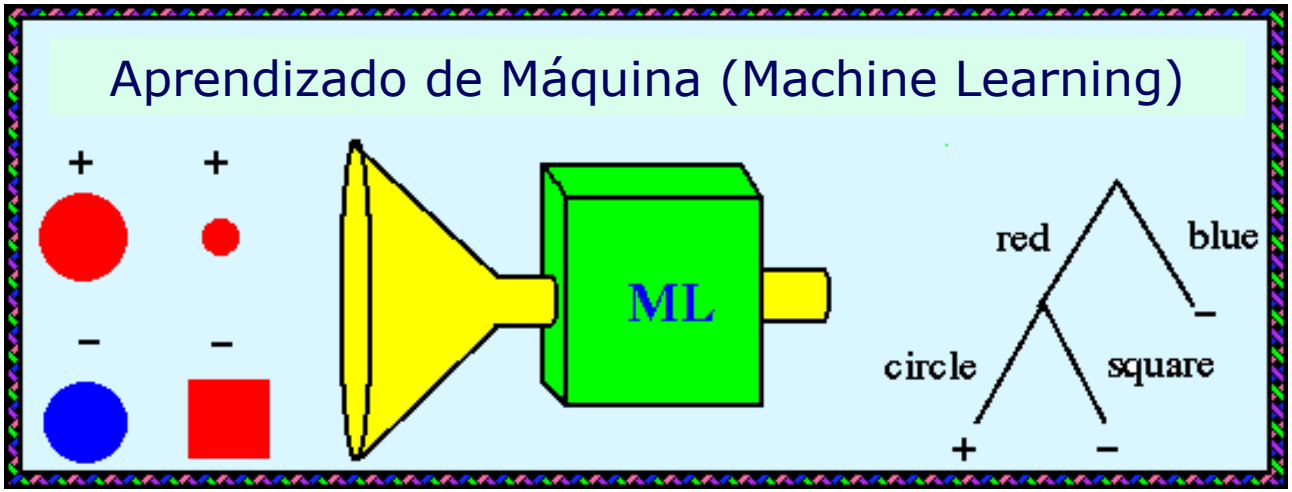
Introdução à Aprendizagem de Máquina

Stanley R. M. Oliveira

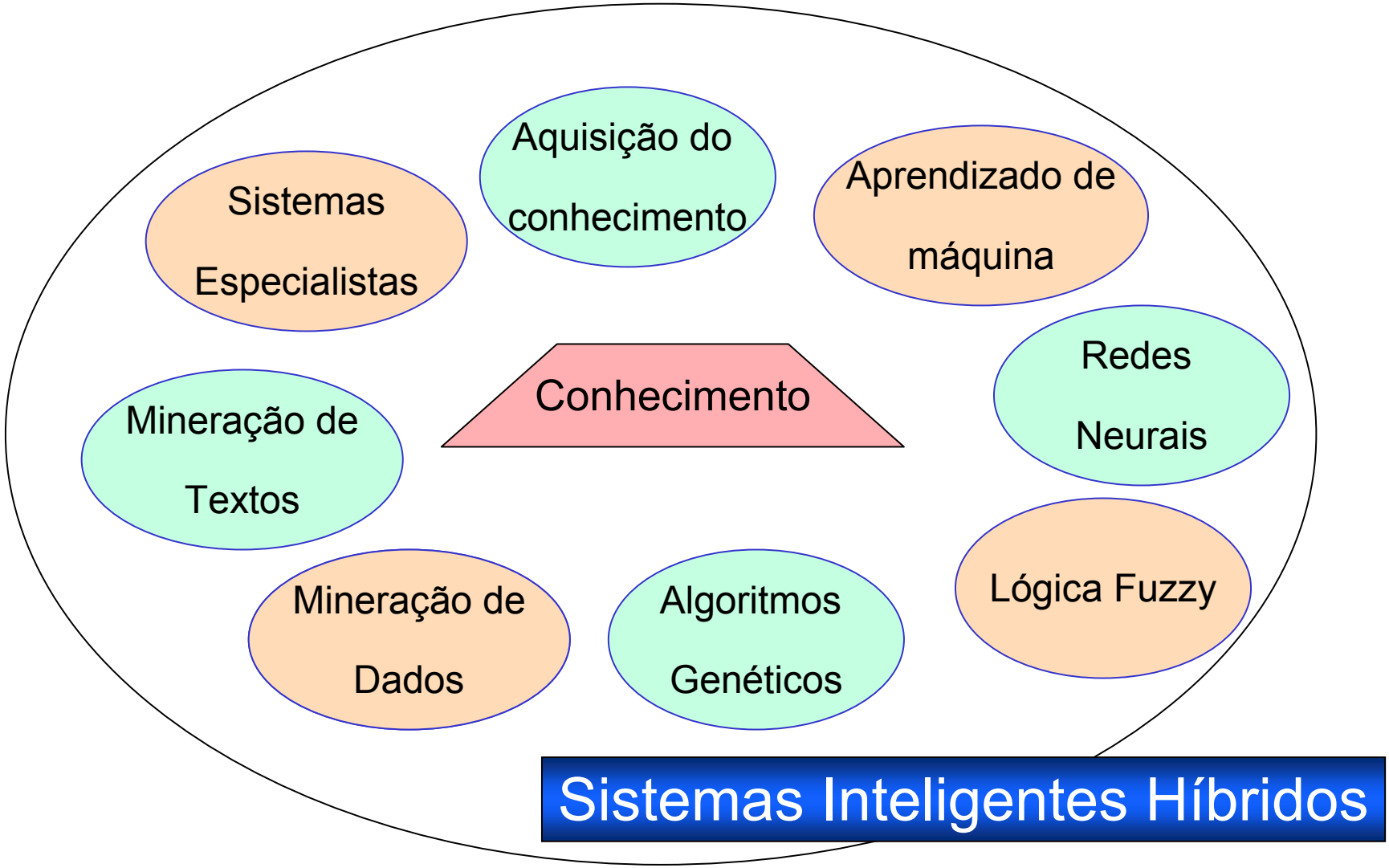


Resumo da Aula

- Sistemas **Híbridos Inteligentes**.
- **Aprendizado de Máquina**:
  - Conceitos.
  - Hierarquiza do aprendizado.
  - Paradigmas.
- **Classificação de Dados**:
  - Processo de classificação.
  - Características de um bom classificador.
  - Principais métodos de classificação.
  - Principais algoritmos existentes.
- **Árvores de Decisão**:
  - Conceitos básicos.
  - Algoritmos mais conhecidos.
  - Mecanismos de poda.
  - Escolha do atributo "**split**".

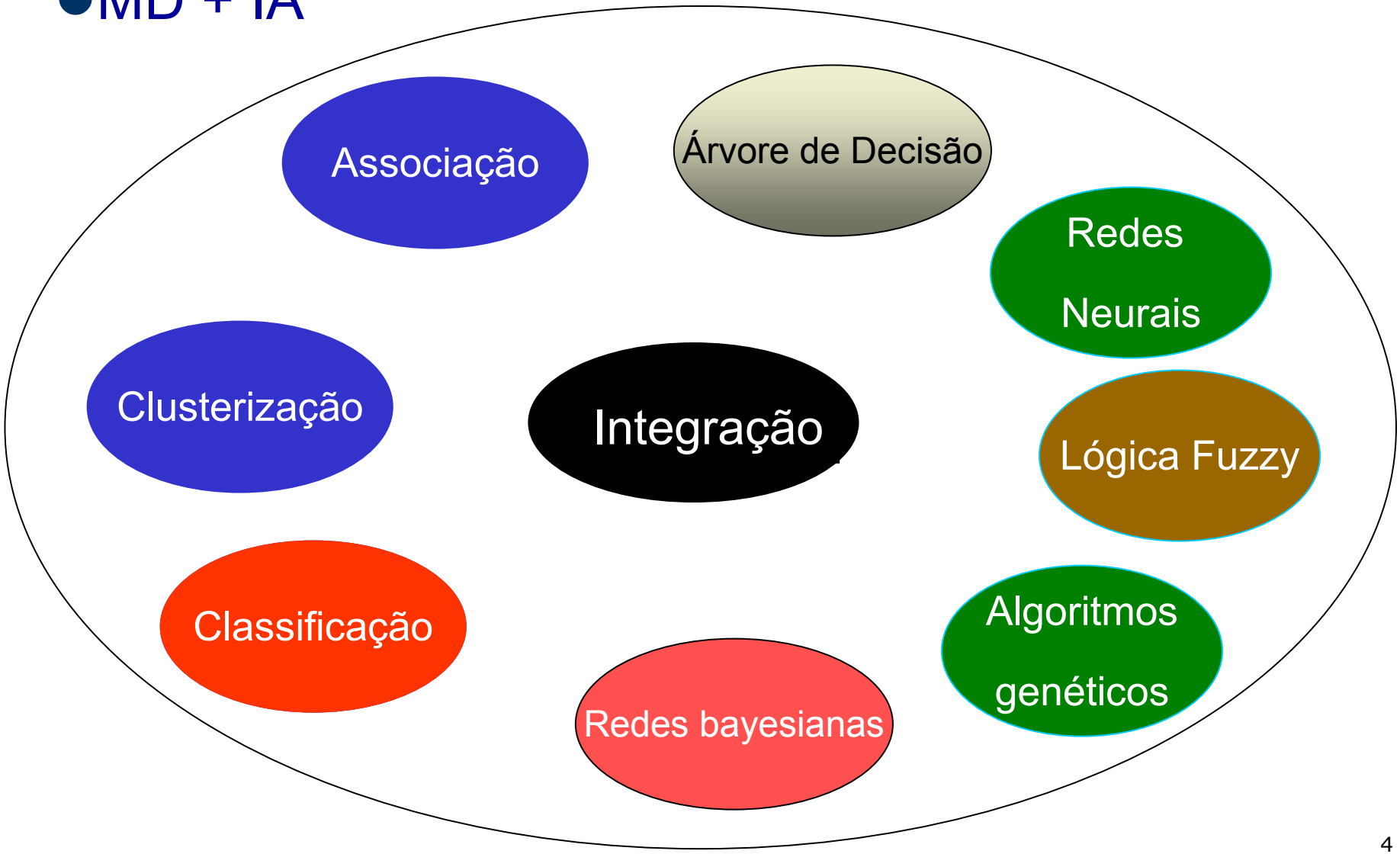
Introdução

- Técnicas-chave em Inteligência Artificial (**IA**)



Introdução

- MD + IA



Aprendizado de Máquina

- **Conceitos de aprendizado de máquina**

  Sistemas de IA

  - Armazenar conhecimento ⇒ **Representação**
  - Aplicar conhecimento para resolver problemas
    - **Raciocínio** (**mecanismo de inferência**)
  - Adquirir novos conhecimentos ⇒ **Aprendizagem**

Aprendizado de Máquina ...

