Análise Exploratória dos Dados (EDA)

Eficiência Energética de Edifícios
Relatório Técnico

Versão 1.0

Londrina, novembro de 2025

Análise Exploratória

O principal objetivo da análise exploratória é compreender a estrutura interna do conjunto de dados, identificar possíveis inconsistências e extrair informações relevantes que orientem o pré-processamento e a implementação dos modelos de forma eficiente.

Nesta etapa, considerou-se o conjunto de dados composto por **768 linhas e 10 colunas**, representando características geométricas e construtivas dos edifícios analisados. Uma primeira inspeção permitiu verificar que todas as variáveis são numéricas e que não há valores faltantes, o que elimina a necessidade de técnicas de imputação ou codificação categórica.

Assim, para a análise exploratória, considerou-se a seguinte estrutura:

Variáveis de saída (dependentes):

- *Y1 — Carga de Aquecimento*
- *Y2 — Carga de Resfriamento*

Variáveis de entrada (independentes):

- X1 — Compacidade Relativa
- X2 — Área Superficial
- X3 — Área de Parede
- X4 — Área de Telhado
- X5 — Altura Total
- X6 — Orientação
- X7 — Área de Vidro
- X8 — Distribuição da Área de Vidro

Metodologia

As análises foram realizadas com Python (Pandas, NumPy) e visualizações com Matplotlib/Seaborn. As figuras e estatísticas foram geradas a partir dos dados originais e as imagens seguintes no relatório correspondem aos gráficos produzidos durante a EDA.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 10 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   X1      768 non-null    float64
 1   X2      768 non-null    float64
 2   X3      768 non-null    float64
 3   X4      768 non-null    float64
 4   X5      768 non-null    float64
 5   X6      768 non-null    int64
 6   X7      768 non-null    float64
 7   X8      768 non-null    int64
 8   Y1      768 non-null    float64
 9   Y2      768 non-null    float64
dtypes: float64(8), int64(2)
memory usage: 60.1 KB
```

Tabela 1 – Info do Dataset

Tabela 1 descreve a estrutura do conjunto de dados, indicando que todas as 10 variáveis possuem 768 registros não nulos, o que confirma a ausência total de dados faltantes**.**

Observa-se também que todas as variáveis são numéricas, sendo oito do tipo *float64* e duas (X6 e X8) originalmente do tipo *int64*. Após o processo de normalização com o MinMaxScaler, todas passam a assumir o tipo *float64*, garantindo uniformidade e compatibilidade para os procedimentos de análise estatística e modelagem. No geral, o dataset apresenta excelente consistência e encontra-se plenamente adequado para as etapas exploratórias e preditivas subsequentes.

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375	2.81250	22.307195	24.587760
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221	1.55096	10.090204	9.513306
