

Árvores de Decisão

THE JOB

strategic
planning
tool
→



©2008 Harry B. Price. Distributed by King Features Syndicate, Inc.

PLANT THE
DECISION TREE
THERE.



LANDSCAPING
AT THE
INDUSTRIAL
PARK.

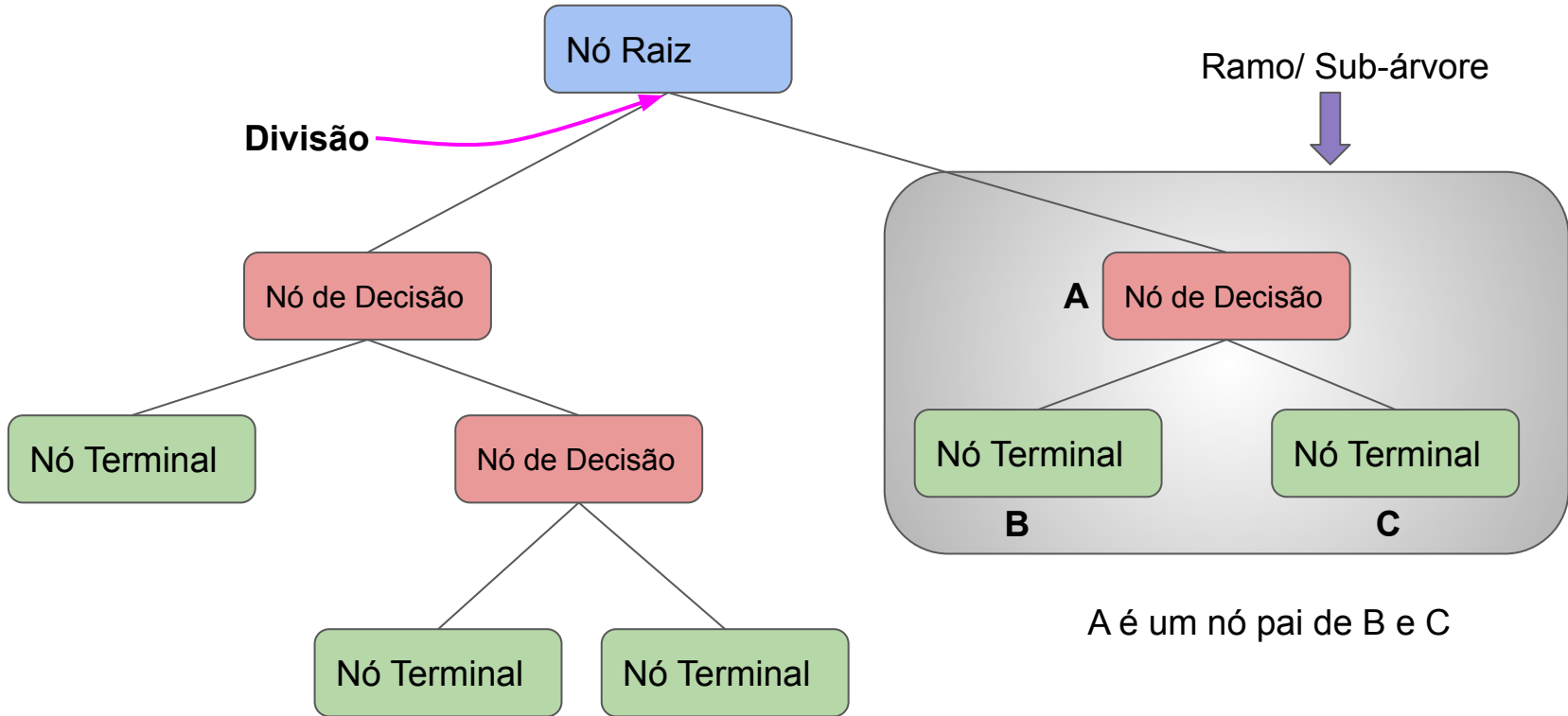


Hi May B. Price 5-28 rhymeswithorange.com

Definição

- Uma árvore de decisão é uma representação de um fluxo de dados e, como o nome indica, baseia-se em uma estrutura de árvore binária.
- Cada nó interno desta árvore representa uma característica (ou atributo) do padrão sendo analisado, cada ramo representa uma regra de decisão e cada nó folha representa a saída do processo.
- A árvore aprende a particionar seus ramos baseando-se nos valores dos atributos.
- Esta partição é realizada de modo recursivo.
- Usa-se este fluxo representado em processos de tomada de decisão e classificações.