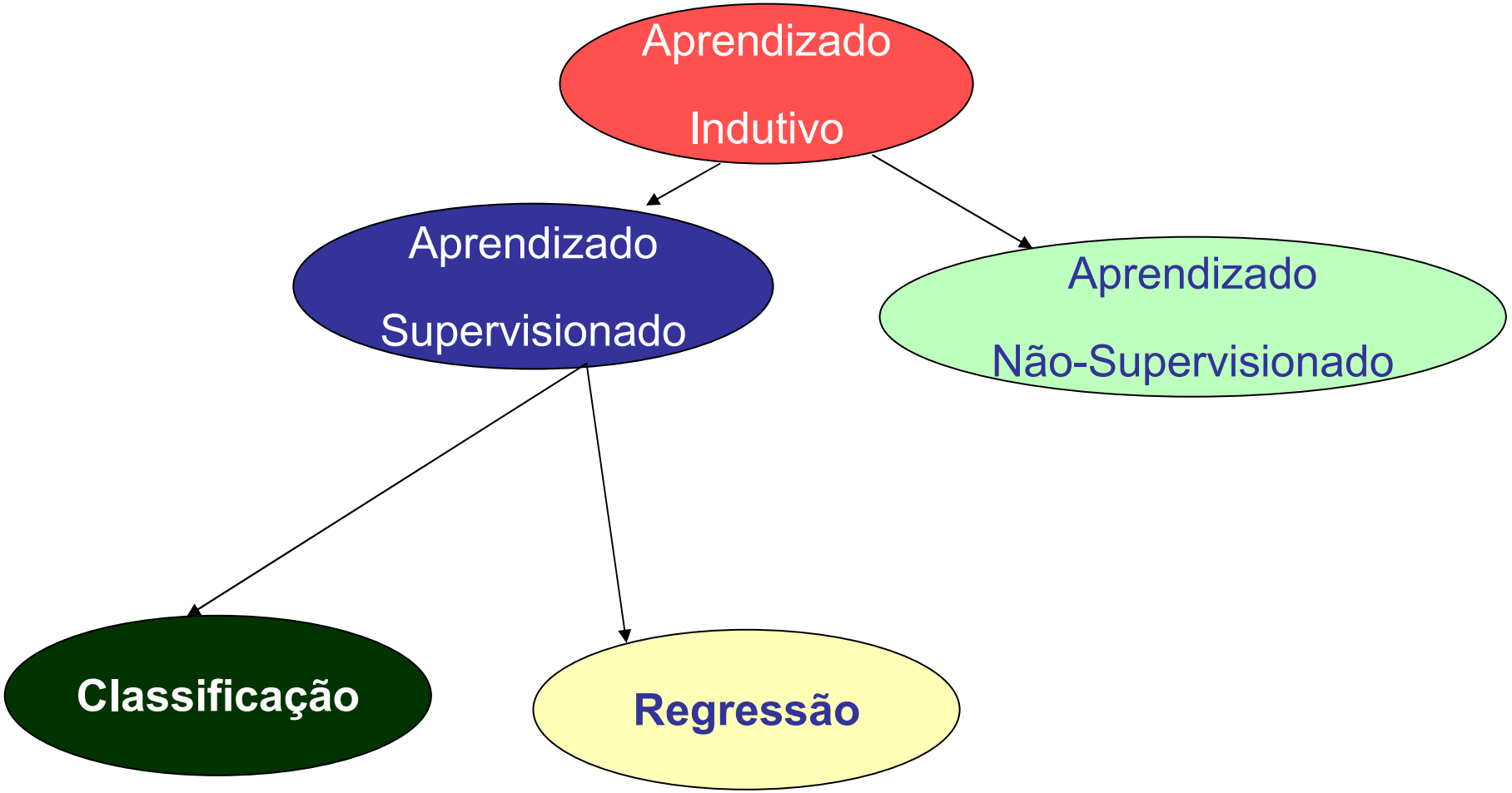
- O que é **Aprendizado de Máquina**?
  - É uma área de IA cujo objetivo é o **desenvolvimento de técnicas computacionais** sobre **o aprendizado** bem como a construção de sistemas capazes **de adquirir conhecimento de forma automática**.
  - Um **sistema de aprendizado** é um programa de computador que toma decisões baseado em experiências acumuladas por meio de solução bem sucedida de problemas anteriores.
  - É uma **ferramenta poderosa para aquisição automática** de conhecimento, entretanto, **não existe um único algoritmo** que apresente **melhor desempenho** para todos problemas.

Aprendizado de Máquina ...

- **Indução:** é a forma de inferência lógica que permite obter conclusões a partir de um conjunto de exemplos.
- Na indução, um conceito é aprendido efetuando-se **inferência indutiva** sobre os **exemplos apresentados** *(cautela na escolha de exemplos)*.
- **Dedução:** Humanos usam raciocínio dedutivo para deduzir nova informação a partir de informação relacionada logicamente.

Aprendizado de Máquina ...

- Hierarquia do aprendizado



Classificação Versus Clusterização

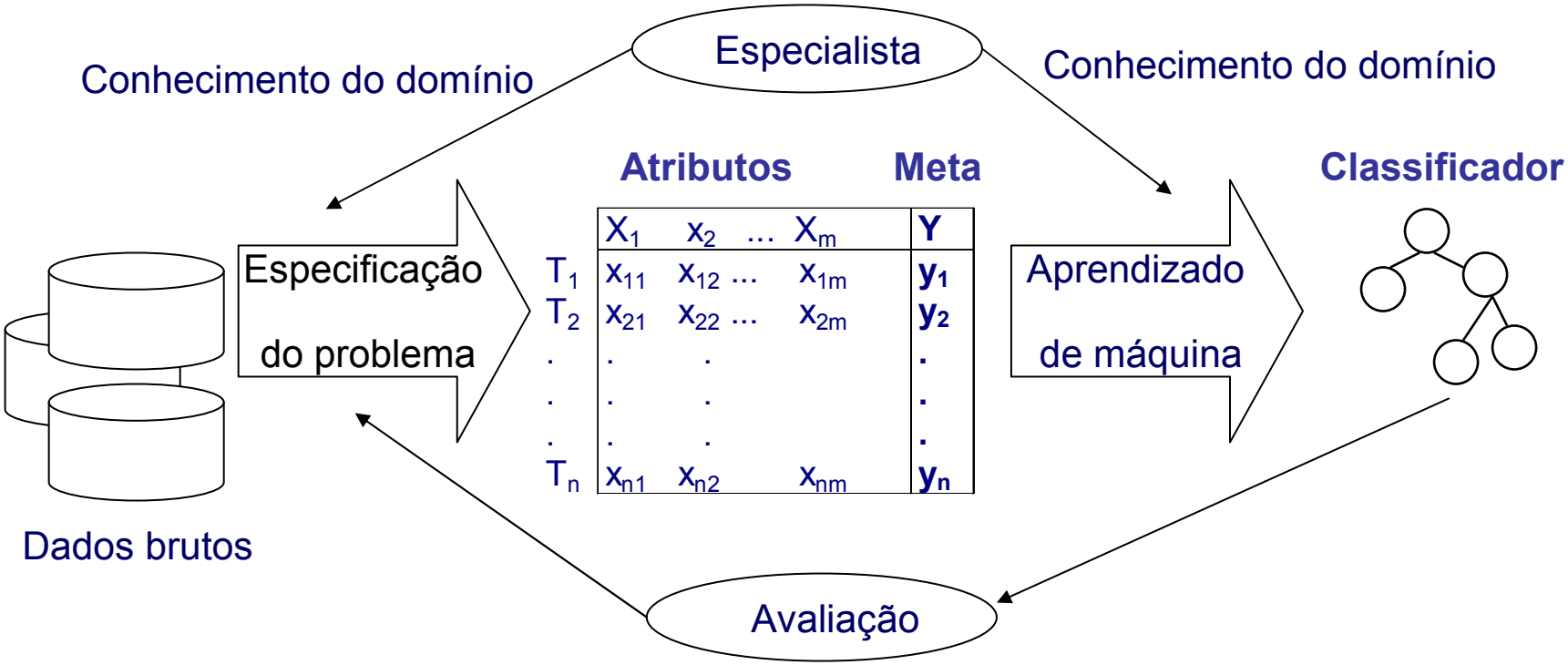
- **Aprendizado supervisionado** (**classificação**)
  - **Supervisão**: As observações no conjunto de treinamento são acompanhadas por "**labels**" indicando a classe a que elas pertencem.
  - Novas ocorrências são classificadas com base no conjunto de treinamento.
- **Aprendizado não-supervisionado** (**clusterização**)
  - Não existe classe pré-definida para nenhum dos atributos.
  - Um conjunto de observações é dado com o propósito de se estabelecer a existência das classes ou clusters.

Aprendizado de Máquina ...

- Hierarquia do aprendizado
  - **Aprendizado Indutivo:** É efetuado a partir de exemplos externos ao sistema de aprendizado.
    - **Aprendizado Supervisionado:** Construir um classificador (indutor) que possa determinar a classe de novos exemplos a partir de exemplos treinamento com classe rotulada.
      - **Classificação: Rótulos para valores discretos.**
      - **Regressão: Rótulos para valores contínuos.**
    - **Aprendizado Não-Supervisionado**: O indutor analisa os exemplos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou *clusters*.

Aprendizado de Máquina ...

- Hierarquia do aprendizado
  - **Processo de Classificação**

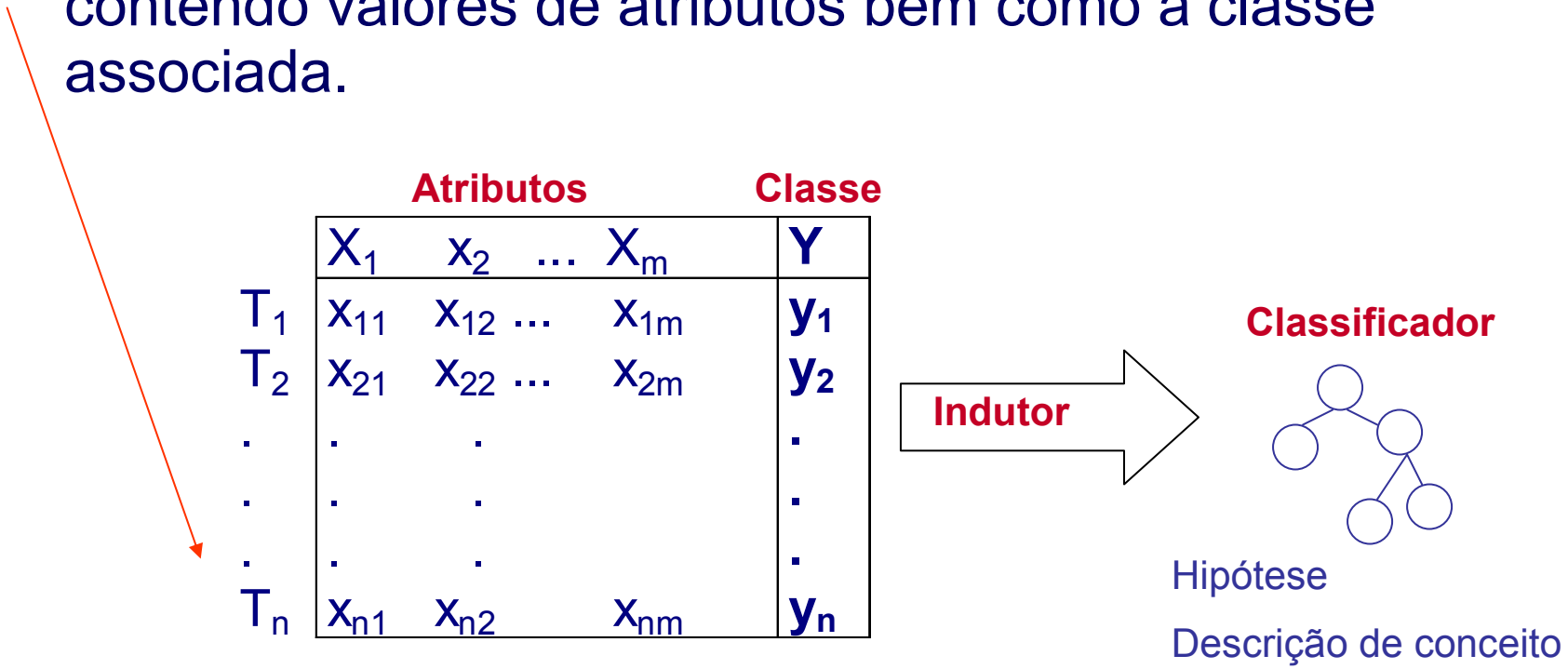


Aprendizado de Máquina ...

- Paradigmas do aprendizado:
  - **Simbólico:** Buscam aprender construindo representações simbólicas (expressão lógica, **árvores de decisão regras**).
  - **Estatístico:** Buscam métodos estatísticos (**Aprendizado bayesiano**)
  - **Baseado em Exemplos:** Sistemas *lazy* (RBC, ***Nearest Neighboors***).
  - **Conexionista:** Modelos inspirados no modelo biológico do sistema nervoso (**Redes Neurais**).
  - **Evolutivo:** Teoria de Darwin (**Algoritmos Genéticos**).

Aprendizado de Máquina: Definições

- Algumas Definições em AM
  - **Conjunto de exemplos:** é um conjunto de exemplos contendo valores de atributos bem como a classe associada.