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000	0.00000	6.010000	10.900000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000	1.75000	12.992500	15.620000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000	3.00000	18.950000	22.080000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000	4.00000	31.667500	33.132500
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000	5.00000	43.100000	48.030000

Tabela 2 – describe

A Tabela 2 apresenta as principais estatísticas descritivas das variáveis que compõem o conjunto de dados utilizado neste estudo. As medidas incluem média, desvio-padrão, valores mínimo e máximo, além dos quartis (25%, 50% e 75%), permitindo compreender a distribuição e o comportamento geral de cada atributo.

Observa-se inicialmente que todas as variáveis apresentam valores distintos ao longo das 768 observações, não havendo qualquer caso de variância nula. Isso indica que todas as variáveis carregam algum nível de informação relevante para o processo de modelagem, já que variáveis constantes não contribuem para a predição.

As variáveis X1 (Compacidade Relativa) e X7 (Área de Vidro) apresentam desvios-padrão reduzidos, indicando baixa variabilidade. Esse comportamento pode limitar o impacto

dessas variáveis nos modelos de regressão, uma vez que atributos com pouca variação tendem a ter menor poder discriminativo. Em contraste, variáveis dimensionais como X2 (Área Superficial)**,** X3 (Área de Parede) e X4 (Área de Telhado) exibem maior amplitude e dispersão, sugerindo maior diversidade estrutural entre as edificações simuladas, o que costuma ser favorável para modelos preditivos.

A variável X5 (Altura Total) apresenta valores discretos e padronizados (3.5, 5.25 e 7.0), evidenciando que sua variação foi definida por parâmetros específicos durante as simulações. De maneira semelhante, X6 (Orientação) também possui comportamento discreto, refletindo diferentes posicionamentos direcionais aplicados aos modelos arquitetônicos avaliados.

As variáveis-alvo Y1 (Carga de Aquecimento) e Y2 (Carga de Resfriamento) apresentam grande amplitude de variação, com valores mínimos e máximos bem distintos, além de desvios-padrão elevados. Esse comportamento é desejável em tarefas de regressão, pois oferece maior riqueza informacional para o aprendizado das relações entre as variáveis de entrada e os resultados térmicos simulados.

Outro aspecto relevante evidenciado pela Tabela 2 é a presença de escalas numéricas muito distintas entre as variáveis. Enquanto algumas assumem valores próximos de zero, outras apresentam magnitudes superiores a centenas. Essa diferença pode influenciar de maneira desproporcional vários algoritmos de aprendizado, especialmente aquelas sensíveis à escala dos dados. Por esse motivo, tornou-se necessário aplicar um procedimento de escalonamento, sendo utilizado o MinMaxScaler, que transforma os dados para o intervalo entre 0 e 1.

