Decision Tree Regression: Tutorial

Aplicação em vendas de sorvetes em João Pessoa (PB)

Resumo

Este documento apresenta, em linguagem acadêmica e formal, a construção de uma *Árvore de Decisão Regressora* aplicada à previsão de **vendas de sorvetes**, tendo como variáveis explicativas a **Temperatura** (quente, ameno, frio) e o indicador **Domingo** (sim/não). A impureza é mensurada pelo **desvio-padrão** (σ), critério que reflete a dispersão dos valores da variável dependente. São discutidos os cálculos, a árvore final (ilustrada com figura), bem como as implicações gerenciais e a comparação com modelos de regressão tradicionais. *Nota didática ☺:* tanto o uso do desvio-padrão como da variância levam à mesma ordenação dos splits.

1. Introdução

Árvores de decisão são modelos consagrados de aprendizado de máquina que permitem segmentar observações em grupos homogêneos. Quando aplicadas à previsão de variáveis contínuas, têm-se as chamadas **árvores de regressão**. Ao contrário das árvores de classificação (que utilizam índices como Gini ou entropia), nas árvores de regressão a medida de impureza é associada à variabilidade dos valores numéricos da variável dependente.

2. Critério de Impureza

Sejam $y_1, \ldots, y_n$ os valores da variável alvo em um nó S. A impureza é definida pelo **desvio-padrão**:

$$\sigma(S) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

onde $\bar{y}$ é a média dos valores em S. Dada uma divisão de S em K subconjuntos $S_1, \ldots, S_K$, define-se a impureza ponderada:

$$\sigma_{\text{pós}} = \sum_{k=1}^{K} \frac{n_k}{n}\,\sigma(S_k),$$

sendo $n_k = |S_k|$. O **ganho de pureza** é:

$$\Delta = \sigma(S) - \sigma_{\text{pós}}.$$

Quanto maior Δ, mais informativo é o split.

3. Base de Dados

A amostra contém 13 dias de vendas, categorizados por temperatura e pelo fato de ser ou não domingo.

Tabela 1: Amostra de dados de vendas de sorvetes

Dia	Temperatura	Domingo	Vendas
1	quente	sim	286
2	frio	não	147
3	ameno	não	169
4	frio	sim	172
5	ameno	não	176
6	quente	não	253
7	quente	não	238
8	frio	não	151
9	frio	sim	168
10	quente	não	264
11	ameno	sim	207
12	quente	sim	309
13	quente	não	245

4. Impureza Inicial

No nó raiz:

$$\bar{y}_{\text{raiz}} \approx 221{,}9, \qquad \sigma_{\text{raiz}} \approx 52{,}35.$$

5. Divisões Possíveis

5.1. Split por Temperatura

Grupo	n	$\bar{y}$	σ
quente	6	265,8	24,63
frio	4	159,5	10,69
ameno	3	184,0	16,51

$$\sigma_{\text{pós}} \approx 18{,}47, \quad \Delta \approx 33{,}88.$$

5.2. Split por Domingo

Grupo	n	$\bar{y}$	σ
sim	5	228,4	58,48
não	8	218,6	45,95

$$\sigma_{\text{pós}} \approx 50{,}77, \quad \Delta \approx 1{,}58.$$

Conclusão: o primeiro split deve ser feito por **Temperatura**.

6. Refinamentos

- **Quente:** subdivisão por Domingo gera médias de 297,5 (sim) e 250,0 (não).
- **Frio:** domingos $\hat{y} = 170$; dias comuns $\hat{y} = 149$.
- **Ameno:** domingo isolado com 207; demais dias em torno de 172,5.

7. Árvore Final

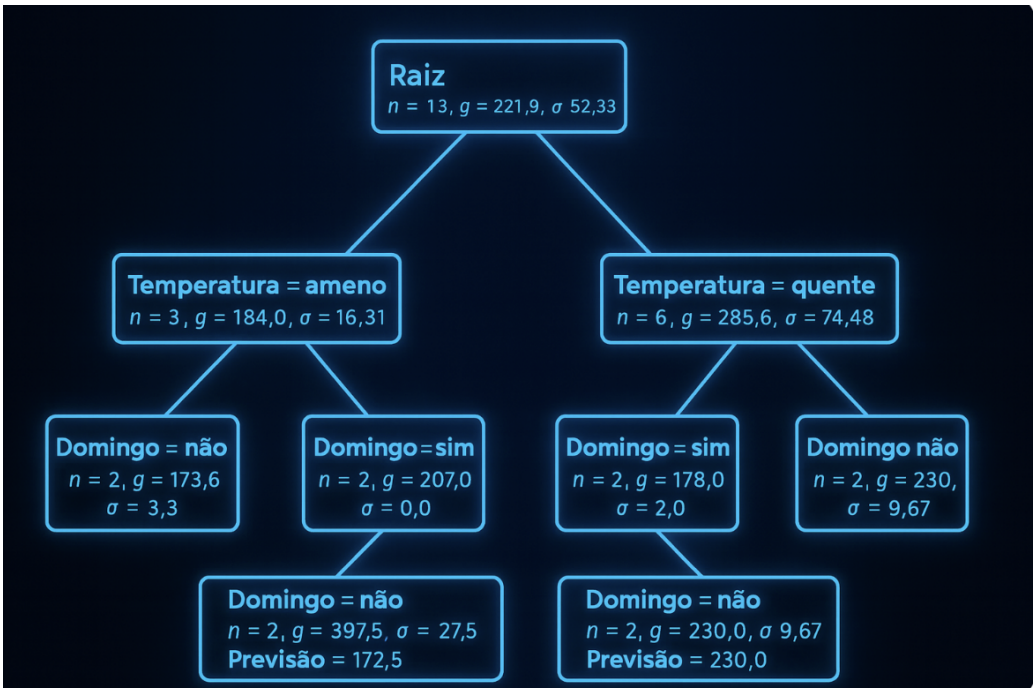

Figura 1: Árvore de Decisão Regressora final (Temperatura e Domingo).

8. Discussão Acadêmica

O critério baseado em σ demonstrou clara capacidade de reduzir a incerteza preditiva. Os resultados evidenciam:

- o papel central da **Temperatura** na explicação das vendas ($\Delta \approx 33{,}9$);
- efeitos condicionais do **Domingo**, mais relevantes em dias quentes;
- a utilidade de modelos interpretáveis para decisões práticas em marketing e planejamento de estoques.

9. Comparação com Regressão Linear

A regressão linear requer especificação funcional e não capta facilmente interações não lineares. Já a árvore segmenta os dados de forma adaptativa, oferecendo regras claras de decisão. A transparência é um diferencial importante em contextos empresariais, embora o controle de complexidade (poda) seja necessário para evitar sobreajuste.

10. Conclusão

O uso do desvio-padrão como métrica de impureza permite construir árvores de regressão intuitivas, interpretáveis e úteis em aplicações gerenciais. No caso estudado, a segmentação revelou padrões consistentes entre clima, dia da semana e vendas de sorvete, com valor agregado para a tomada de decisão.