Aprendizado de Máquina: Definições

- Algumas Definições em AM
  - **Indutor:** programa de aprendizado ou algoritmo de indução que visa extrair um bom classificador a partir de um conjunto de exemplos rotulados.
  - **Exemplo:** caso, dado ou registro.
  - **Atributo:** descreve uma característica ou aspecto de um exemplo.
  - **Classe (Atributo-Meta):** atributo especial no qual se pretende aprender a fazer previsões a respeito.

Aprendizado de Máquina: Definições

- Algumas Definições em AM
  - **Bias:** qualquer preferência de uma hipótese sobre a outra.
  - **Modo de aprendizado:**
    - todo conjunto de treinamento presente no aprendizado (**não incremental**).
    - quando novos exemplos de treinamento são adicionados (**incremental**).

Aprendizado de Máquina: Definições

Erro (err(h))

Medida de desempenho de um Classificador.

Considerando $||E||$ = { 1 se a expressão for verdadeira
0, caso contrário

$$err(h) = \frac{1}{n} \sum_{i=1}^{n} || \; y_i \neq h(x_i) ||$$

Acurácia (acc(h))

Complemento da Taxa de Erro, representa a Precisão do Classificador.

$$acc(h) = 1 - err(h)$$

Aprendizado de Máquina: Definições

Distribuição de Classes (distr(Cj))

Para cada Classe Cj, sua distribuição distr(Cj) é calculada como sendo o número de exemplos em T que possuem classe Cj dividido pelo número total de exemplos (n), ou seja, a proporção de exemplos em cada classe

$$distr(C_j) = \frac{1}{n} \sum_{i=1}^{n} \| y_i = C_j \|$$

Exemplo: Conjunto com 100 Exemplos: 60 Classe C1
15 Classe C2
25 Classe C3

distr (C1, C2, C3) = (60%, 15%, 25%)

Neste exemplo, **Classe Majoritária** (ou **Prevalente**) é C1.
Classe **Minoritária** é C2.

Aprendizado de Máquina: Definições

Erro Majoritário (maj-err(T))

Limite Máximo abaixo do qual o erro de um Classificador deve ficar

$$maj - err(T) = 1 - \max_{i=1,\ldots k} distr(C_i)$$

No Exemplo anterior: maj-err(T) = 1 - 0,60 = 0,40

Erro Majoritário **INDEPENDE** do algoritmo de aprendizado.

Aprendizado de Máquina: Definições

Prevalência de Classe

Problema com **desbalanceamento** de classes em conjunto de exemplos.

Exemplo: distr (C1, C2) = (99,75%, 0,25%)

Neste exemplo, Classe **Majoritária** (ou **Prevalente**) é C1
Classe **Minoritária** é C2

Classificador que classifique novos exemplos como C1 teria uma precisão de 99,75%.

Se a Classe C2 fosse, por exemplo, ocorrência de Geada ...

Aprendizado de Máquina: Definições

“Overfitting”

- É possível que o Classificador faça uma indução muito específica para o conjunto de treinamento utilizado (**“Overfitting”**).

- Como este é apenas uma amostra de dados, é possível que a indução tenha bom desempenho no conjunto de treinamento, mas um desempenho ruim em exemplos diferentes daqueles pertencentes ao conjunto de treinamento.

- **Cálculo do Erro** em um conjunto de teste independente evidencia a situação de **“Overfitting”**.

- **Under** e **overfitting**: ajusta-se em muito pouco ou em excesso ao conjunto de treinamento.

O que é classificação?

- Dado um conjunto de registros (dataset):
  - Cada registro contém um conjunto de **atributos**, em que um dos atributos é o **atributo-meta** (**variável resposta**).
  - O conjunto de dados é dividido em dois subconjuntos: **conjunto de treinamento** para construir o modelo e **conjunto de teste** para validar o modelo.
- **Passo 1**: encontrar um **modelo** para o atributo-meta (ou atributo-classe) como uma função dos valores dos outros atributos.
- **Passo 2**: registros não conhecidos devem ser associados à classe com a maior precisão possível.

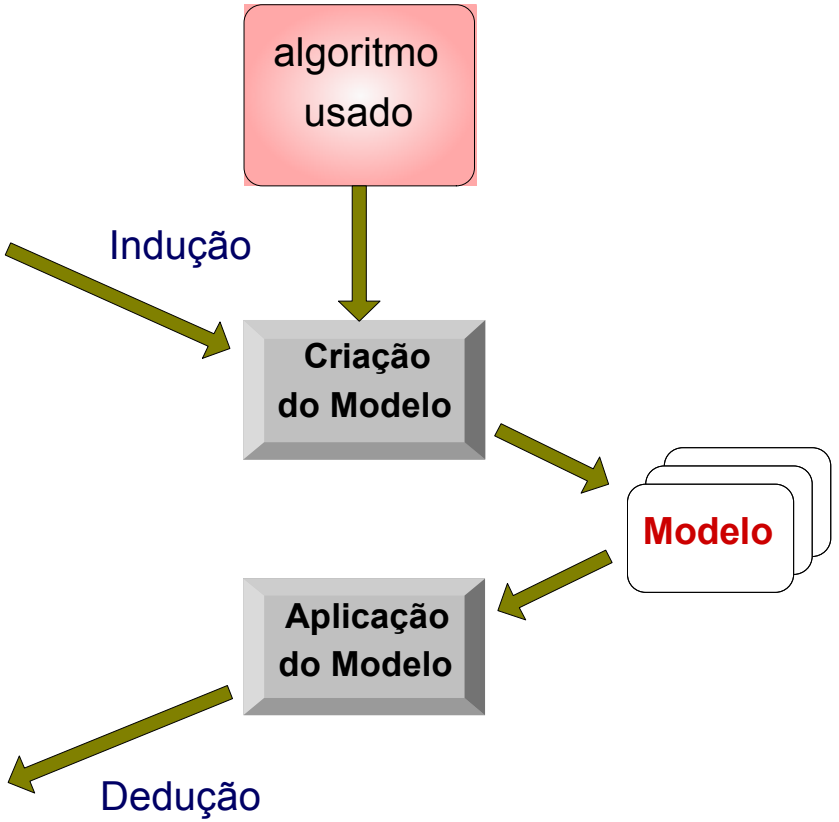
O que é classificação? ...

Tid	Atrib1	Atrib2	Atrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Conjunto de treinamento

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Conjunto de Teste

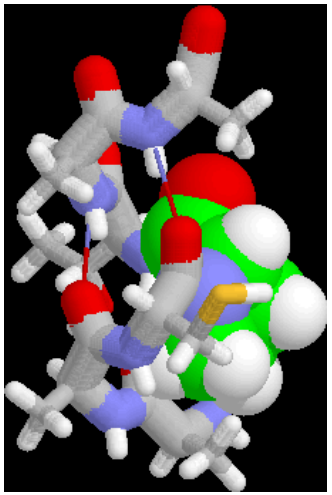


Observação Importante

- **Qualidade** do Exemplo de Treinamento → **Qualidade das Regras**.
- **Não** é possível **descobrir** algo que **não** esteja nos exemplos.
- **Seleção dos exemplos** e das **características** é fundamental neste processo.
  - Daí a **importância** do **conhecimento** e da **experiência** do **Especialista**.

Classificação: Aplicações

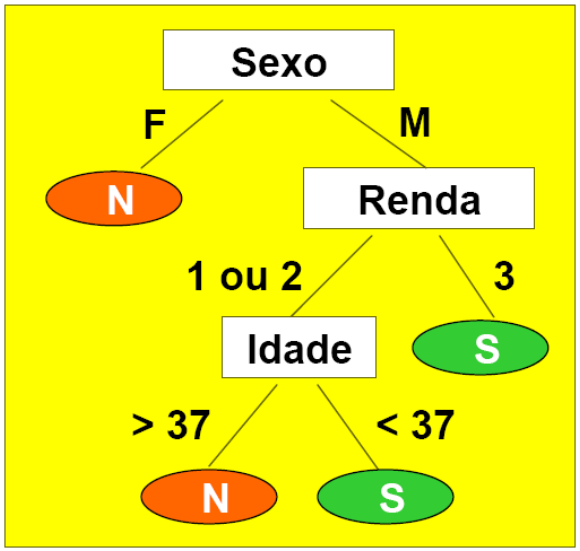
- Classificar tumores como benigno ou maligno.
- Classificar transações de cartão de crédito como legítima ou fraudulenta.
- Classificar estruturas secundárias de proteínas como alpha-helix, beta-sheet ou random coil.
- Avaliar riscos de empréstimos, previsão de tempo, etc.
- Sistema de alerta de geada.
- Previsão de mortalidade de frangos, etc.



Características de um bom classificador

- **Precisão**
- **Velocidade**
  - Tempo para construir o modelo.
  - Tempo para usar o modelo.
- **Robustez**
  - Capacidade de lidar com ruídos e valores faltantes (**missing**).
- **Escalabilidade**
  - Eficiência em banco de dados residentes em disco.
- **Interpretabilidade**
  - Clareza fornecida pelo modelo.
- **Relevância na seleção de regras**
  - Tamanho da árvore de decisão.
  - Regras de classificação compactas.

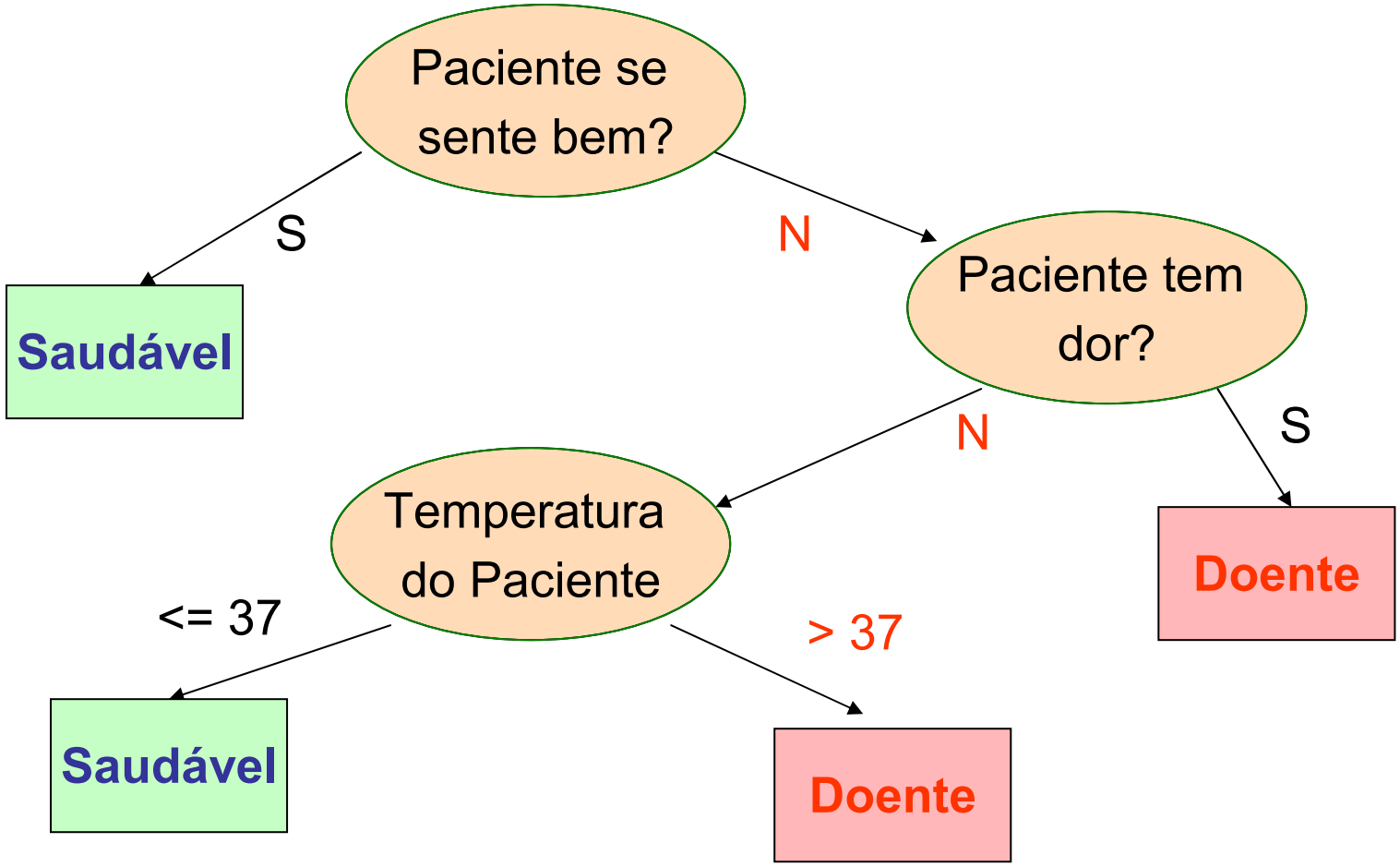
Árvores de Decisão



Árvore de Decisão

- Árvore de decisão
  - Um fluxograma com a estrutura de uma árvore.
  - Nó interno representa um testes sobre um atributo.
  - Cada ramo representa um resultado do teste.
  - Folhas representam as classes.
- A geração de uma árvore consiste de duas fases:
  - **Construção da árvore**
    - Particionamento de atributos (**best fit**).
  - **Fase da poda** (**Tree pruning**).
    - Identifica e remove ramos que refletem ruídos ou outliers.
- **Uso da árvore**: Classificação de amostras desconhecidas
  - Testa os valores dos atributos da amostra "**contra**" a árvore.

