

Árvore de decisão para regressão

Temperatura	Domingo	Vendas
quente	sim	286
frio	não	147
ameno	não	169
frio	sim	172
ameno	não	176
quente	não	253
quente	não	238
frio	não	151
frio	sim	168
quente	não	264
ameno	sim	207
quente	sim	309
quente	não	245

Dado o dataset acima, queremos criar uma árvore de decisão para esse problema de regressão, para prever o número de vendas de sorvete baseado na temperatura do dia e se é domingo ou não.

Diferente da classificação onde apenas separamos os nós pelas features separando a quantidade de resultados, na regressão devemos considerar um intervalo. Portanto usamos a seguinte fórmula:

$$\sum_{c \in X} P(c) \cdot S(c)$$

A primeira coisa que devemos fazer então é calcular o desvio padrão da variável target "vendas":

Svendas = 52,35

Depois disso calculamos o desvio padrão de cada variável. Começando pela temperatura:

TEMP.	S	AMOSTRAS
quente	24,65	6
frio	10,69	4
ameno	16,51	3

O desvio padrão é calculado pelas amostras, por exemplo, temos 6 vendas em dias quentes, portanto calculamos o desvio padrão das 6 vendas realizadas.

$$S_{TEMP.} = 24,65 \cdot \frac{6}{13} + 10,69 \cdot \frac{4}{13} + 16,51 \cdot \frac{3}{13}$$

Acima podemos ver então que o desvio padrão da variável temperatura é dado pelo somatório dos pesos de cada categoria multiplicado pelo seu respectivo desvio padrão. Exemplo:

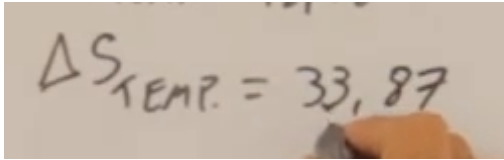
$$\text{Quente} = 24,65 \cdot 6 / 13$$

São 6 amostras de vendas quando está quente de um total de 13 vendas

Basta somar isso e o resultado será:

$$S_{TEMP.} = 24,65 \cdot \frac{6}{13} + 10,69 \cdot \frac{4}{13} + 16,51 \cdot \frac{3}{13}$$
$$S_{TEMP.} = 18,48$$

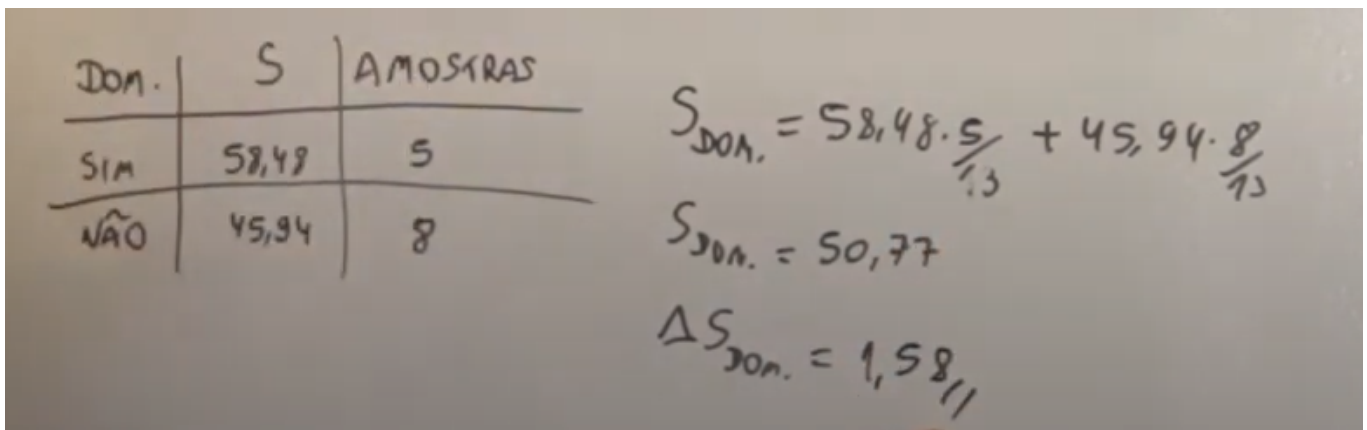
Por fim devemos calcular a diferença entre o desvio padrão das vendas totais com o desvio padrão da variável temperatura: $52,35 - 18,48$


$$\Delta S_{TEMP} = 33,87$$

Isso é chamado de redução de desvio padrão. Sabemos que o desvio padrão é o valor que indica o quanto os valores estão distantes da média, se for muito alto, mais longe da média, se for baixo mais próximo da média.