Depois do escalonamento, tornou-se possível comparar diretamente a dispersão das variáveis por meio dos boxplots apresentados na Figura 1, permitindo identificar outliers, amplitudes relativas e padrões de assimetria importantes para o pré-processamento.

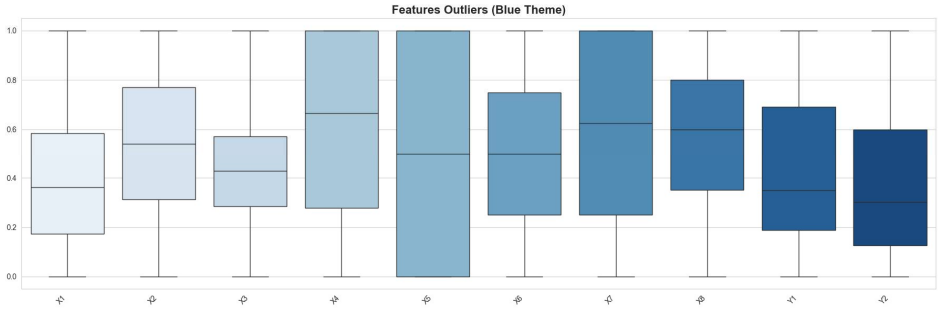

Figura 1 – Boxplots

A Figura 1 apresenta os boxplots das variáveis após o escalonamento MinMaxScaler, o que permite comparar diretamente suas distribuições usando a mesma escala (0 a 1). Esse procedimento facilita identificar padrões importantes, como amplitudes relativas, presença de valores extremos e diferenças na variabilidade entre os atributos.

Observa-se que X1, X2, X3 e X4 possuem distribuições bem equilibradas: seus valores estão espalhados ao longo de toda a faixa normalizada, com poucos sinais de outliers. Isso indica que essas variáveis têm boa diversidade e representam diferentes

características estruturais das edificações simuladas — aspecto positivo para modelos de regressão.

As variáveis X5 (Altura Total) e X6 (Orientação) apresentam padrões discretos, formando boxplots mais segmentados. Esse comportamento reforça que seus valores foram definidos por parâmetros específicos nas simulações, o que limita a variabilidade, mas ainda preserva informações relevantes.

A variável X7 (Área de Vidro) mostra baixa dispersão, com valores concentrados em uma região estreita do gráfico. Esse padrão indica variabilidade limitada, o que pode reduzir sua capacidade de diferenciar os casos no processo de modelagem.

Já as variáveis-alvo Y1 (Carga de Aquecimento) e Y2 (Carga de Resfriamento) mantêm grande amplitude mesmo após o escalonamento, refletindo forte variação térmica entre as diferentes configurações arquitetônicas simuladas — algo essencial para que modelos de regressão aprendam relações mais complexas.

No conjunto, os boxplots permitem visualizar de forma clara quais variáveis apresentam maior ou menor variabilidade, possíveis assimetrias e características que podem influenciar diretamente o desempenho dos modelos. Essa análise é fundamental para orientar decisões de pré-processamento, seleção de atributos e verificação de outliers antes da modelagem final.

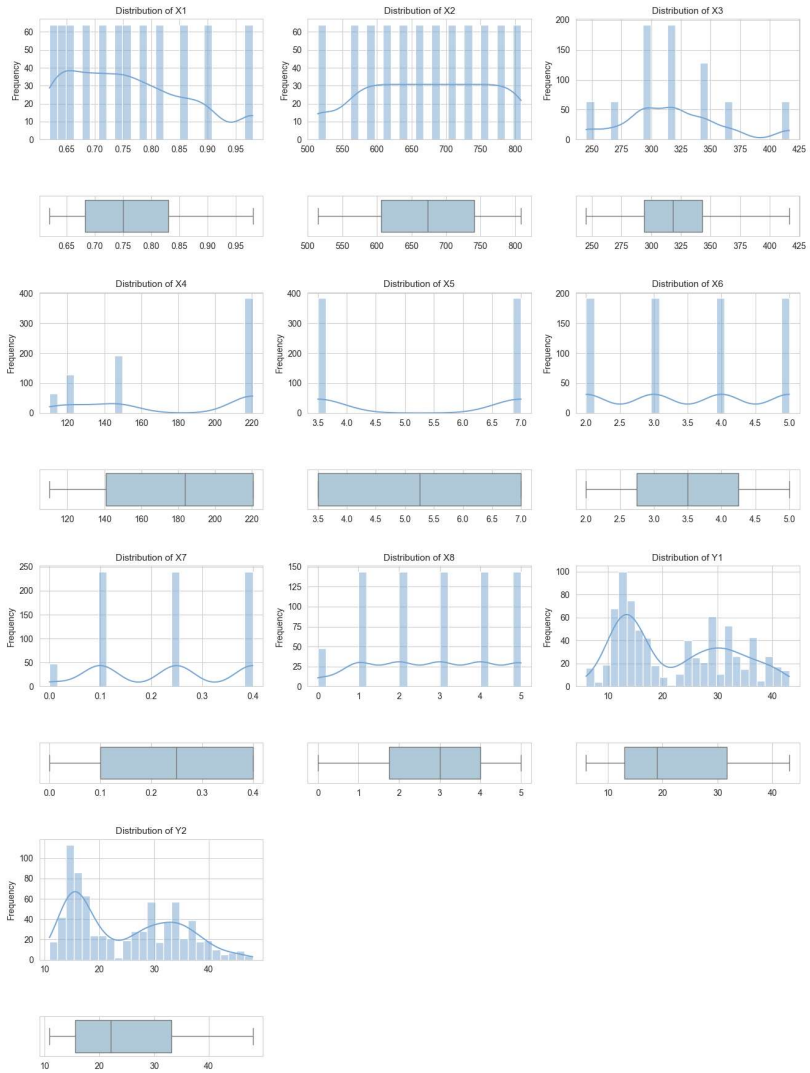

Figura 2 – Distribuição das variáveis

A Figura 2 apresenta os histogramas e boxplots das variáveis do conjunto de dados, exibindo apenas aquelas que possuem variação real, já que variáveis constantes não agregariam informação visual. As distribuições revelam comportamentos distintos entre os atributos, o que auxilia na compreensão do dataset e orienta etapas de pré-processamento.