Árvore de Decisão – Exemplo



Árvore de Decisão – Exemplo ...

- **Geração de regras**

Se paciente se sente bem = ***sim***
    então classe = ***saudável***
fim se.
Se paciente se sente bem = **não**
    e paciente tem dor = **sim**
    então classe = **doente**
fim se.

...

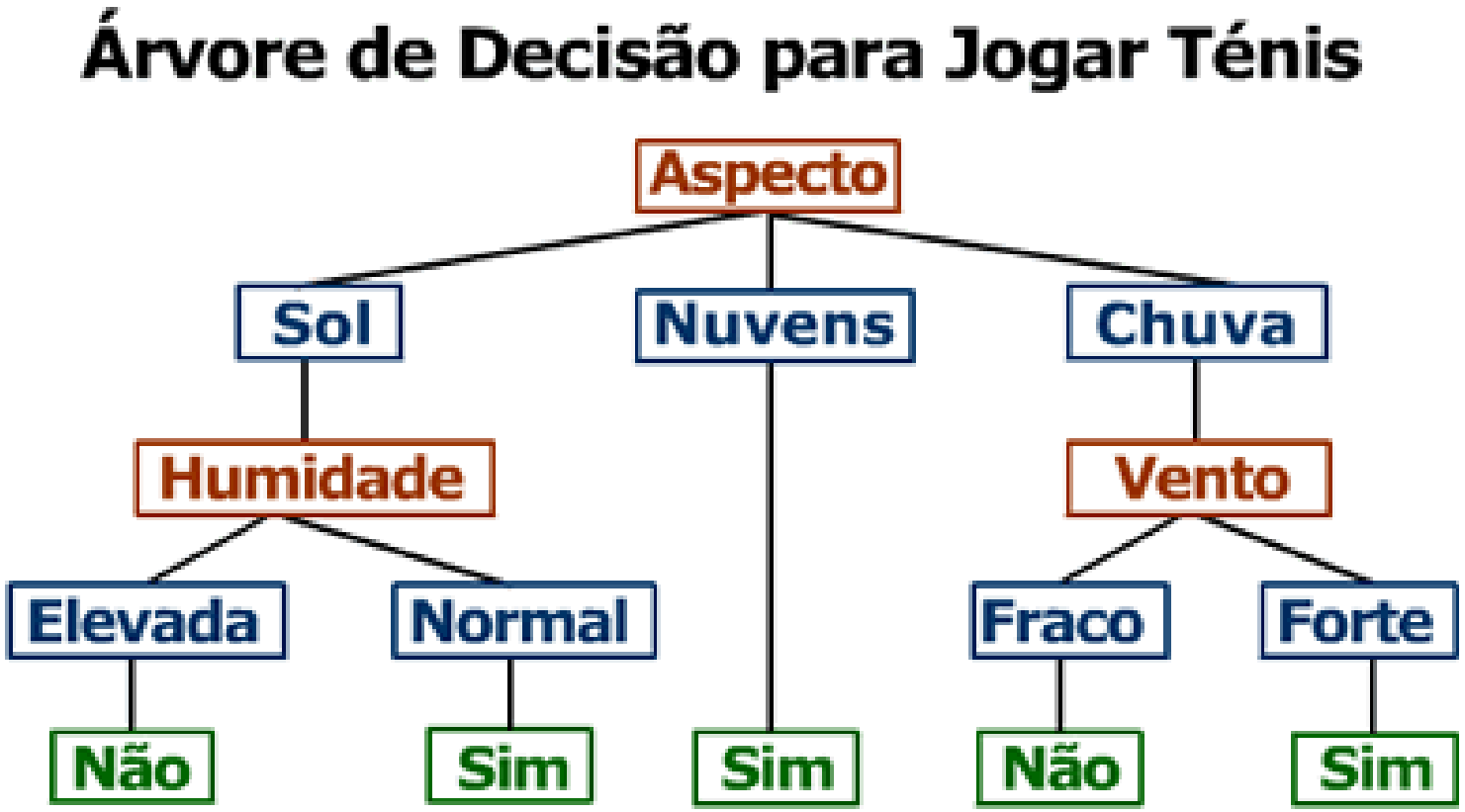
Árvore de Decisão – Exemplo ...

- **Exemplo**: Condições do tempo para **Jogar Tênis**.

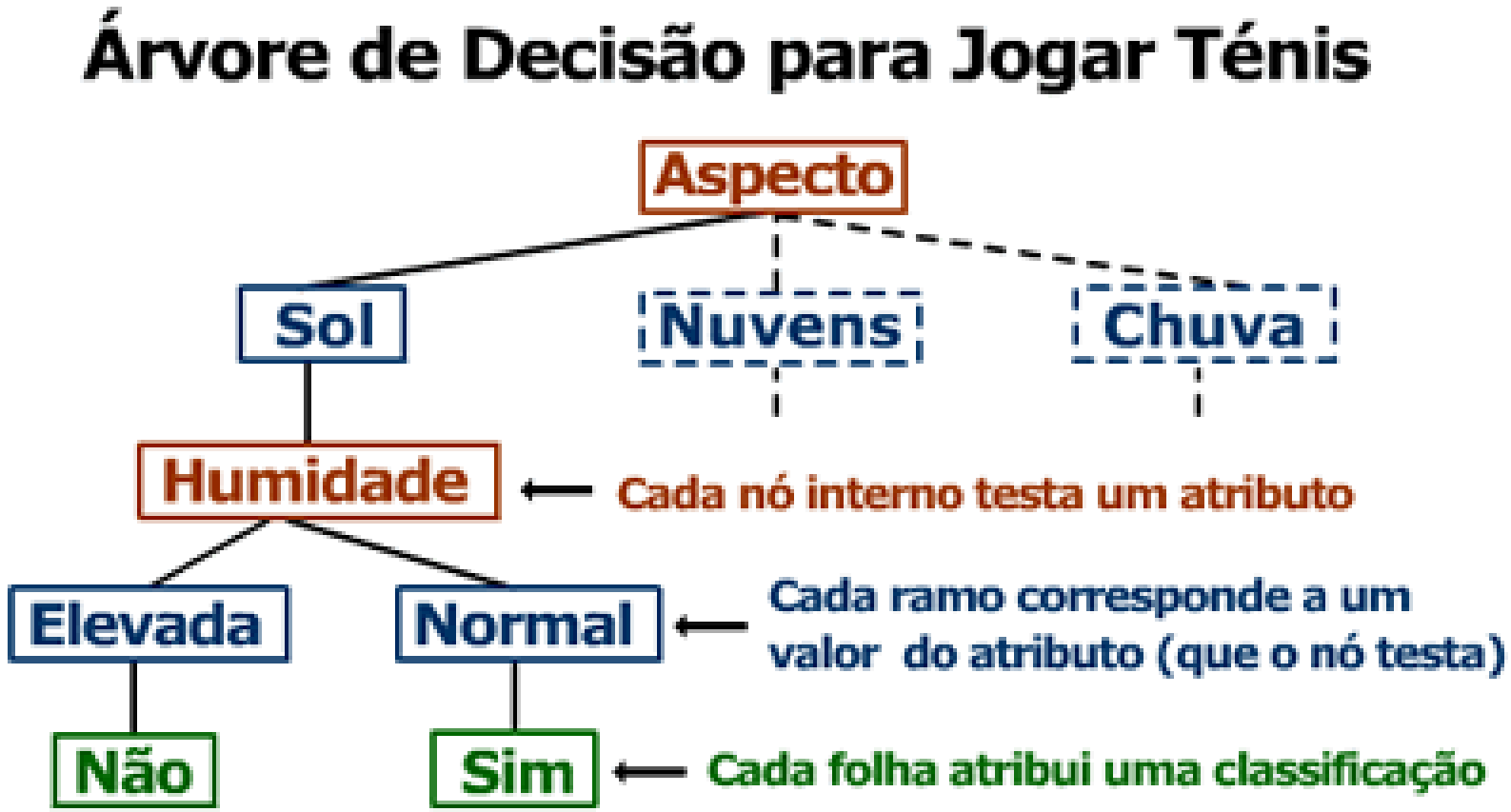
Exemplos de Treino

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

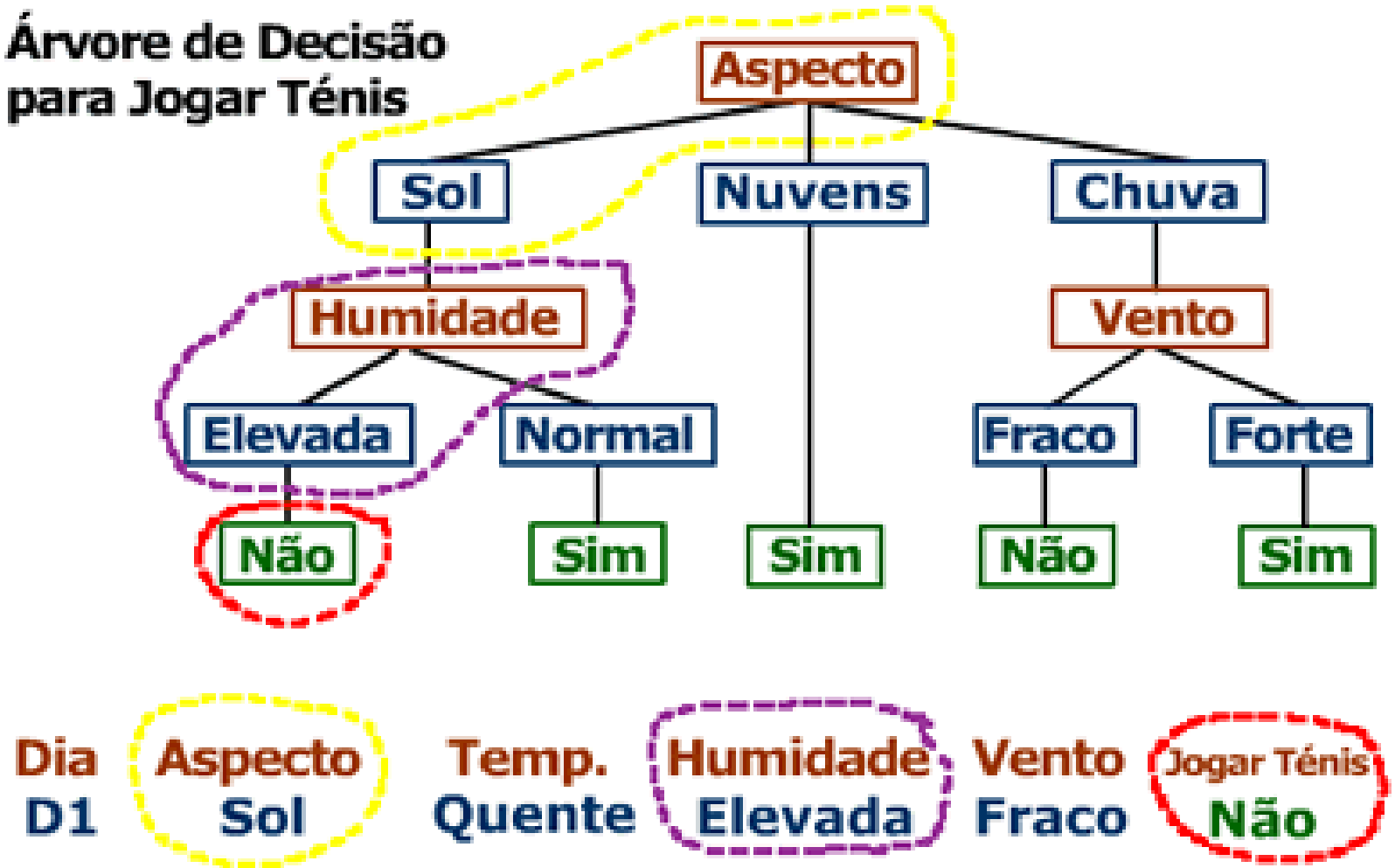
Árvore de Decisão – Exemplo ...



