Decision Tree Regression: Tutorial

Análise com dados de vendas de sorvetes em João Pessoa (PB)

Resumo

Este documento apresenta a construção formal de uma *Árvore de Decisão Regressora* aplicada à previsão de **vendas de sorvetes** com base em duas variáveis explicativas: **Temperatura** e **Domingo**. A medida de impureza adotada é o **desvio-padrão**, que quantifica a dispersão dos valores em cada nó. O estudo descreve o cálculo da impureza inicial, a avaliação dos possíveis splits e a montagem da árvore final. O resultado é interpretado de forma gerencial, evidenciando as vantagens do modelo em termos de interpretabilidade e segmentação de mercado. *Nota didática ☺:* a utilização de σ ou σ^2 (variância) leva à mesma ordenação de splits.

1. Conceito de Árvores de Regressão

Uma árvore de regressão é um modelo que segmenta os dados em regiões homogêneas quanto à variável dependente Y. Em cada nó, o critério de divisão busca minimizar a variabilidade dos valores de Y nos nós filhos. A previsão em cada folha corresponde à média dos valores da variável alvo naquela região.

2. Medida de Impureza: Desvio-Padrão

O desvio-padrão (σ) expressa a dispersão dos valores em torno da média. Numa árvore de regressão:

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

onde Δ representa o ganho de pureza. Quanto maior Δ, maior a redução da incerteza na previsão.

3. Base Empírica

Foram coletados 13 dias de vendas (Y) com duas variáveis explicativas:

- **Temperatura**: quente, ameno ou frio;
- **Domingo**: sim ou não.

4. Impureza no Nó Raiz

Considerando os 13 dias:

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

Tabela 1: Observações de Temperatura, Domingo e Vendas

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

5. Avaliação de Splits

5.1. 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. 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: Temperatura é a variável de maior poder explicativo no nó raiz.

6. Refinamentos

6.1. Temperatura = quente

σ cai de 24,63 para $\approx 10,28$ ao dividir por Domingo (sim: $\hat{y} \approx 297{,}5$; não: $\hat{y} \approx 250{,}0$).

6.2. Temperatura = frio

Domingo distingue vendas médias de 170 (sim) e 149 (não), ambas com $\sigma = 2$.

6.3. Temperatura = ameno

Divisão por Domingo gera folhas puras ou quase puras: sim ($\hat{y} = 207$); não ($\hat{y} = 172{,}5$).

7. Árvore Final

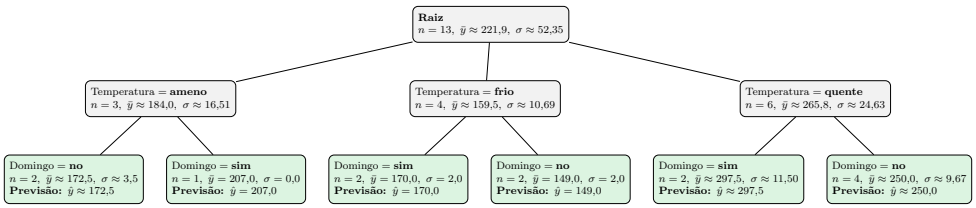

Figura 1: Árvore de Decisão Regressora final (Temperatura e Domingo).

8. Discussão Acadêmica

Os resultados confirmam que:

- **Temperatura** é o principal fator, responsável pela maior redução da dispersão ($\Delta \approx 33{,}9$);
- **Domingo** apresenta efeitos condicionais: relevante em dias quentes, marginal em dias frios;
- o modelo fornece **regras interpretáveis**, facilitando comunicação com gestores.

9. Comparação com Regressão Linear

Enquanto a regressão linear impõe forma funcional única, a árvore adapta-se a interações não lineares e categorias múltiplas, com regras intuitivas. É particularmente vantajosa em contextos de decisão gerencial, onde transparência e explicabilidade são fundamentais.

10. Conclusão

O critério baseado em σ permite identificar divisões que reduzem significativamente a incerteza preditiva. A árvore obtida mostra-se robusta, interpretável e aplicável à gestão de vendas de sorvetes, ilustrando o potencial das árvores de regressão em contextos práticos de negócios.