada.

- **Aplicação do Web Scraping**

- Com o crescimento dos dados online, o Web Scraping torna-se uma ferramenta eficaz para a **coleta automatizada** de grandes volumes de dados, oferecendo informações detalhadas sobre o mercado imobiliário.

01 - Contextualização e Relevância do Estudo



- **Integração de Inteligência Artificial**

- A inteligência artificial generativa, como o ChatGPT, possibilita a automação de tarefas, desde o **processamento de dados** até a geração de relatórios, tornando o processo de avaliação mais rápido e preciso.

- **Teste Prático de aplicação de I.A. para Avaliações**

- Possibilidade de **(1)** Avaliar a capacidade de **cumprimento de diretrizes normativas**, da **(2)** Realização de testes e adequação de pressupostos estatísticos, do **(3)** entendimento e **processamento de informações** quantitativas e qualitativas, bem como do **(4)** **Treinamento de um modelo ou rotina** a partir de um conjunto robusto de dados.

01 - Objetivo do Estudo



- 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.

01 - Objetivo do Estudo

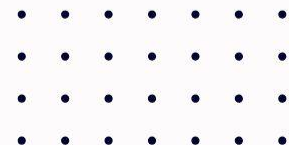


- Este estudo busca não apenas apresentar uma **metodologia prática e replicável** para avaliadores de imóveis, mas também fornecer insights sobre a importância da análise de dados e do uso de técnicas avançadas de machine learning e inteligência artificial generativa na avaliação imobiliária.

2 - Metodologia

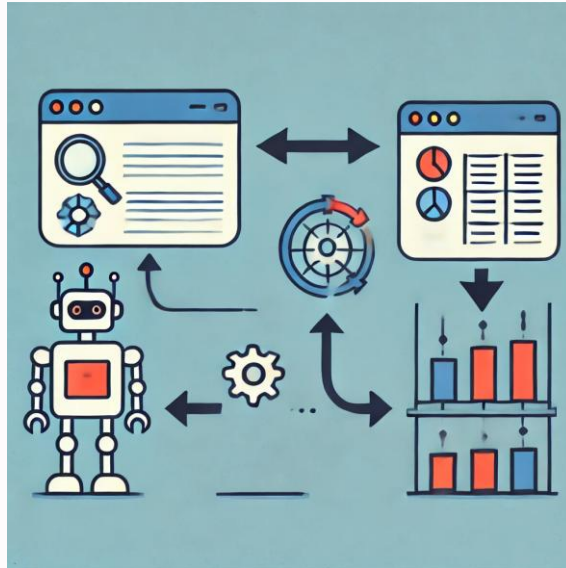
Detalha-se a seguir, a **metodologia desenvolvida** para o Estudo de Caso desenvolvido neste trabalho.

02 - Metodologia



02 – Escolha da Plataforma

- **Plataforma Utilizada: Octoparse**



- Para a Raspagem de Dados foi escolhida a plataforma “**Octoparse**”. 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.

02 - Fluxograma



3 – Estudo de Caso

Mercado Imobiliário Residencial de *imóveis de pequena metragem* em São Paulo – SP - Brasil

03 – Escolha de Tipologia



Definiu-se inicialmente a obtenção de dados de apartamentos que apresentassem áreas no intervalo de 20 a 50 m².

- 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.

03 – Refinamento e Tratamento

Classe
Textual

- Verificação de duplicidade de oferta através de verificação do campo de identificação da oferta.

Classe
Numérica

- Verificação de preenchimento incorreto ou inadequado dos campos numéricos;

Classe
Qualitativa

- Análise de declaração de campos e qualidade das informações.

03 – Amostragem Definitiva



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

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

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

- Amostra Final

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

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



03 – Imóvel Avaliando



Para fins de prática, definiu-se o avaliando com características não presentes na amostra quanto à variável localização, cujas características são:

- I. **Área do avaliando:** 43,00m²;
- II. **Quantidade de Dormitórios:** 01 dormitório;
- III. **Quantidade de Banheiros:** 01 banheiro;
- IV. **Vagas de Garagem:** 01 vaga de garagem.