Podemos observar que na variável temperatura o desvio padrão de quando o dia está quente é maior do que quando está frio, ou seja, ela explicaria muito bem a divisão dos dados, uma vez que queremos que o desvio padrão seja o menor possível em relação ao desvio padrão da variável target, uma vez que explica melhor a relação das separações de cada variável.

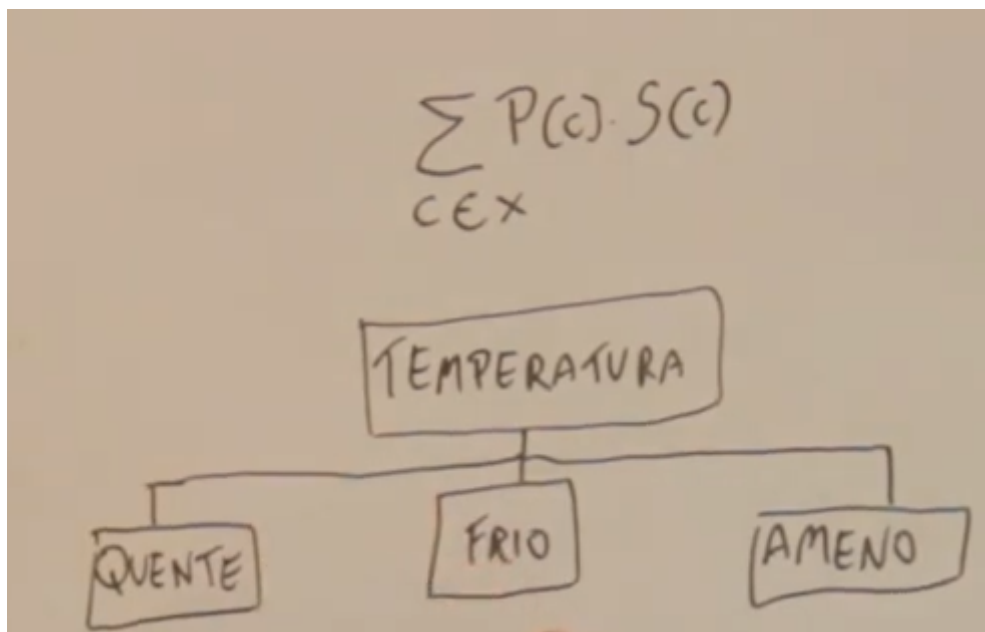
Agora calculando tudo novamente para a variável domingo:



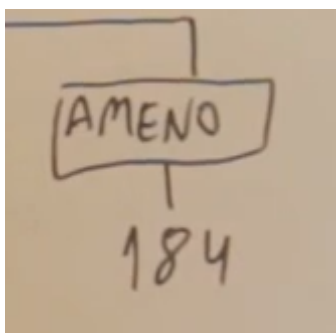
DOM.	S	AMOSTRAS
SIM	58,48	5
NÃO	45,94	8

$$S_{DOM.} = 58,48 \cdot \frac{5}{13} + 45,94 \cdot \frac{8}{13}$$
$$S_{DOM.} = 50,77$$
$$\Delta S_{DOM.} = 1,5811$$

Portanto a maior redução do desvio padrão foi a variável temperatura, ou seja, ela explica melhor os dados e usaremos ela para começar a árvore de decisão.