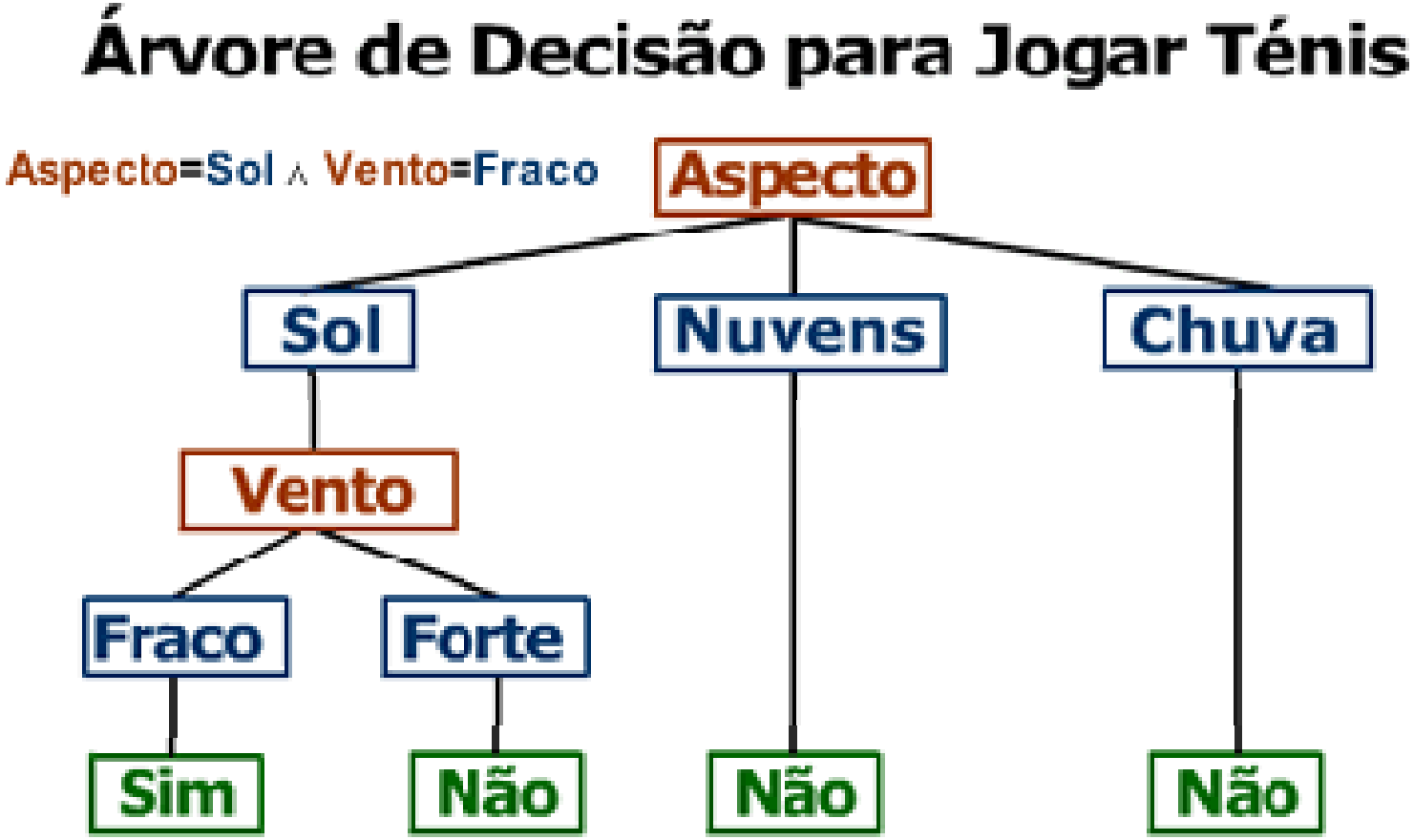
Árvore de Decisão – Exemplo ...



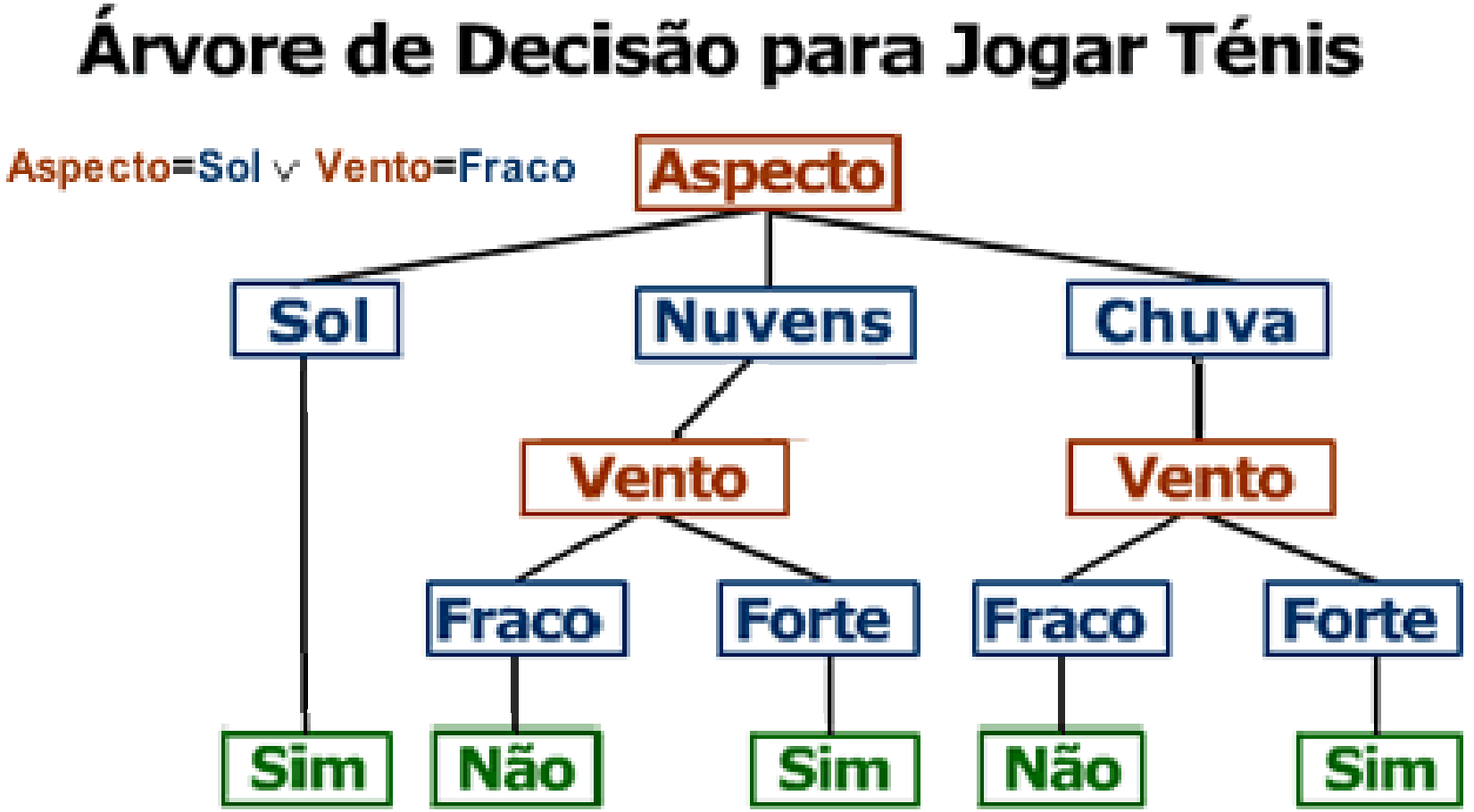
Árvore de Decisão – Exemplo ...



Árvore de Decisão – Exemplo ...



Árvore de Decisão – Exemplo ...



Algoritmos para árvores de decisão

- Algoritmo Básico (**algoritmo guloso**)
  - A árvore é construída recursivamente no sentido **top-down (divisão para conquista)**.
  - No início, todas as amostras estão na raiz.
  - Os atributos são nominais (se numéricos, eles são discretizados).
  - Amostras são particionadas recursivamente com base nos atributos selecionados.
  - Atributos "**testes**" são selecionados com base em heurísticas ou medidas estatísticas (ex., **ganho de informação**) **[ID3 / C4.5]**
- Condições de parada do particionamento
  - Todas as amostras de um nó pertencem a mesma classe.
  - Não existem mais atributos para particionamento.
  - Não existem mais amostras no conjunto de treinamento.

Determinando o tamanho da árvore

As seguintes abordagens podem ser usadas:

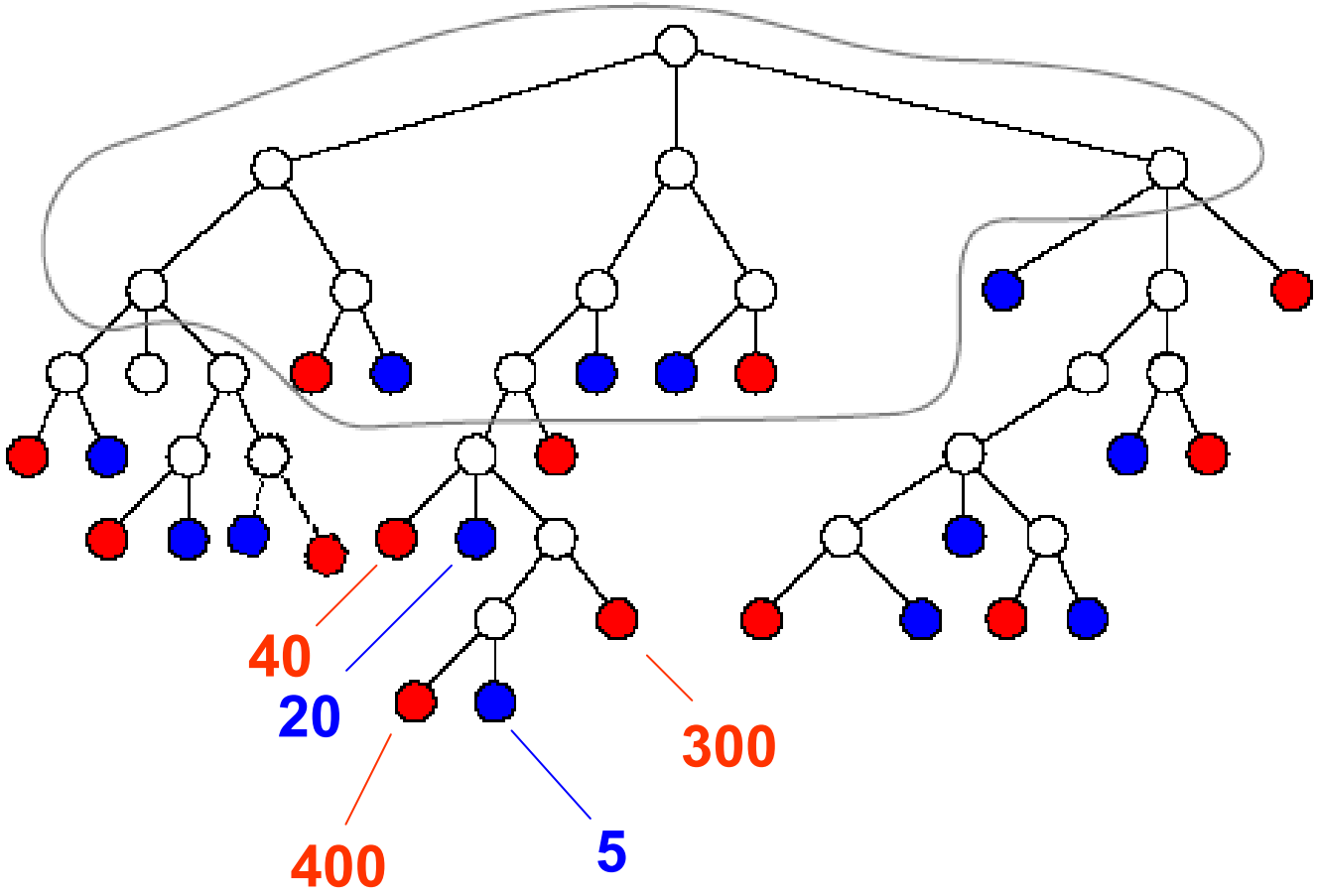
- **Divisão do dataset**: conjunto de treinamento (2/3) e conjunto de teste (1/3) – **Percentage split**.
- Uso de validação cruzada (**10-fold cross validation**).
- Uso de conjunto de teste extra (**Supplied test set**).
- Uso de todo o dataset para treinamento
  - Aplicação de um **teste estatístico** (ex:, **qui-quadrado**) para estimar se expandindo ou podado um nó pode melhorar a distribuição total.

