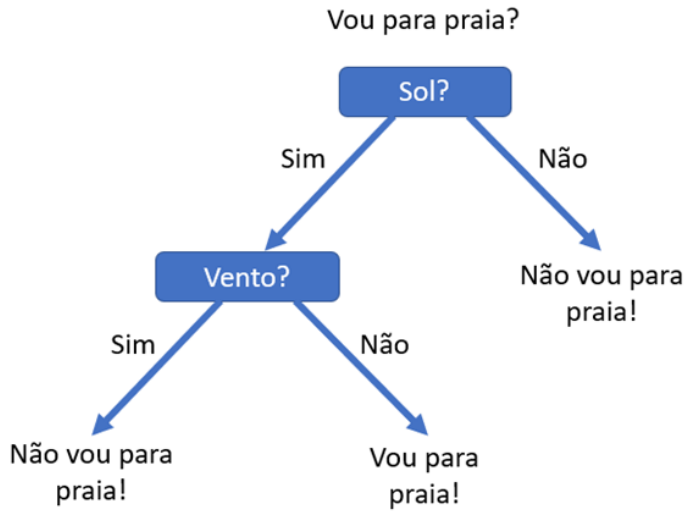
# Ciência de Dados

## WEBINAR 06 – Unidade 04

UNIVERSIDADE CANDIDO MENDES | EAD

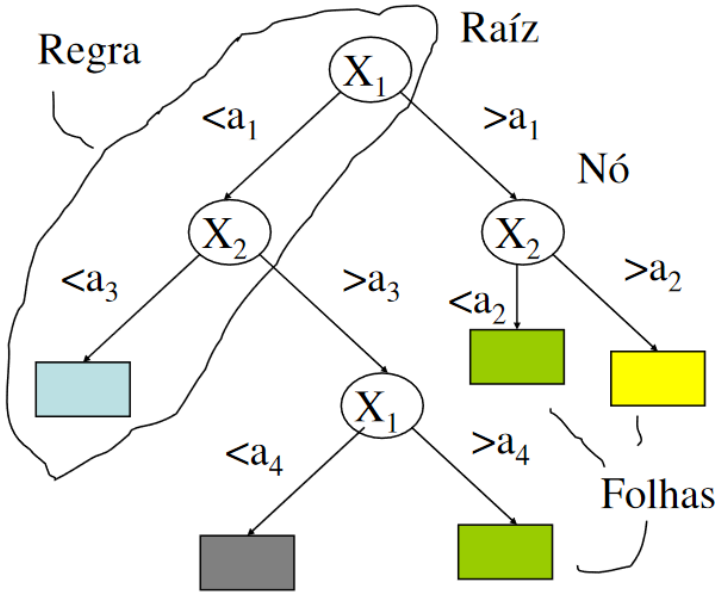
# Introdução ao Big Data

- Aprendizado supervisionado:
  - Algoritmos de classificação:
    - Árvores de decisão ou regressão:
      - Os algoritmos de árvores de decisão ou regressão utilizam a estratégia de dividir para conquistar;
        - Um problema complexo é decomposto em sub-problemas menores e mais simples;
        - Recursivamente a mesma estratégia é aplicada a cada sub-problema.
      - As árvores de decisão ou regressão são estruturas de dados formadas por um conjunto de elementos que armazenam informações chamadas nós.
      - Em uma árvore de decisão, uma decisão é tomada através do caminhamento a partir do nó raiz até o nó folha.



# Introdução ao Big Data

- Aprendizado supervisionado:
  - Algoritmos de classificação:
    - Árvores de decisão ou regressão:
      - Cada nó de decisão contem um teste de um atributo;
      - Cada ramo descendente corresponde a um possível valor deste atributo;
      - Cada folha está associada a uma classe;
      - Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.



# Introdução ao Big Data

- Aprendizado supervisionado:
  - Algoritmos de classificação:
    - Árvores de decisão ou regressão:
      - Entropia é uma medida da aleatoriedade (impureza) de uma variável;
      - Quanto maior a entropia, maior a desordem dos dados;
      - Quanto menor, mais puro é o dado;
      - Partindo da entropia, o algoritmo confere o ganho de informação de cada variável. Aquela que apresentar maior ganho de informação será a variável do primeiro nó da árvores.
      - Podemos entender o ganho de informação como a medida de quão bem relacionados os dados das variáveis independentes ou preditoras estão com os dados da variável dependente ou rótulo
      - Ou seja, a variável independente com melhor desempenho (maior impacto na predição da variável dependente) será a escolhida para iniciar a árvore.

# Introdução ao Big Data

- Aprendizado supervisionado:
  - Algoritmos de classificação:
    - Árvores de decisão ou regressão:

$$entropia(X) = -\sum_i p_i \log_2 p_i$$

Suponha que S é uma coleção de 14 exemplos, incluindo 9 positivos e 5 negativos

– Notação: [9+,5-]

A entropia de S em relação a esta classificação booleana é dada por:

$$Entropy\ ([9+,5-]) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$$
$$= 0.940$$

# Introdução ao Big Data

- Aprendizado supervisionado:
  - Algoritmos de classificação:
    - Árvores de decisão ou regressão:
      - Ganho de informação: mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo;

$$ganho(Exs, Atri) = entropia(Exs) - \sum_{v} \frac{\#Exs_v}{\#Exs} entropia(Exs_v)$$

- Informação da Classe:
  - p(sim) = 9/14
  - p(não) = 5/14
  - Ent(joga) = - 9/14 $\log_2$ 9/14
  - 5/14 $\log_2$ 5/14 = 0.940
- Informação nas partições:
  - p(sim | tempo=sol) = 2/5
  - p(não | tempo=sol) = 3/5

- Informação nas partições:
  - Ent(joga | tempo=sol)
  - = -2/5 $\log_2$ 2/5 –3/5 $\log_2$ 3/5 = 0.971
  - Ent(joga | tempo=nublado) = 0.0
  - Ent(joga | tempo=chuva) = 0.971
  - Info(tempo) = 5/14*0.971 + 4/14*0+5/14*0.971= 0.693
- Ganho de Informação obtida neste atributo:
  - Ganho(tempo) = Ent(joga)-Info(tempo)
  - Ganho(tempo) = 0.940 – 0.693 = 0.247

Ganho

| | Sol | Nublado | Chuva |
|---|---|---|---|
| Sim | 2 | 4 | 3 |
| Não | 3 | 0 | 2 |

# Introdução ao Big Data

- Aprendizado supervisionado:
  - Algoritmos de classificação:
    - Árvores de decisão ou regressão:
      - Exemplo Excel.
      - Exemplo Python.

UNIVERSIDADE
CANDIDO
MENDES

EAD