Terminologia



Terminologia

Nó Raiz : indica toda a população (amostra) e a partir dele a árvore divide-se em dois ou mais conjuntos homogêneos

Divisão : é o processo de particionar um nó em dois ou mais subnós.

Nó de Decisão: quando um subnó se divide em mais subnós, ele é chamado de nó de decisão

Folha/ Nó Terminal: são nós que não se dividem

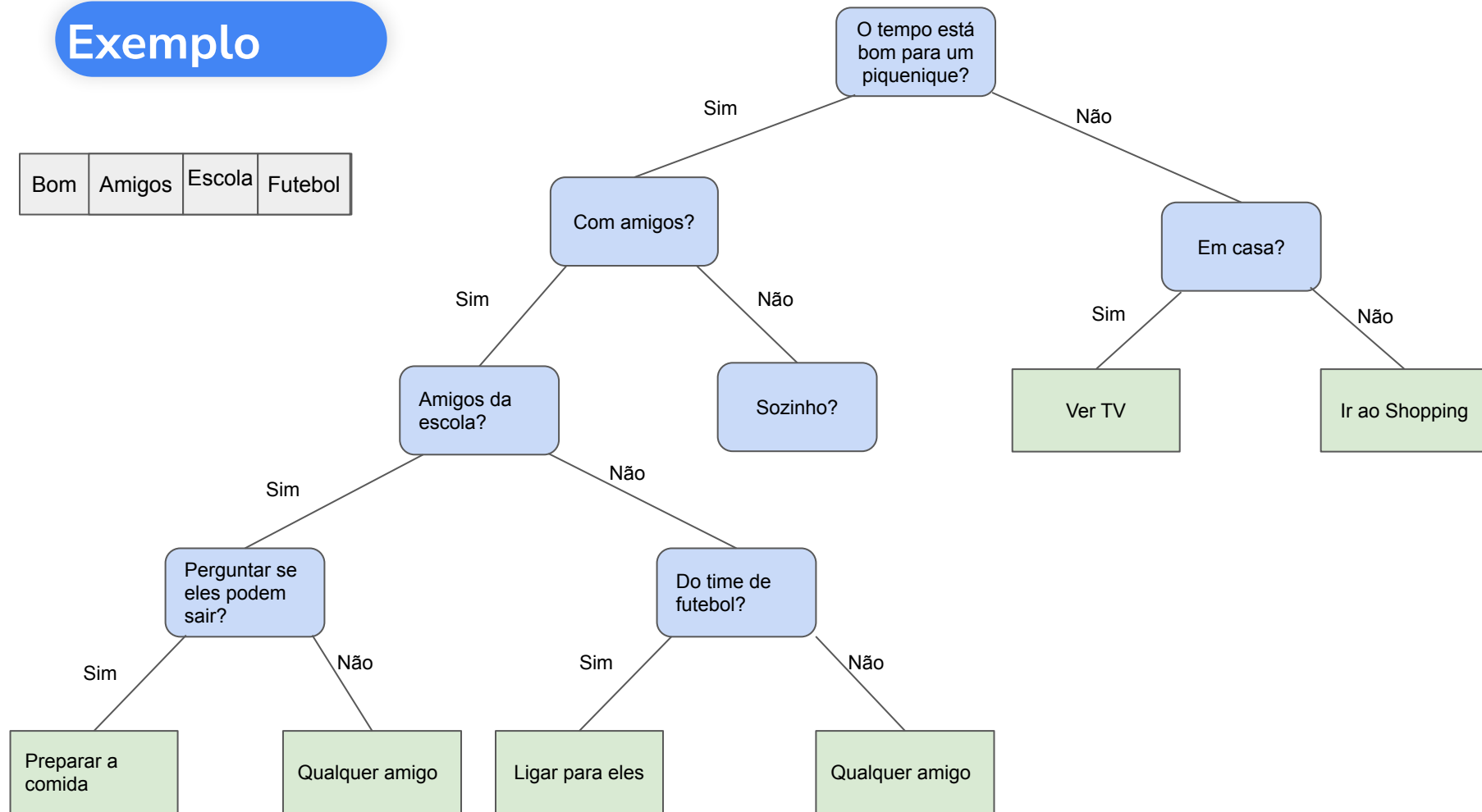
Poda: Acontece quando removem-se subnós de um nó de decisão. Pode-se dizer que é o contrário da divisão.

