# Análise da Regressão Linear Múltipla para Eficiência Energética de Edifícios

Fernando Andrade Lima Tavares

3 de Junho de 2025

## 1 Introdução

Este relatório apresenta a análise do conjunto de dados de eficiência energética de edifícios e a aplicação de regressão linear múltipla para prever a carga térmica de aquecimento (*Heating Load*). A análise inclui a descrição do dataset, a formulação do problema de mínimos quadrados e a interpretação dos resultados gerados pelo modelo.

## 2 Descrição do Dataset

O dataset contém 768 amostras, cada uma representando um edifício simulado com diferentes características. Os atributos considerados são:

- X1: **Compacidade Relativa** - Mede a eficiência térmica do edifício.
- X2: **Área Superficial** - Influencia a troca de calor com o ambiente.
- X3: **Área das Paredes** - Relacionada à retenção de calor.
- X4: **Área do Telhado** - Impacta na absorção de calor.
- X5: **Altura Total** - Pode afetar a circulação de ar e a dispersão de calor.
- X6: **Orientação** - Determina a exposição ao sol.
- X7: **Área de Vidros** - Influencia o ganho/perda de calor.
- X8: **Distribuição da Área de Vidros** - Define a localização dos vidros no edifício.

Esses atributos constituem as variáveis independentes (vetor $X$), e o objetivo é prever a variável dependente $y$, correspondente à carga térmica de aquecimento (*Heating Load*).

# 3 Formulação do Problema de Mínimos Quadrados

A regressão linear múltipla busca encontrar um vetor de coeficientes $\beta$ tal que a função de custo (ou erro) quadrático médio entre as predições $\hat{y}$ e os valores reais $y$ seja minimizada. A função objetivo é definida por:

$$F(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - x_i^\top \beta)^2 = \frac{1}{2m} \|y - X\beta\|^2 \tag{1}$$

Onde:

- $m$ é o número de amostras.
- $x_i^\top$ é o vetor linha correspondente à $i$-ésima observação de atributos.
- $X$ é a matriz de dados com dimensão $m \times n$.
- $y$ é o vetor de saídas reais, com dimensão $m \times 1$.
- $\hat{y} = X\beta$ representa o vetor de predições do modelo.

Para encontrar os coeficientes $\beta$ que minimizam a função de custo, derivamos $F(\beta)$ em relação a $\beta$:

$$\nabla_\beta F(\beta) = \frac{1}{2m} \nabla_\beta (y - X\beta)^\top (y - X\beta) \tag{2}$$

$$= \frac{1}{2m} [-2X^\top (y - X\beta)] \tag{3}$$

$$= -\frac{1}{m} X^\top (y - X\beta) \tag{4}$$

Igualando o gradiente a zero, obtemos a condição de mínimo:

$$X^\top X\beta = X^\top y \tag{5}$$

Assumindo que $X^\top X$ é inversível, a solução analítica (chamada de equação normal) é dada por:

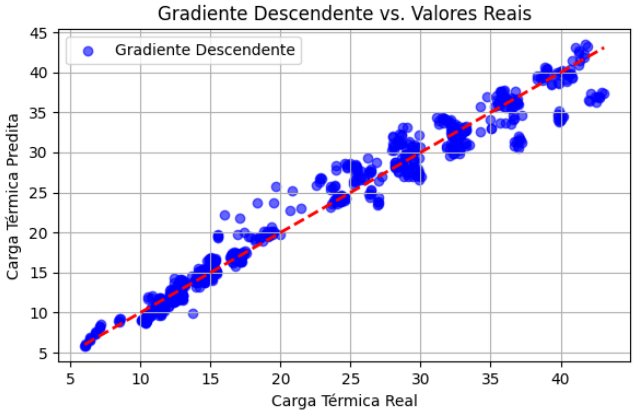
$$\beta = (X^\top X)^{-1} X^\top y \tag{6}$$

No código, optou-se por utilizar o algoritmo *SGDRegressor*, que aproxima a solução por meio do método de Gradiente Descendente Estocástico (SGD), ajustando os coeficientes $\beta$ iterativamente com base em amostras individuais. Essa abordagem é eficiente para grandes conjuntos de dados e pode ser combinada com técnicas de regularização.

# 4 Análise dos Resultados

A predição da carga térmica foi comparada aos valores reais através do gráfico de dispersão. A linha vermelha representa a relação ideal onde $\hat{y} = y$. Observa-se:

Figure 1: Gráfico de dispersão entre os valores reais da carga térmica e as predições do modelo (SGDRegressor).




O modelo apresenta uma distribuição bem ajustada em torno da reta ideal, indicando uma boa precisão. Algumas pequenas discrepâncias sugerem que características adicionais ou técnicas de regularização, como Ridge ou Lasso, podem melhorar o desempenho do modelo. Como medida quantitativa de desempenho, foi utilizada a métrica de Erro Médio Quadrático (MSE), que avalia a proximidade entre os valores reais e preditos.

# 5 Conclusão

A regressão linear múltipla demonstrou desempenho satisfatório na tarefa de predição da carga térmica de aquecimento, capturando relações lineares relevantes entre os atributos.