Árvore de Decisão: Poda

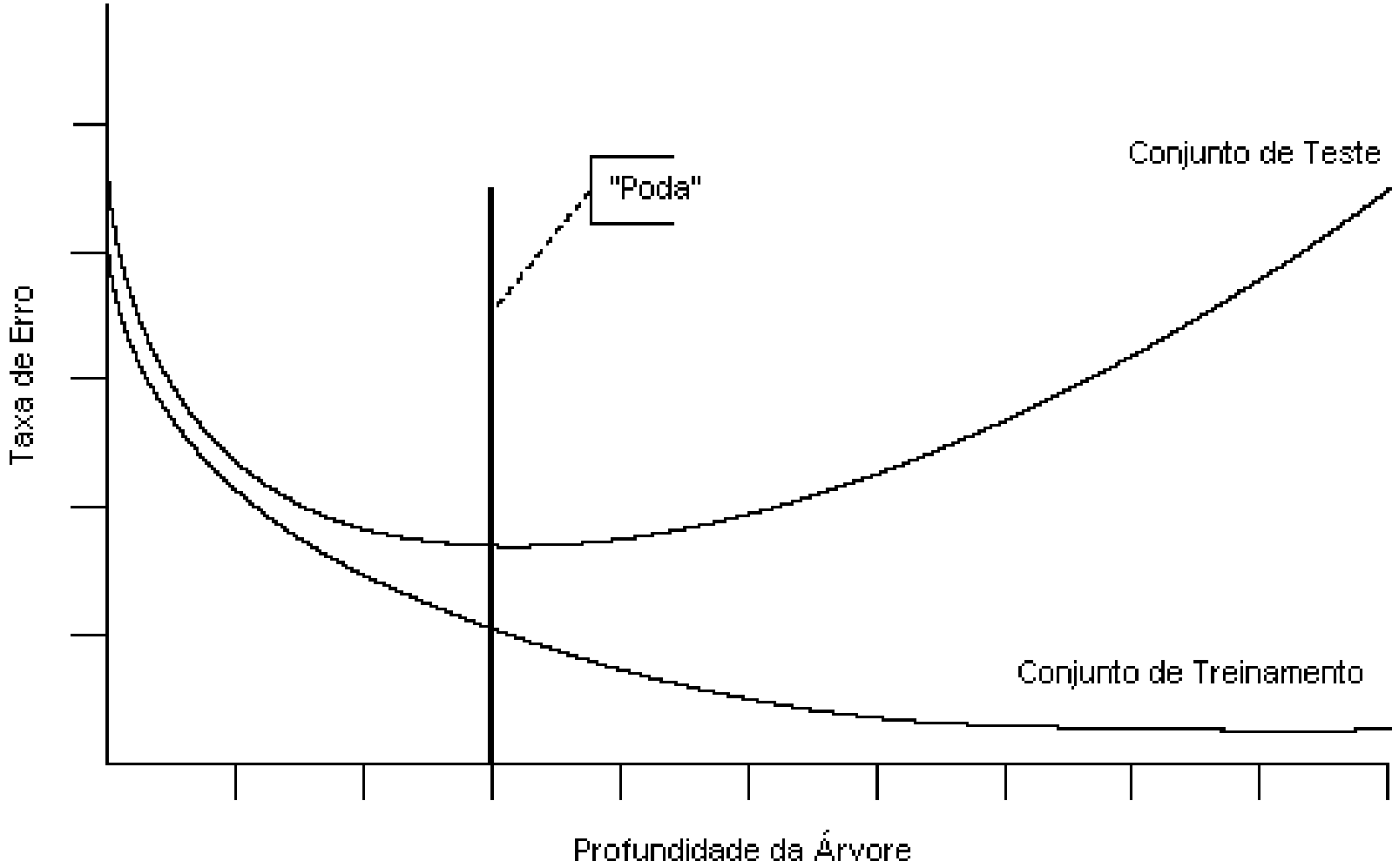
PODA

- Técnica para lidar com ruído e **"Overfitting"**
- **Pré-Poda**: Durante a geração da Hipótese.
  - Alguns exemplos de treinamento são deliberadamente ignorados.
- **Pós-Poda**: Inicialmente, é gerada um Classificador que explique os exemplos.
  - Após isso, elimina-se algumas partes (**cortes em ramos da árvore**) generalizando a Hipótese.

Árvore de Decisão: Poda ...



Árvore de Decisão: Poda ...



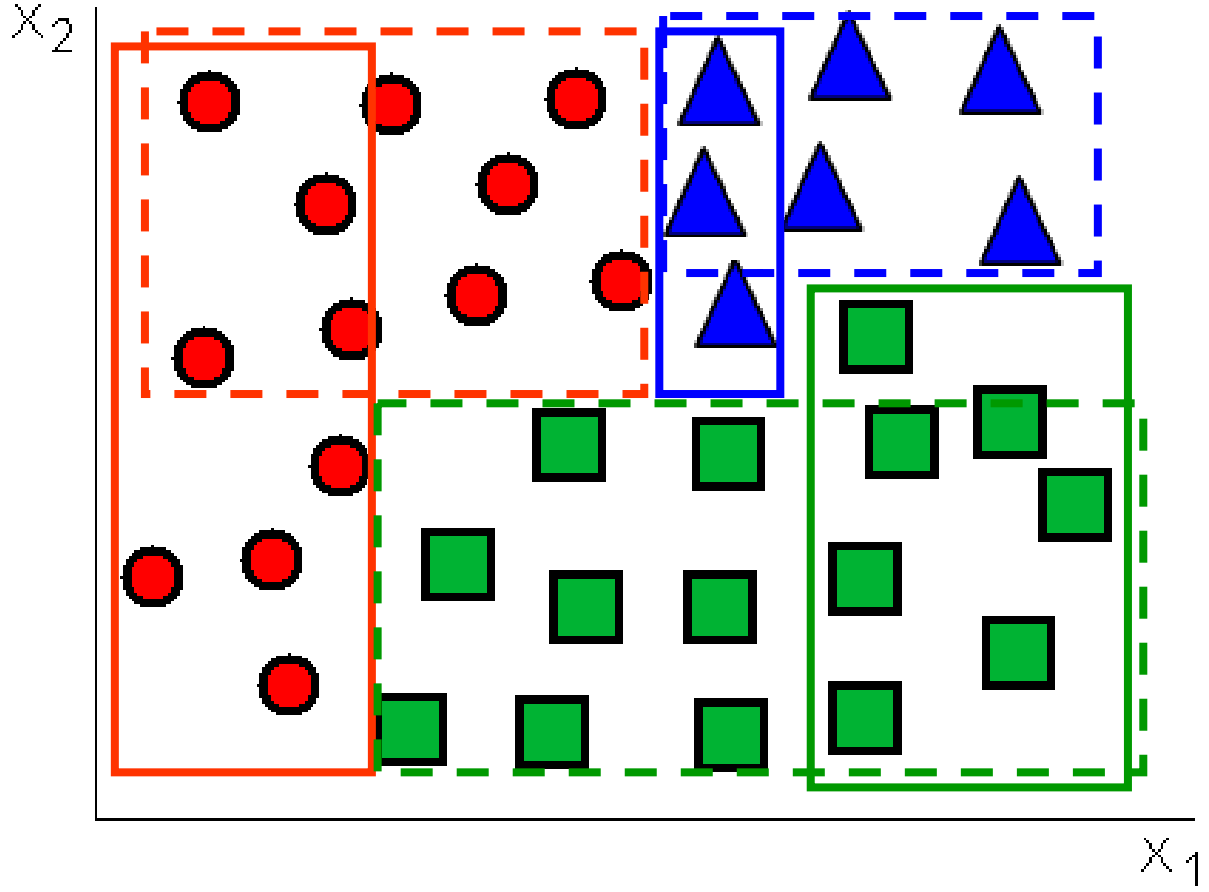
Completude e Consistência

COMPLETUDE E CONSISTÊNCIA

- **COMPLETUDE**: Se a Hipótese gerada pelo Classificador classifica **TODOS** os exemplos.
- **CONSISTÊNCIA**: Se a Hipótese gerada pelo Classificador classifica **CORRETAMENTE** os exemplos.
- Uma Hipótese gerada pelo Classificador pode ser:
  - Completa e Consistente.
  - Incompleta e Consistente.
  - Completa e Inconsistente.
  - Incompleta e Inconsistente.