Ramo/ Sub-árvore: é uma subseção da árvore

Nós pais e nós filhos: um nó que se divide em subnós é chamado de nó pai. Subnós são filhos de nós pai.

Exemplo

Bom	Amigos	Escola	Futebol
-----	--------	--------	---------



Vantagens

- Uma árvore de decisão é fácil de entender e de interpretar. É um modelo do tipo “caixa-branca”, ou seja, explicita sua lógica interna de decisão.
- Sua complexidade é função do número de vetores (registros) de entrada e do número de atributos (características) dos dados.
- É um método não-paramétrico, ou seja, não depende da distribuição de probabilidade dos dados de entrada
- Podem aliar alta dimensionalidade dos dados com uma boa acurácia

Vantagens

- Podem-se incluir opiniões de especialistas e preferências, bem como dados concretos
- Podem ser usadas juntamente com outras técnicas de decisão
- Novos cenários podem ser facilmente adicionados

Desvantagens

- Se uma árvore de decisão for usada com variáveis categóricas multinível, aquelas variáveis com mais níveis possuirão maior ganho de informação
- Os cálculos podem rapidamente tornarem-se muito complexos, embora isto geralmente apenas seja um problema se ela estiver sendo criada manualmente

Exemplo de dataset

Cor	Forma	Possui caroço?	Diâmetro (mm)	Classe
Red	Esférico	Sim	120	Maçã
Red	Esférico	Sim	30	Cereja
Red	Cônico	Não	30	Morango
Red	Esférico	Sim	50	Ameixa
Verde	Esférico	Sim	120	Laranja
Verde	Esférico	Sim	120	Maçã
Verde	Esférico	Sim	500	Melancia
Verde	Esférico	Sim	20	Abacate
Amarelo	Esférico	Sim	50	Pêssego
Amarelo	Esférico	Sim	120	Laranja
Amarelo	Esférico	Sim	200	Melão

Entropia

Mede a quantidade de desordem de um grupo → o quanto “misturado” está o grupo ?

Uma função de entropia calcula a frequência de cada item (quantidade de ocorrências dividida pelo total de linhas)

$$p(i) = \text{frequência}(\text{resultado}) = \text{contagem}(\text{resultado}) / \text{contagem}(\text{total de linhas})$$

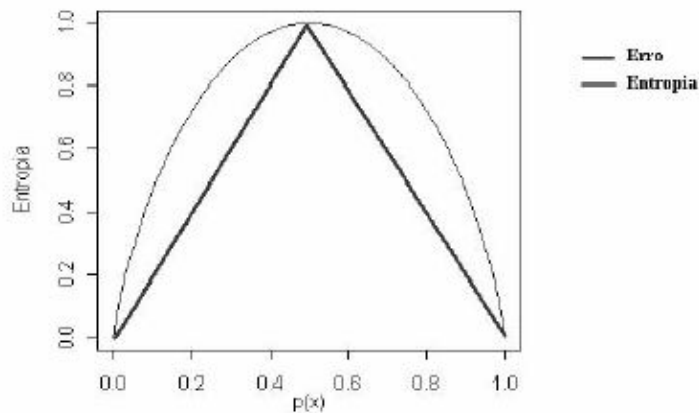
$$\text{entropia} = \sum_{i=1}^c - (p(i) \cdot \log_2(p(i)))$$

A entropia aumenta mais lentamente que a impureza de Gini → ela penaliza grupos mais fortemente misturados

Entropia

Entropia é a medida da aleatoriedade na informação sendo processada.