As variáveis X1 e X2 apresentam distribuições próximas ao uniforme, indicando boa diversidade nos valores simulados e favorecendo a modelagem. Já X3 e X4 exibem picos mais concentrados, refletindo combinações geométricas específicas utilizadas durante a geração dos dados. As variáveis X5 e X6 possuem distribuições claramente discretas, com poucos valores possíveis, evidenciando características parametrizadas das simulações. X7 apresenta baixa variabilidade, com valores concentrados em uma faixa estreita, enquanto X8 mostra categorias discretas com frequências relativamente equilibradas.

Entre as variáveis-alvo, Y1 apresenta distribuição multimodal e assimétrica, indicando diferentes regimes térmicos decorrentes das configurações simuladas. Y2 também é

assimétrica, com forte concentração em valores mais baixos e uma cauda longa, refletindo maior variabilidade nas condições de resfriamento.

De modo geral, as distribuições da Figura 2 evidenciam a heterogeneidade do conjunto de dados e reforçam a necessidade de cuidados no pré-processamento, incluindo normalização, verificação de assimetrias e avaliação de potenciais outliers antes da modelagem preditiva.

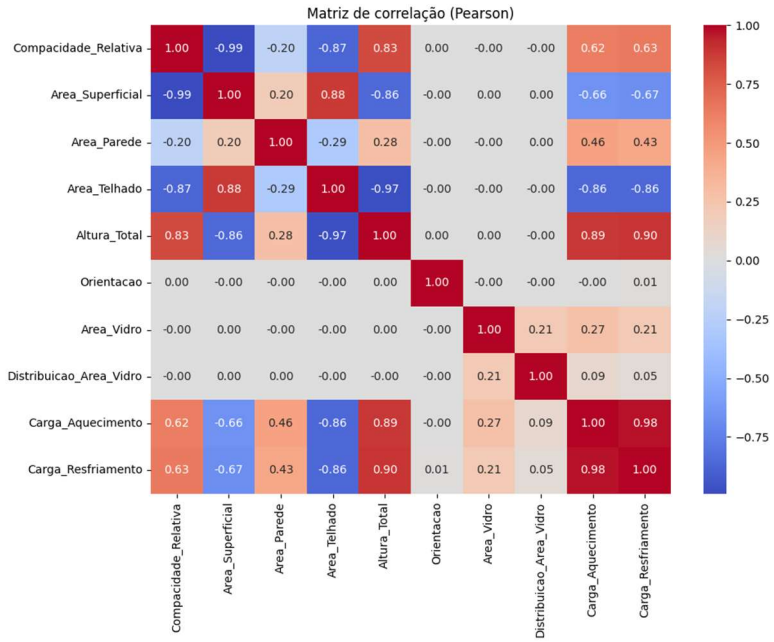

Figura 3 – Matriz de correlações

A Figura 3 apresenta a matriz de correlação de Pearson, utilizada para identificar relações lineares entre as variáveis do conjunto de dados. A análise revela forte correlação positiva entre as variáveis geométricas (Área Superficial, Área Parede, Área Telhado e Altura Total), resultado esperado devido à dependência direta entre as dimensões físicas do edifício. Em contraste, a Compacidade Relativa apresenta correlação fortemente negativa com a Área Superficial, refletindo sua definição geométrica — edifícios mais compactos possuem menor área exposta.

As variáveis relacionadas ao vidro (Área_Vidro e Distribuicao_Area_Vidro) exibem correlações fracas com as dimensões geométricas, indicando independência entre a proporção de superfícies envidraçadas e o tamanho geral das edificações. A variável Orientação também apresenta correlações próximas de zero, o que é consistente com sua natureza categórica e sua falta de relação direta com medidas geométricas.

Quanto às variáveis-alvo, Carga_Aquecimento e Carga_Resfriamento demonstram correlação extremamente alta entre si, sugerindo comportamentos térmicos fortemente interligados. Ambas também apresentam correlação significativa com as variáveis geométricas, indicando que edifícios maiores tendem a demandar mais energia para aquecimento e resfriamento, enquanto edificações mais compactas requerem menos energia.

No conjunto, a matriz evidencia a existência de multicolinearidade entre as variáveis geométricas, aspecto relevante para modelos lineares ou sensíveis à redundância entre atributos. Essa observação reforça a necessidade de seleção de variáveis ou de técnicas avançadas, como regularização ou redução de dimensionalidade, para mitigar possíveis efeitos negativos na modelagem.