03 – Conjunto de Dados



Ainda, para fins de estudo, foram informadas características quanto a:

- I. **Localização dos Imóveis:** Em classe textual;
- II. **Benfeitorias presentes:** Em formato pré qualitativo, em campos distintos, mas ainda em classe textual;
- III. **Descrições internas dos anúncios:** Classe textual declarada pelos autores dos anúncios

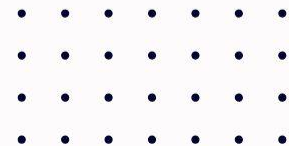
4 – Preparação da Modelagem

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.

04 – Integração com I.A. – L.L.M.



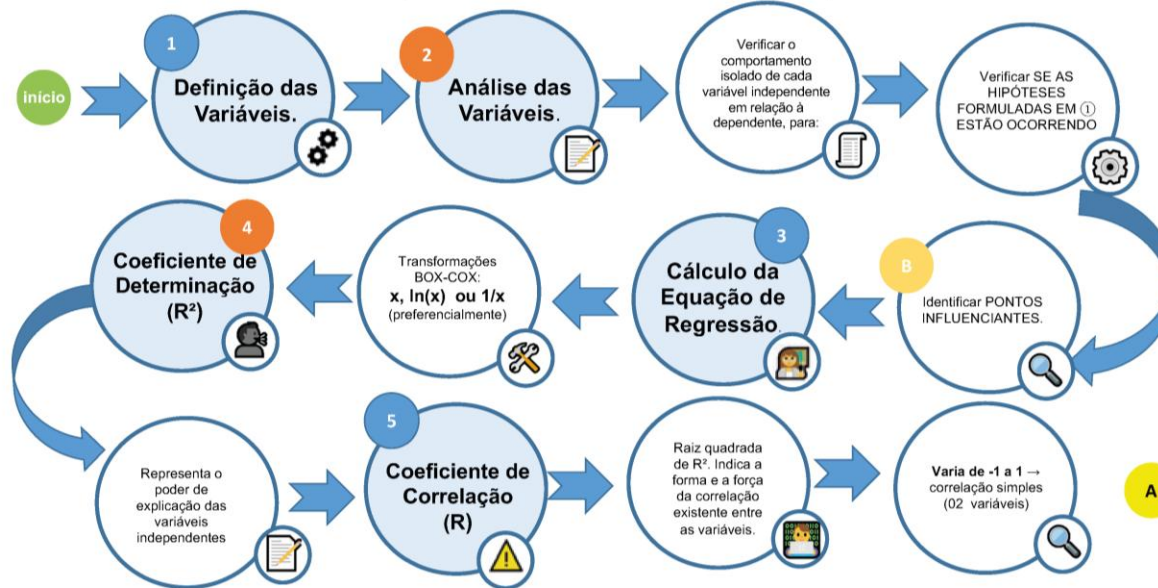
- Iniciou-se um processo de integração de processos a partir do uso do **Chat GPT 4o**, na versão paga do ambiente.
- Para tanto, estruturou-se um processo inicial de **alimentação de informações e diretrizes** para posterior prosseguimento das análises estatísticas.
- A integração fora efetivada na prática a partir de (i) **comandos diretos**, bem como do (ii) **upload de arquivos** em **.xlsx (Excel)** e **.pdf (Adobe Acrobat)** consolidando uma base de dados, bem como da definindo diretrizes técnicas.



04 – Integração com I.A. – L.L.M.