Quanto **mais alta a entropia**, **mais difícil é tirar alguma conclusão** dessa informação. Por exemplo, jogar uma moeda é uma ação que produz uma informação aleatória, sua entropia pode ser calculada:



Índice de Gini

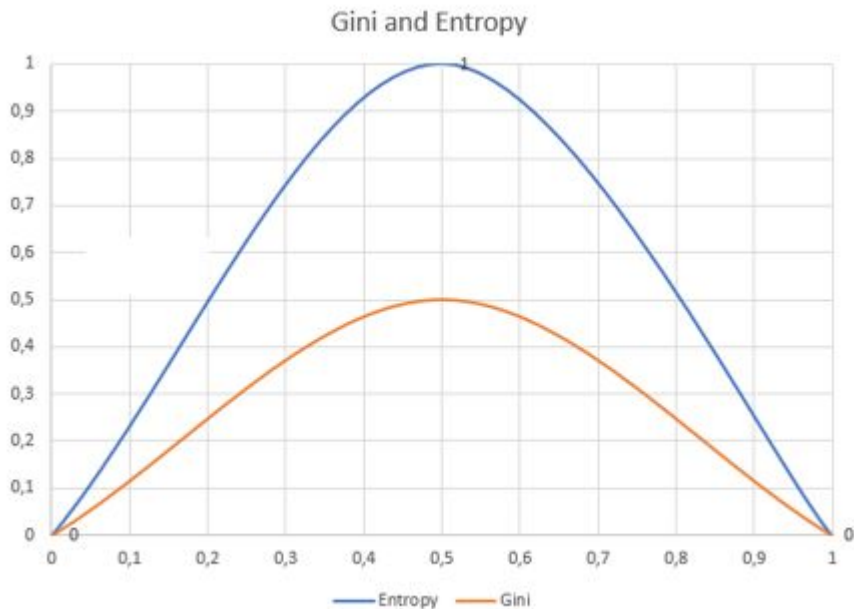
- É o erro esperado se um resultado de um grupo é aplicado aleatoriamente a um dos itens do grupo.
- Se todos os itens pertencerem à mesma classe, a suposição será sempre correta □ o erro será zero
- Se existirem quatro resultados possíveis igualmente distribuídos no grupo, haverá uma probabilidade de $\frac{3}{4}$ da suposição estar errada □ o erro será 0.75

Impureza de Gini

A impureza de Gini representa o valor de uma função de custo usada para avaliar partições em um conjunto de dados. Ela é calculada subtraindo-se de 1 (um) a soma dos quadrados das probabilidades de cada classe. É fácil de calcular e privilegia maiores partições enquanto o ganho de informação favorece menores partições com valores distintos.