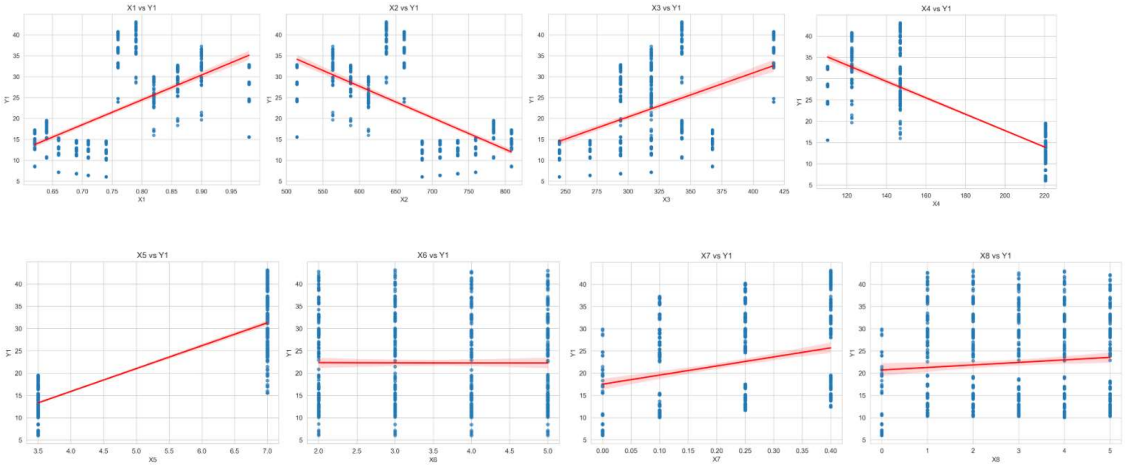

Figura 4: Scatterplots de Y1 versus variáveis preditoras

A Figura 4 apresenta os gráficos de dispersão entre a variável-alvo Y1 (Carga de Aquecimento) e cada uma das variáveis preditoras. Esses scatterplots permitem avaliar visualmente a existência de tendências lineares, relações fracas ou padrões não lineares entre os atributos geométricos e a resposta térmica simulada.

Variáveis discretas como X5 (Altura Total), X6 (Orientação) e X8 (Distribuição da Área de Vidro) formam agrupamentos verticais bem definidos, refletindo o fato de que assumem apenas um conjunto restrito de valores. Nesses casos, a linha de regressão tende a apresentar pouca inclinação, indicando ausência de uma relação linear clara.

Entre as variáveis contínuas, observam-se tendências coerentes com a matriz de correlação: X1 (Compacidade Relativa) e X3 (Área da Parede) exibem relação positiva com Y1, sugerindo que edificações menos compactas ou com maior área de parede tendem a demandar mais energia para aquecimento. Em contraste, X2 (Área Superficial) e X4 (Área de Telhado) apresentam relação negativa, indicando que, em determinadas configurações geométricas, aumentos nessas áreas podem estar associados à redução da carga de aquecimento.

As variáveis relacionadas ao vidro, X7 e X8, demonstram relações visivelmente fracas, com grande dispersão e linhas de tendência pouco inclinadas, reforçando que seu impacto direto na carga térmica é limitado e possivelmente dependente de interações com outros fatores.

No conjunto, os scatterplots sugerem que a carga de aquecimento é influenciada por múltiplos atributos simultaneamente, e que relações lineares simples capturam apenas parte do comportamento observado. Esses resultados reforçam a necessidade de

modelos capazes de considerar padrões mais complexos ou combinações de variáveis para representar adequadamente o fenômeno térmico analisado.

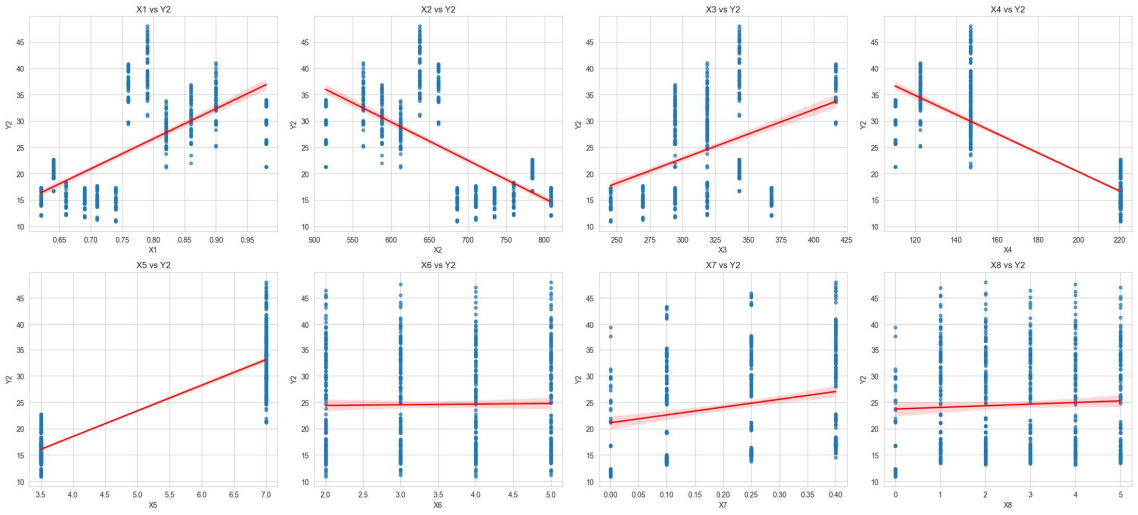

Figura 5 — Scatterplots de Y2 versus variáveis preditoras

A Figura 5 apresenta os scatterplots entre a variável-alvo Y2 (Carga de Resfriamento) e as oito variáveis preditoras. Esses gráficos permitem visualizar tendências lineares e padrões de dispersão que ajudam a compreender como cada característica geométrica influencia a carga térmica de resfriamento.

Variáveis discretas como X5 (Altura Total), X6 (Orientação) e X8 (Distribuição da Área de Vidro) geram agrupamentos verticais nos gráficos, pois assumem apenas poucos valores possíveis. Esse comportamento já observado anteriormente limita a observação de relações contínuas e resulta em linhas de regressão com pouca inclinação.