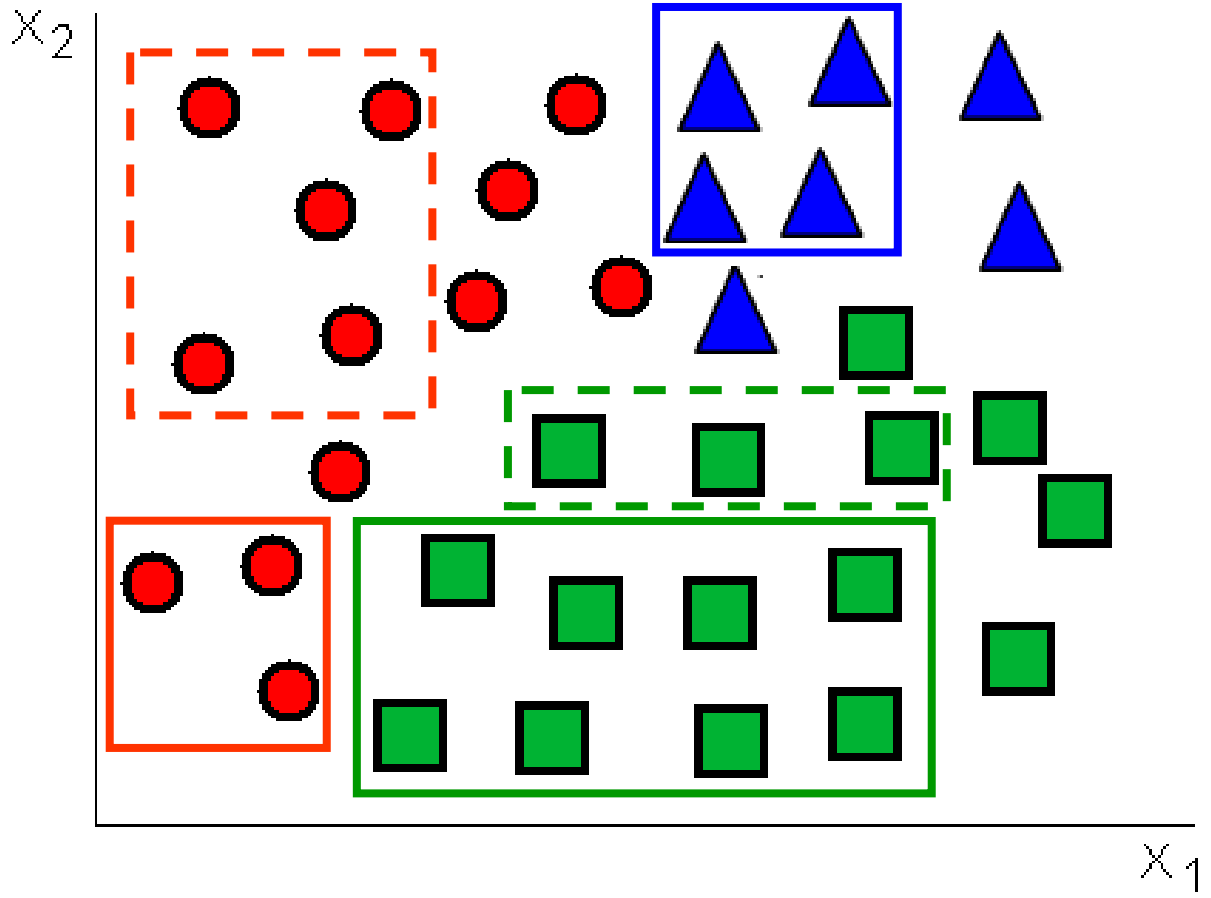
Completude e Consistência ...

COMPLETO e CONSISTENTE



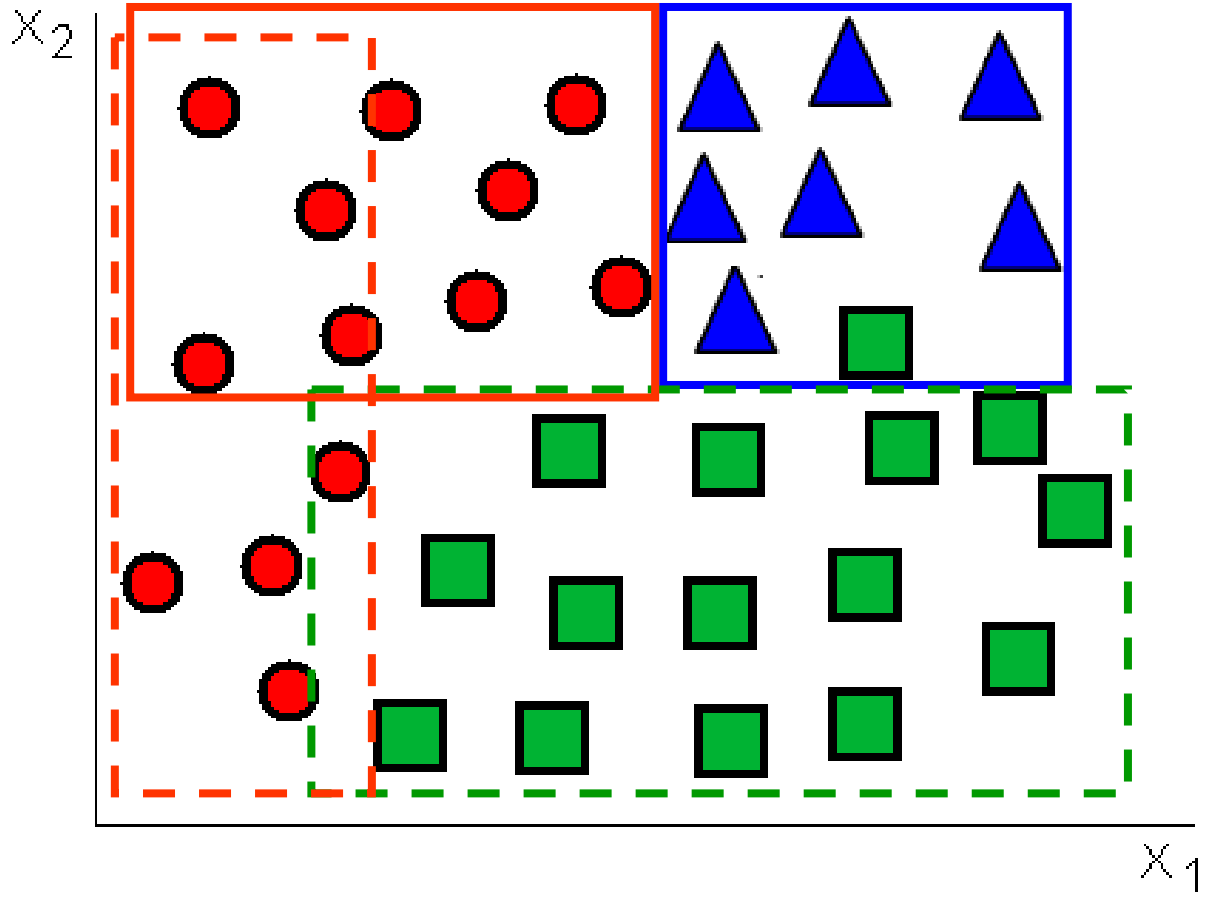
Completude e Consistência ...

INCOMPLETO e CONSISTENTE



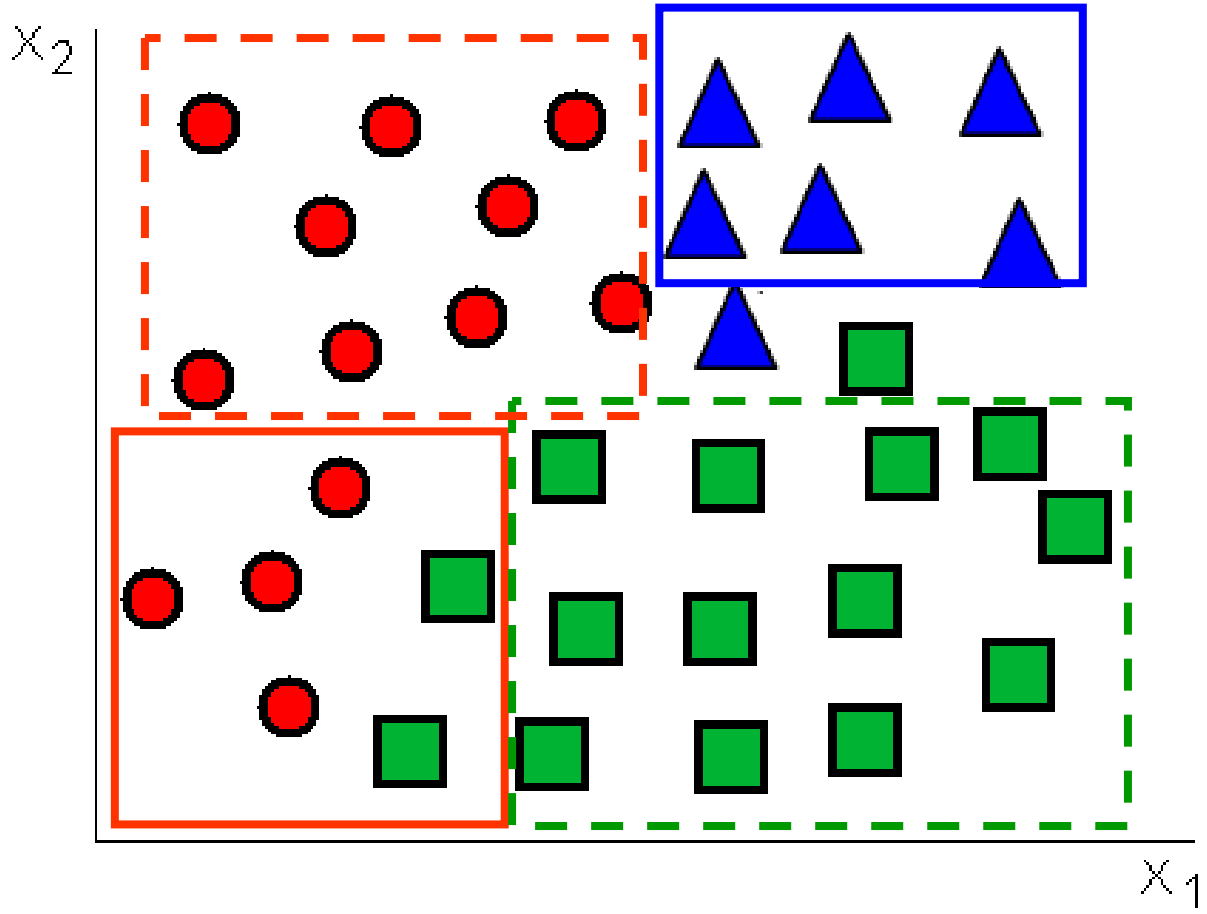
Completude e Consistência ...

COMPLETO e INCONSISTENTE



Completude e Consistência ...

INCOMPLETO e INCONSISTENTE



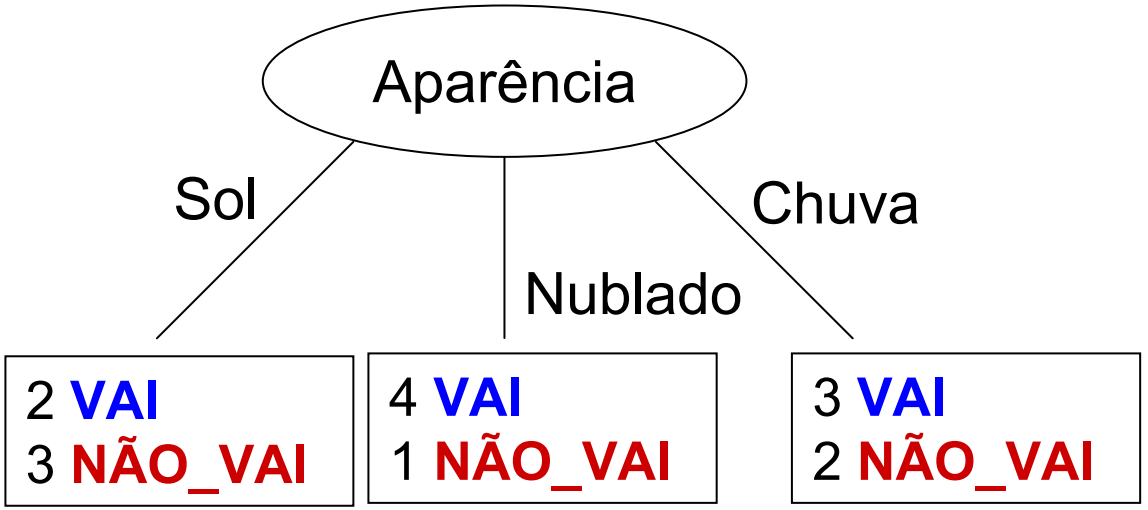
Completude e Consistência: Exemplo

Exemplo	Aparência	Temperatura	Umidade	Ventando	Viajar
T1	sol	25	72	sim	VAI
T2	sol	28	91	sim	NÃO_VAI
T3	sol	22	70	não	VAI
T4	sol	23	95	não	NÃO_VAI
T5	sol	30	85	não	NÃO_VAI
T6	nublado	23	90	sim	VAI
T7	nublado	29	78	não	VAI
T8	nublado	19	65	sim	NÃO_VAI
T9	nublado	26	75	não	VAI
T10	nublado	20	87	sim	VAI
T11	chuva	22	95	não	VAI
T12	chuva	19	70	sim	NÃO_VAI
T13	chuva	23	80	sim	NÃO_VAI
T14	chuva	25	81	não	VAI
T15	chuva	21	80	não	VAI