Fluxo Análise de Regressão - 1/3

Conforme práticas de mercado & ABNT NBR-14653:2/2011

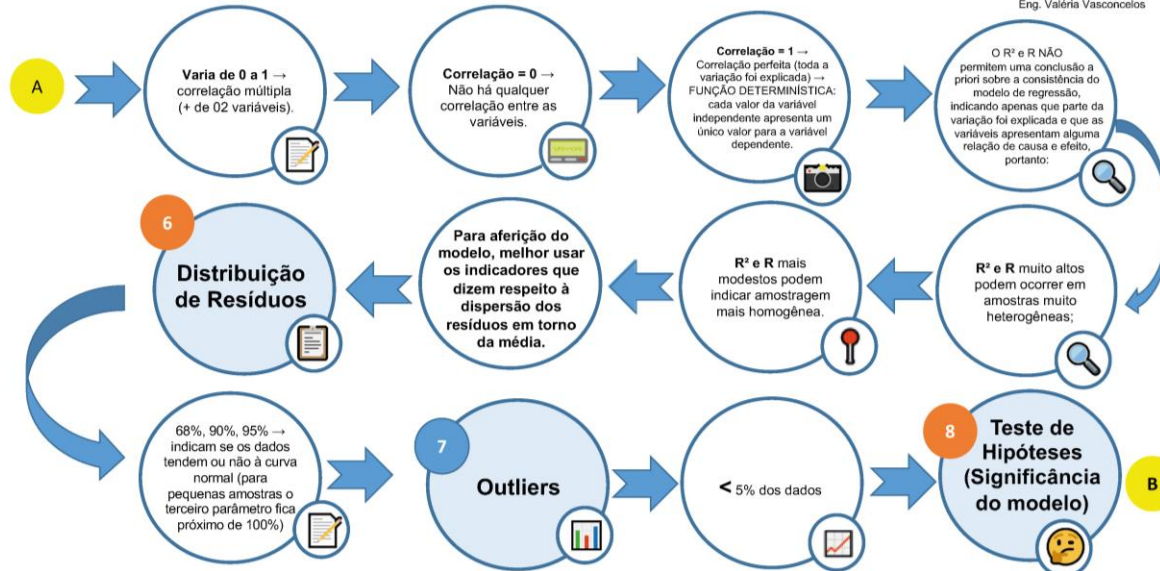


04 – Integração com I.A. – L.L.M.

Fluxo Análise de Regressão - 2/3

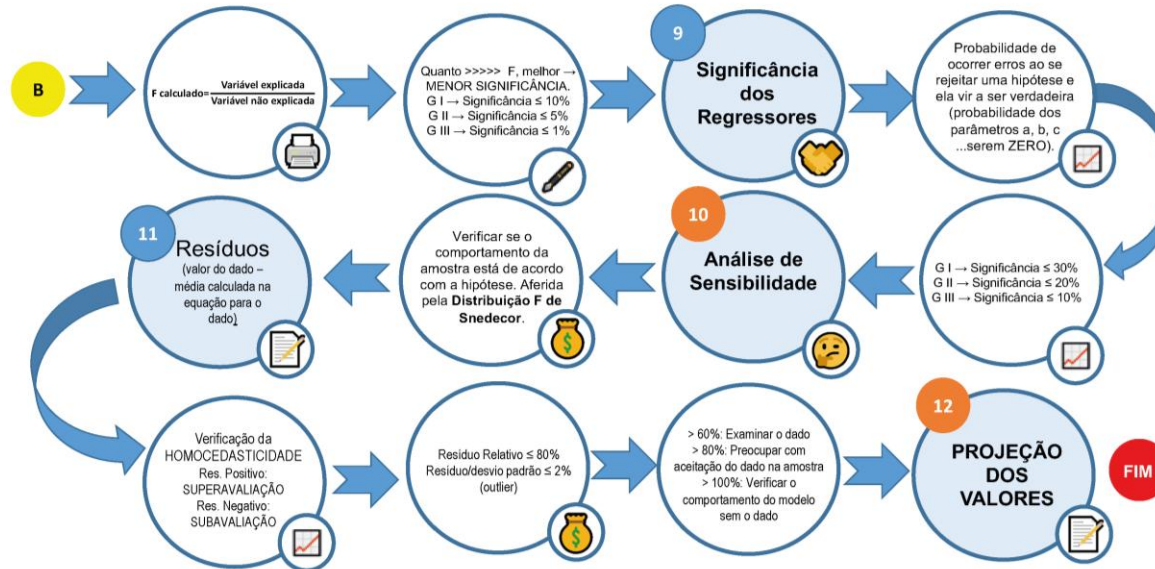
Conforme práticas de mercado & ABNT NBR-14653:2/2011

Revisão: 00
Data: 28/10/2020
Elaborado por:
Eng. Valéria Vasconcelos
Verificado por:
Eng. Valéria Vasconcelos



04 – Integração com I.A. – L.L.M.

Fluxo Análise de Regressão - 3/3 Conforme práticas de mercado & ABNT NBR-14653:2/2011



04 – Fluxo de Análise

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

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.

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: $\leq 10\%$

Grupo II: $\leq 5\%$

Grupo III: $\leq 1\%$

Análise de Sensibilidade: Verificar o comportamento da amostra conforme a hipótese.