Entre as variáveis contínuas, X1 (Compacidade Relativa) apresenta uma tendência positiva em relação a Y2, indicando que edificações menos compactas — e, portanto, com maior área exposta — tendem a exigir maior carga de resfriamento. Já X3 (Área de Parede) também exibe relação positiva, sugerindo que superfícies externas maiores intensificam os fluxos de calor para o interior, aumentando a demanda de resfriamento.

Em contraste, as variáveis X2 (Área Superficial) e X4 (Área de Telhado) mostram tendência negativa com Y2, comportamento compatível com a matriz de correlação. Isso indica que, em determinadas combinações geométricas, aumentos nessas áreas podem estar associados à redução da necessidade de resfriamento.

As variáveis relacionadas às superfícies envidraçadas, X7 (Área de Vidro) e X8 (Distribuição da Área de Vidro), apresentam relações fracas com Y2, evidenciadas pela ampla dispersão dos pontos e linhas de regressão pouco inclinadas. Isso sugere que essas variáveis isoladas têm impacto limitado, podendo depender de interações com outros fatores, como orientação solar ou compacidade.

No geral, os scatterplots indicam que, embora algumas variáveis apresentem tendências lineares moderadas, a carga de resfriamento depende da combinação simultânea de múltiplos atributos. Esse padrão reforça a importância do uso de modelos capazes de capturar relações não-lineares e interações estruturais para representar adequadamente o comportamento térmico do sistema.

	Variável	VIF
1	X1	105.524054
2	X2	inf
3	X3	inf
4	X4	inf
5	X5	31.205474
6	X6	1.000000
7	X7	1.047508
8	X8	1.047508

Tabela 3 – Fator de Inflação da Variância (VIF)

A Tabela 3 apresenta os valores de VIF das variáveis preditoras, utilizados para avaliar o grau de multicolinearidade no conjunto de dados. Os resultados indicam a presença de forte redundância entre alguns atributos geométricos. As variáveis X2, X3 e X4 exibem VIF igual a infinito, evidenciando colinearidade quase perfeita entre si. Esse comportamento confirma que essas variáveis carregam essencialmente a mesma informação estrutural e não devem ser utilizadas simultaneamente em modelos sensíveis à multicolinearidade, como regressão linear.

A variável X1 também apresenta VIF extremamente elevado (≈ 105), reforçando sua alta correlação com outras variáveis geométricas. Já X5, com VIF em torno de 31, mostra multicolinearidade relevante, embora em menor intensidade do que as variáveis anteriores.

Por outro lado, X6, X7 e X8 apresentam valores de VIF próximos de 1, indicando baixa colinearidade e contribuição informacional mais independente para modelos preditivos. Esses atributos tendem a oferecer maior estabilidade estatística.

No geral, a análise de VIF confirma a presença de multicolinearidade estrutural entre variáveis relacionadas às dimensões físicas do edifício. Para mitigar esse problema, recomenda-se remover variáveis redundantes ou aplicar técnicas como regularização ou análise de componentes principais (PCA) antes da modelagem linear.