$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

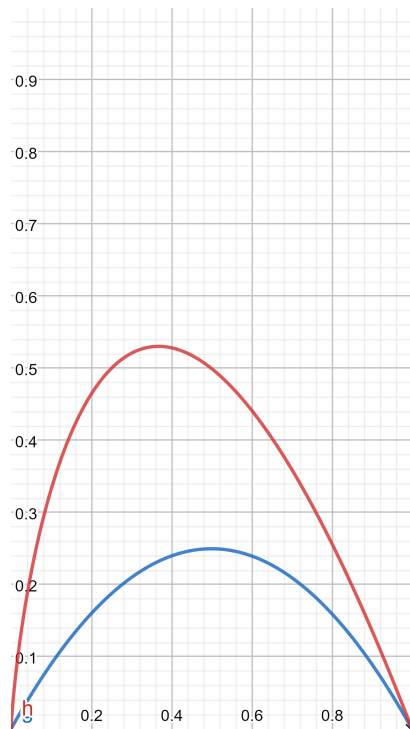
Entropia X Desigualdade



$$G = \sum_{i=1}^C p(i) * (1 - p(i))$$

$$entropia = \sum_{i=1}^c - (p(i) \cdot \log_2(p(i)))$$

Entropia X Desigualdade



$$G = \sum_{i=1}^C p(i) * (1 - p(i))$$

$$entropia = \sum_{i=1}^c - (p(i) \cdot \log_2(p(i)))$$

Ganho de Informação

Ganho de Informação (GI) é uma propriedade estatística que mede o quão bem um dado atributo divide os exemplos de treinamento de acordo com a classificação desejada. O ato de construir uma árvore de decisão consiste em encontrar um atributo que retorna o maior ganho de informação e a menor entropia.

$$\text{Ganho de Informação} = \text{Entropia}(\text{antes}) - \sum_{j=1}^K \text{Entropia}(j, \text{depois})$$

Redução da Variância

É um algoritmo usado para em variáveis objetivo contínuas (problemas de regressão). Este algoritmo usa a fórmula padrão de variância para escolher a melhor partição. A partição com a menor variância é escolhida como o critério para dividir a população

$$Variância = \frac{\sum (X - \bar{X})^2}{n}$$

$\bar{X} \rightarrow$ média dos valores

$X \rightarrow$ o valor da medida

$n \rightarrow$ quantidade de valores

Poda

Prevenindo o *overfitting* (sobreajuste)

Pode ser realizada em dois casos:

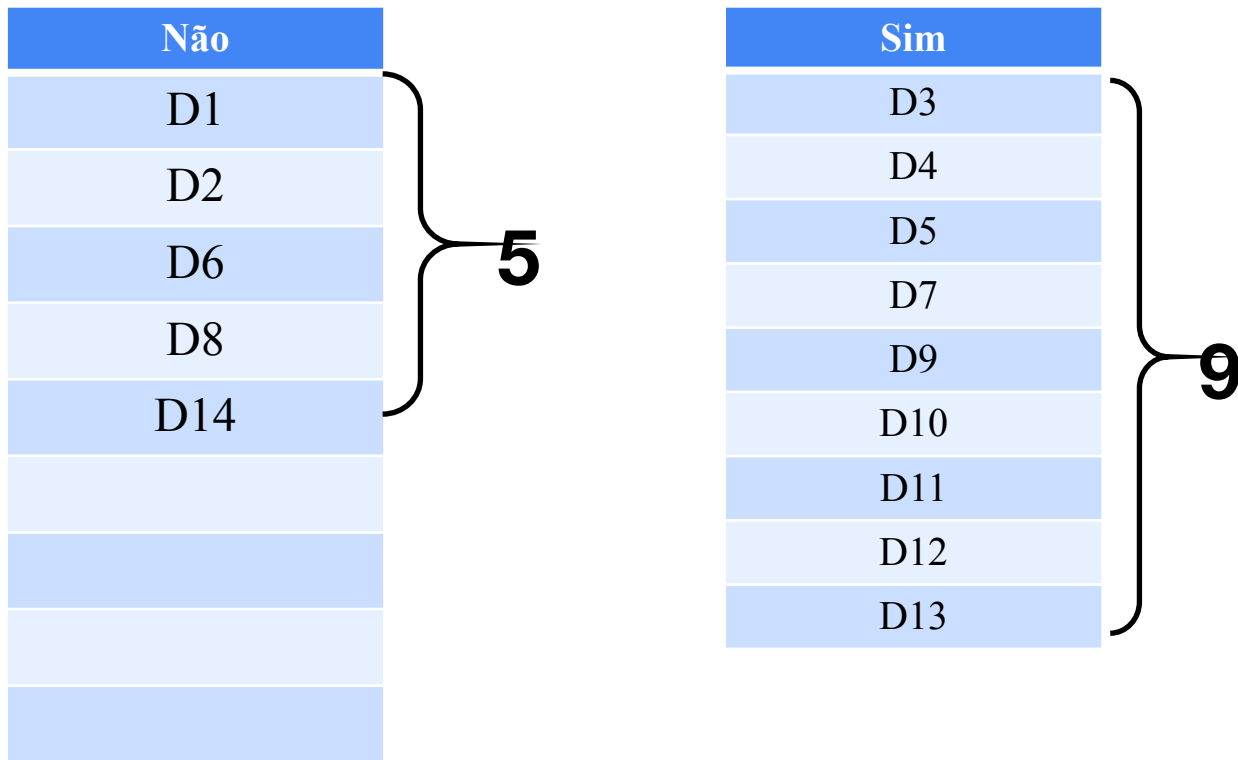
- Para controlar precocemente o crescimento da árvore (pré-poda, poda descendente)
- Para reduzir o tamanho final da árvore (pós-poda, poda ascendente)