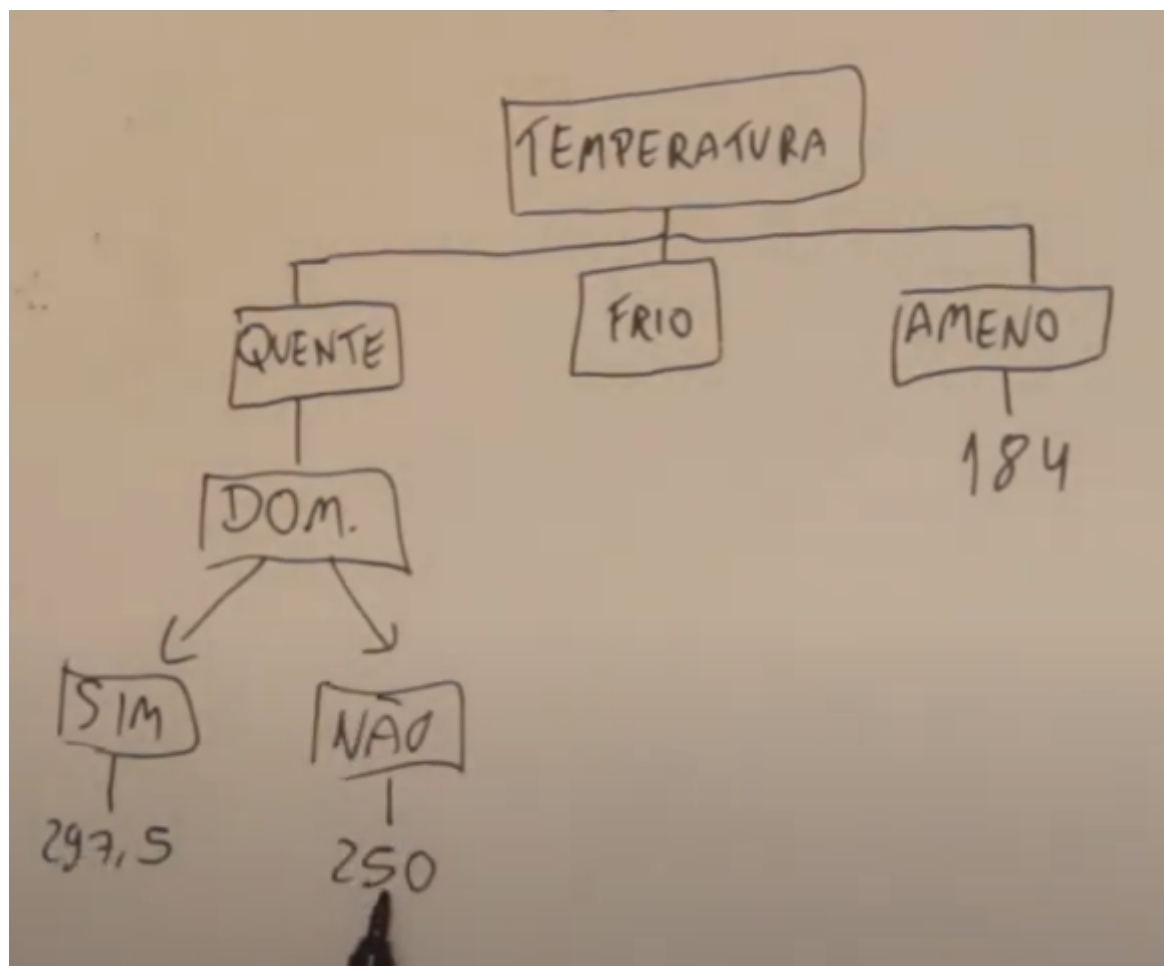
Considerando um mínimo de 4 amostras e analisando nó filho Ameno, temos somente 3 amostras de vendas quando a temperatura pé amena, ou seja, não precisamos analisar a variável domingo para esse caso, basta tirar uma média de vendas quando a temperatura é amena, resultando em 184:



Agora analisando o nó quente temos que obter todos os valores da tabela original:

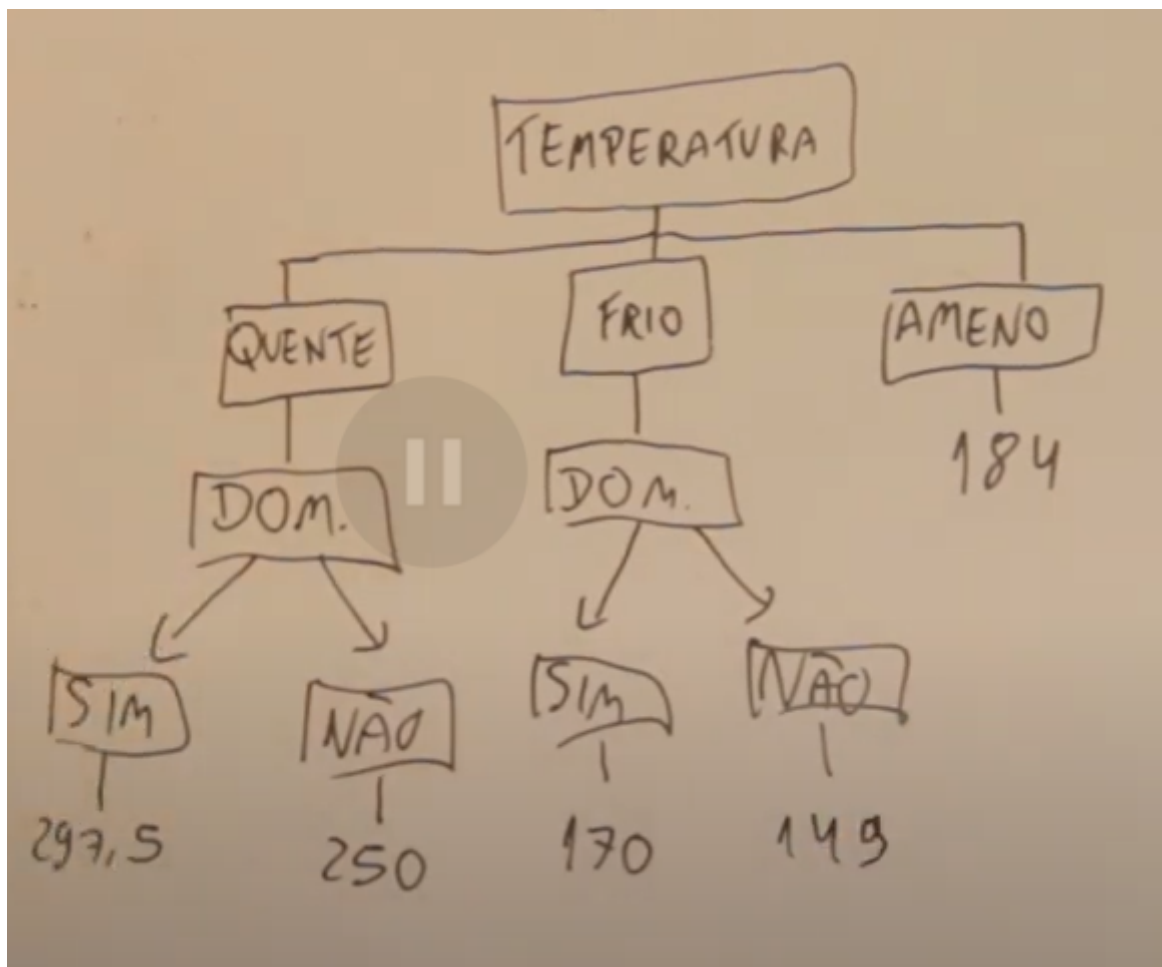
Temperatura	Domingo	Vendas
quente	sim	286
quente	não	253
quente	não	238
quente	não	264
quente	sim	309
quente	não	245

Se tivéssemos mais de uma variável além da temperatura, teríamos que calcular a redução de desvio padrão para cada uma delas e definir como filho do nó quente, porém só temos a variável domingo portanto fica dessa forma:



Basta calcular a média de vendas quando a temperatura é quente e é domingo ou não é domingo.

Agora a mesma coisa para variável frio:



Essa é a forma de montar uma árvore de decisão para um problema de regressão.