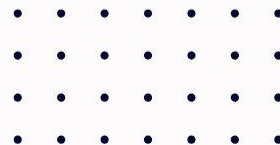
04 – Regressão Linear Simples



- 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.

- A regressão linear seguiu, em ordem:
 - I. Preparação dos Dados;
 - II. Execução da Regressão Linear;
 - III. Avaliação dos Resultados;
 - IV. Análise da Importância das Variáveis Independentes



04 – Resultados – R.L.S.

Resultados da Regressão Linear Simples – Via Chat GPT 4o

Modelo

- Variável dependente: Valor do Imóvel

- **Área: 6,735**

- Variáveis independentes: Área, Dormitórios,
Banheiros, Vagas

- Erro padrão: 991.897

- t: 6.791

- $P > |t|$: 0.000

Resultados

- Intervalo de confiança 95%: [4,789, 8,682]

- Constante (Intercepto): 396,700

- Erro padrão: 36,500

- **Dormitórios: -255,000**

- t: 10.878

- Erro padrão: 13,700

- $P > |t|$: 0.000

- t: -18.658

- Intervalo de confiança 95%: [325,000, 468,000]

- $P > |t|$: 0.000

- Intervalo de confiança 95%: [-282,000, -228,000]

04 – Resultados – R.L.S.

Resultados da Regressão Linear Simples – Via Chat GPT 4o

- Banheiros: 95,650

- Erro padrão: 25,500

- t: 3.751

- $P > |t|$: 0.000

- Intervalo de confiança 95%: [45,600, 146,000]

- 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

04 – Interpretação dos Resultados



O modelo explica aproximadamente 30% da variação no valor do imóvel (R^2 -quadrado = 0.300).

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.

04 – Aprimoramento Técnico



- **Regressão Linear Múltipla**

- Incluir **múltiplas variáveis independentes** para capturar melhor as interações entre diferentes características dos imóveis. A regressão linear múltipla apresentou uma melhora no coeficiente de determinação (R^2).



- **Adoção de Técnicas de Regularização**

- Reduzir a complexidade do modelo e evitar overfitting. Ajudaram a **estabilizar os coeficientes de regressão**, reduzindo a influência de multicolinearidade e melhorando a robustez do modelo.

- **Análise de Resíduos e Transformações**

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



04 – Aprimoramento Técnico



- **Testes de Significância dos Regressores**

- Verificar a **significância estatística** de cada variável independente no modelo. (Área, Dormitórios, Banheiros e Vagas de Estacionamento).



- **Validação Cruzada**

- Avaliar a **estabilidade** e a generalização do modelo.

- **Comparação de Modelos**

- Utilizamos métricas como R^2 ajustado, **AIC** (Akaike Information Criterion) e **BIC** (Bayesian Information Criterion) para avaliar a qualidade dos modelos.



5 – LLM - 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.

05 – Apresentação de Testes



Random Forest

- **Objetivo:** Utilizar um conjunto de árvores de decisão para capturar interações complexas e não lineares entre as variáveis.
- **Resultados:** Mostrou um aumento significativo na precisão das previsões. Múltiplas árvores de decisão para minimizar o erro e evitar overfitting.

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.

05 – Apresentação de Testes



- **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.



- **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.



05 – Modelo Mais Eficiente



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



- I. **Coeficiente de Determinação (R^2)** - Avaliamos a proporção da variabilidade do valor dos imóveis que cada modelo conseguiu explicar;
- II. **Erro Médio Absoluto (MAE)** - Medimos a média dos erros absolutos das previsões para avaliar a precisão do modelo;



05 – Modelo Mais Eficiente



III. **Erro Quadrático Médio (MSE)** - Cálculo a média dos quadrados dos erros para penalizar mais severamente grandes erros de previsão;



IV. **Raiz Quadrada do Erro Médio (RMSE)** - Forneceu uma métrica na mesma escala dos valores preditos, facilitando a interpretação;

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



05 – Modelo Mais Eficiente



- ✓ 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 (qualitativas) e localização a nível bairro (qualitativa)**, melhorando ainda mais a precisão das nossas avaliações imobiliárias.