Onde devemos criar um novo ramo?

	Tempo	Temperatura	Umidade	Vento	Jogar tênis?
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Amena	Alta	Fraco	Sim
D5	Chuvoso	Frio	Normal	Fraco	Sim
D6	Chuvoso	Frio	Normal	Forte	Não
D7	Nublado	Frio	Normal	Fraco	Sim
D8	Ensolarado	Amena	Alta	Fraco	Não
D9	Ensolarado	Frio	Normal	Fraco	Sim
D10	Chuvoso	Amena	Normal	Forte	Sim
D11	Ensolarado	Amena	Normal	Forte	Sim
D12	Nublado	Amena	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Amena	Alta	Forte	Não

Escolhendo onde ramificar

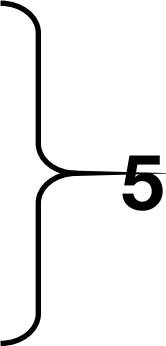
- Para cada classe, divide-se o conjunto em duas listas



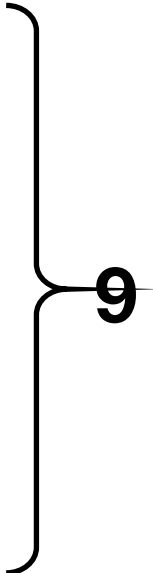
Escolhendo onde ramificar

- Para o atributo VENTO

Tempo X Não
D1
D2
D6
D8
D14



Tempo X Sim
D3
D4
D5
D7
D9
D10
D11
D12
D13

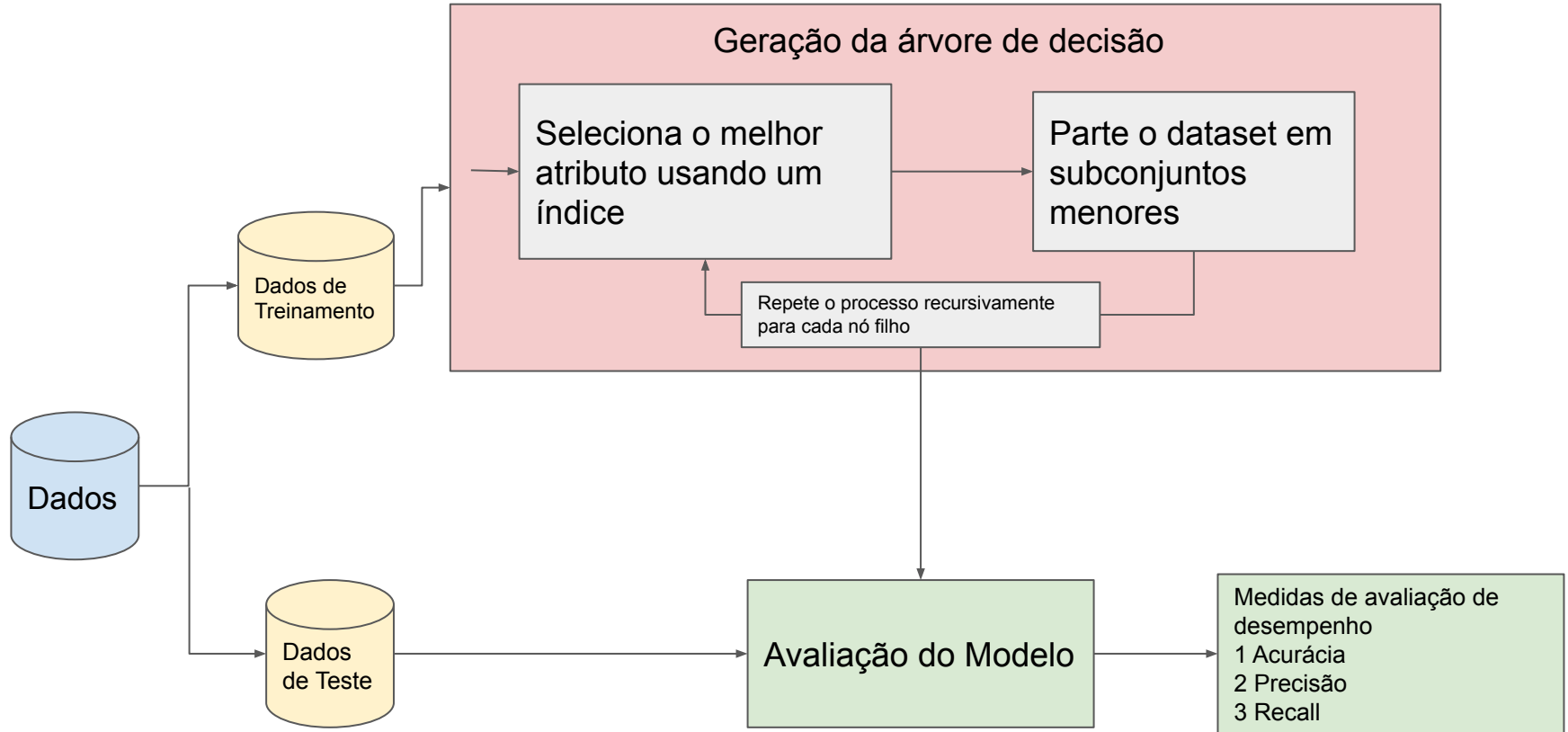


Avaliando a “mistura”

$$\begin{aligned} \text{Entropia} &= \underbrace{5/14 \cdot \log_2 5/14}_{\substack{\text{Classe} \\ \text{Não}}} - \underbrace{9/14 \cdot \log_2 9/14}_{\substack{\text{Classe} \\ \text{Sim}}} \\ &= \boxed{0,94} \end{aligned}$$

Gini = ? (calcule)

Algoritmo Fundamental



Algoritmo Fundamental

A ideia básica em qualquer algoritmo de árvores de decisão é :

1. Selecione o melhor atributo usando Medidas de Seleção de Atributos, para particionar os registros
2. Torne este atributo um nó de decisão e articione o dataset em subconjuntos menores
3. Comece a construir a árvore repetindo este processo recursivamente para cada nó filho até que uma das seguintes condições seja atingida:
 - Todas as tuplas possuam os mesmos valores de atributos
 - Não existam mais atributos restantes
 - Não existam mais registros

Algoritmos mais comuns

ID3 → Extensão do D3

C4.5 → sucessor do ID3

CART → **C**lassification **A**nd **R**egression **T**ree

CHAID → (**CH**i-square **A**utomatic **I**nteraction **D**etection) executa divisões multinível ao computar árvores de classificação

MARS → **M**ultivariate **A**daptive **R**egression **S**plines

ID3

- Inicia-se com o conjunto original S como o nó raiz
- A cada iteração do algoritmo, escolhe-se o atributo menos usado do conjunto S e calculam-se a Entropia(H) e o Ganho de Informação (IG) deste atributo
- Seleciona-se o atributo que possui o menor valor de H ou o maior valor de IG
- O conjunto S é então particionado pelo atributo selecionado de modo a gerar um subconjunto de dados
- O algoritmo continua a recorrer em cada subconjunto, levando em consideração apenas atributos ainda não selecionados

Florestas