6 – Teste Prático – Random Forest

A adoção do modelo de Random Forest foi etapa crucial na análise, permitindo ***capturar interações complexas entre as variáveis*** e melhorar significativamente a precisão das previsões do valor dos imóveis.

06 – Teste Prático



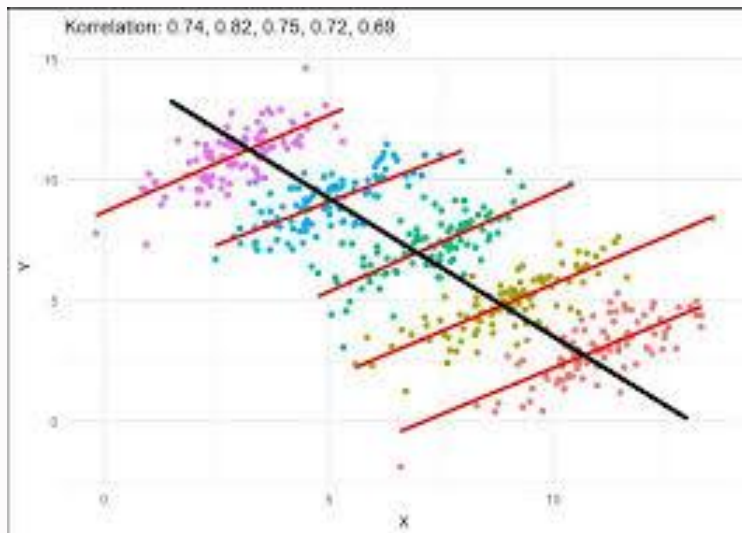
Seria possível, a partir de um modelo de treinamento gerado pela capacidade de processamento de dados e redes neurais de uma inteligência artificial, **gerar novas variáveis relevantes** para o modelo, mesmo que declaradas textualmente?

Caso positivo, seriam **estatisticamente válidas?**

06 – Teste Prático – Paradoxo de Simpsom

Regressão Linear em Preto: Dados do Modelo

Regressões em **Vermelho**: Análises a partir de adições de variáveis



06 – Teste Prático – 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**.
- 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.

06 – Teste Prático



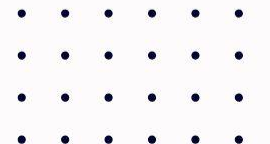
- A amostra coletada continha **1042** registros de imóveis, com variáveis textuais, numéricas e qualitativas.
- **Variáveis Textuais:** Logradouro, Número, Bairro, Município, Estado, Chamada.
- **Variáveis Numéricas:** Área, Dormitórios, Banheiros, Vagas, Valor Imóvel, Condomínio.

- **Valor Previsto:** R\$ 308.079,24



Intervalo de Confiança (80%):

- **Limite Inferior:** R\$ 265.440,66
- **Limite Superior:** R\$ 357.567,00
- Percentual de **Amplitude do Intervalo** de Confiança: 29,90%



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.**

07 – Riscos e Desafios



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.

07 – Oportunidades



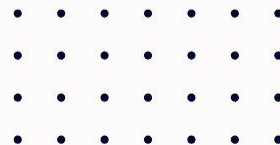
Possível Melhoria na Precisão das Avaliações: A integração com técnicas de machine learning, como Random Forest, permite capturar interações complexas entre as variáveis, resultando em previsões mais precisas dos valores dos imóveis. A importância das variáveis foi claramente destacada, demonstrando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



07 – Oportunidades



Possibilidade de **Análise Detalhada e Customizada**: A inteligência artificial generativa e o uso de redes neurais, 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.



BÔNUS DE REFLEXÃO – O pode ser o futuro?

8 – Reflexões Humanas

O que move o mundo?

Perguntas?

Respostas!



8 – Reflexões Humanas

O que a inteligência artificial gera¹?

Perguntas?

Respostas!

¹Data Base 20/11/2025



8 – Reflexões Humanas



Somos nós, engenheiros e arquitetos de avaliações, à

luz do conhecimento técnico, quem:

- Planejamos pesquisa de dados;
- Compreendemos contexto de mercado;
- Analisamos importância de variáveis;
- Construimos modelos estatísticos;
- Realizamos as avaliações!



AGRADECIMENTO



Bruno Henrique Gazzinelli



(31) 99300-9609



bruno@bhgengenharia.com