Completude e Consistência: Exemplo

Exemplo	Aparência	Temperatura	Umidade	Ventando	Viajar
T1	sol	25	72	sim	VAI
T2	sol	28	91	sim	NÃO_VAI
T3	sol	22	70	não	VAI
T4	sol	23	95	não	NÃO_VAI
T5	sol	30	85	não	NÃO_VAI
T6	nublado	23	90	sim	VAI
T7	nublado	29	78	não	VAI
T8	nublado	19	65	sim	NÃO_VAI
T9	nublado	26	75	não	VAI
T10	nublado	20	87	sim	VAI
T11	chuva	22	95	não	VAI
T12	chuva	19	70	sim	NÃO_VAI
T13	chuva	23	80	sim	NÃO_VAI
T14	chuva	25	81	não	VAI
T15	chuva	21	80	não	VAI

Completude e Consistência: Exemplo



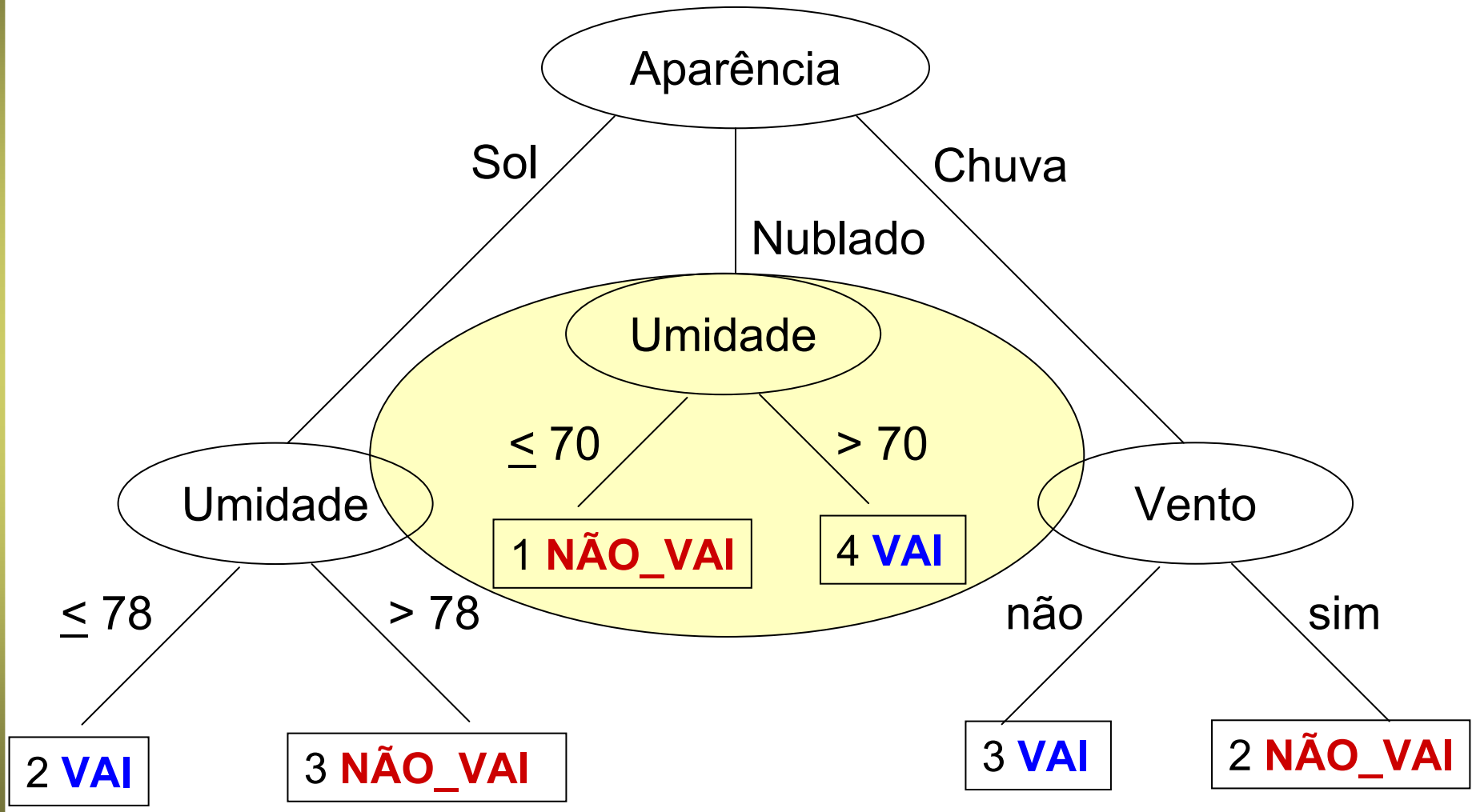
Completude e Consistência: Exemplo

Exemplo	Aparência	Temperatura	Umidade	Ventando	Viajar
T1	sol	25	72	sim	VAI
T2	sol	28	91	sim	NÃO_VAI
T3	sol	22	70	não	VAI
T4	sol	23	95	não	NÃO_VAI
T5	sol	30	85	não	NÃO_VAI
T6	nublado	23	90	sim	VAI
T7	nublado	29	78	não	VAI
T8	nublado	19	65	sim	NÃO_VAI
T9	nublado	26	75	não	VAI
T10	nublado	20	87	sim	VAI
T11	chuva	22	95	não	VAI
T12	chuva	19	70	sim	NÃO_VAI
T13	chuva	23	80	sim	NÃO_VAI
T14	chuva	25	81	não	VAI
T15	chuva	21	80	não	VAI

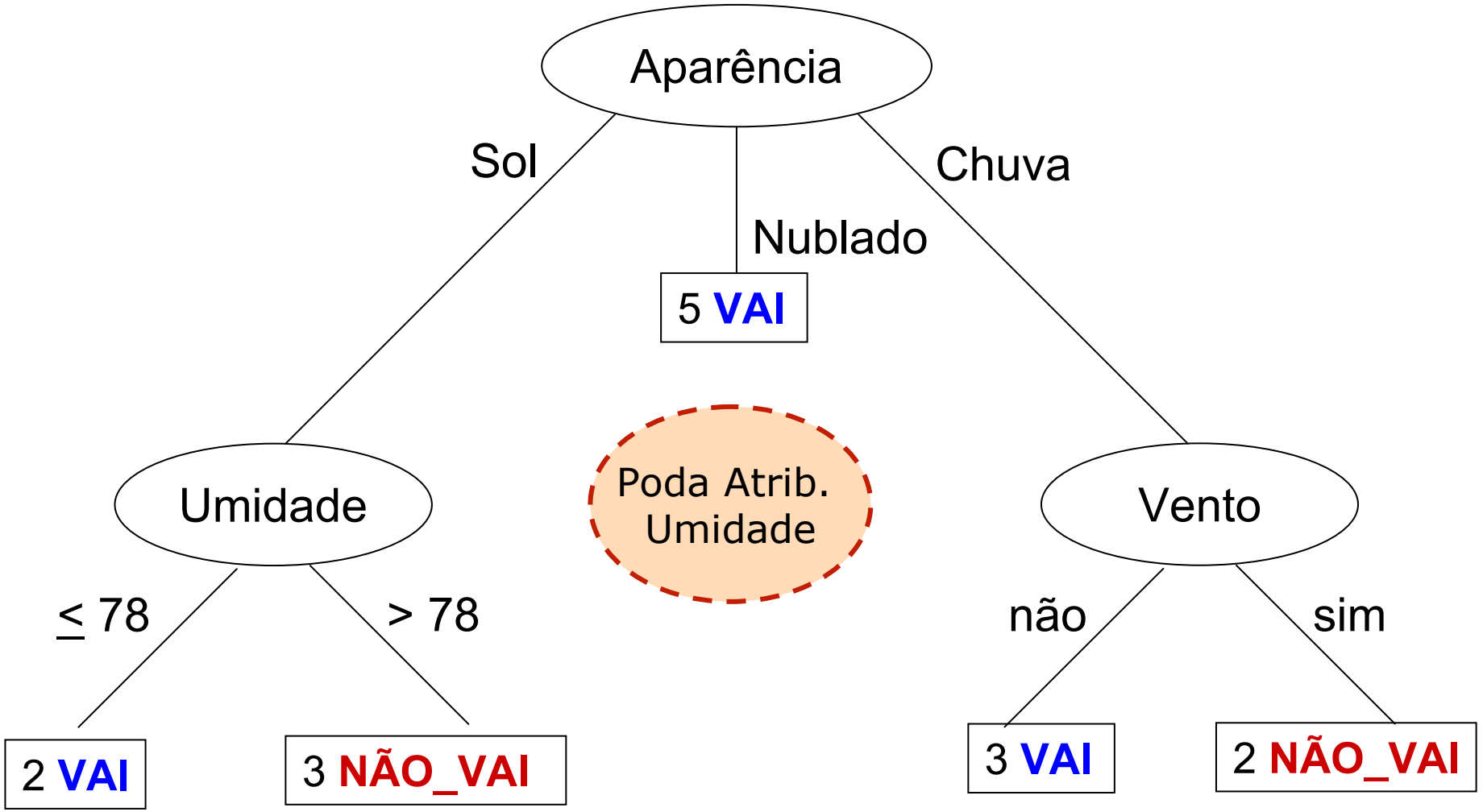
Completude e Consistência: Exemplo

Exemplo	Aparência	Temperatura	Umidade	Ventando	Viajar
T1	sol	25	72	sim	VAI
T2	sol	28	91	sim	NÃO_VAI
T3	sol	22	70	não	VAI
T4	sol	23	95	não	NÃO_VAI
T5	sol	30	85	não	NÃO_VAI
T6	nublado	23	90	sim	VAI
T7	nublado	29	78	não	VAI
T8	nublado	19	65	sim	NÃO_VAI
T9	nublado	26	75	não	VAI
T10	nublado	20	87	sim	VAI
T11	chuva	22	95	não	VAI
T12	chuva	19	70	sim	NÃO_VAI
T13	chuva	23	80	sim	NÃO_VAI
T14	chuva	25	81	não	VAI
T15	chuva	21	80	não	VAI

Completude e Consistência: Exemplo



Completude e Consistência: Exemplo



Aspectos Importantes

Natureza eliminatória do processo

- Exemplos do conjunto de treinamento vão sendo **descartados** à medida que são utilizados. Isso causa um efeito importante na definição dos nós mais profundos da árvore, uma vez que o suporte estatístico para a tomada de decisão vai sendo progressivamente diminuído, colocando em dúvida a representatividade do conjunto de elementos remanescentes.

- Possibilidade de utilizar recursos para descarte (**"poda"**) dos ramos tecnicamente menos representativos da árvore da decisão.

Algoritmos mais conhecidos

- **ID3** (**Iterative Dichotomiser 3**) (Quilan,1986):
  - Um algoritmo usado para gerar árvores de decisão. Os atributos do conjunto de dados devem ser obrigatoriamente categóricos.
- **C4.5** (**J48 no Weka**) (Quilan, 1993):
  - Um algoritmo para geração de árvores de decisão, sucessor do algoritmo ID3.
  - O algoritmo C4.5 considera atributos numéricos e categóricos.
- **CART** (**C**lassification **A**nd **R**egression **T**rees) (Breiman et al., 1984):
  - Técnica não-paramétrica que produz árvores de classificação ou regressão, dependendo se as variáveis são categóricas ou numéricas, respectivamente.

Como escolher o melhor atributo?

Escolha do melhor atributo “split”

- **Não existe solução computacionalmente viável** para que se obtenha sempre a melhor árvore de decisão possível (problema **NP-completo**: custo de proceder buscas exaustivas da melhor solução cresce a taxas exponenciais à medida que o tamanho do conjunto de treinamento aumenta).

- **Utilização de heurísticas**: soluções baseadas em algum tipo de conhecimento prévio sobre as propriedades dos dados, na procura de uma boa solução (mas não necessariamente a melhor).

Como escolher o melhor atributo?

Exemplo: Conjunto de todas soluções possíveis (**floresta de decisão**).

BUSCA EXAUSTIVA:
Correr todo esse conjunto, comparando cada elemento, até que todos tenham sido avaliados, e selecionar a melhor solução.
SOLUÇÃO ÓTIMA GARANTIDA.

BUSCA HEURÍSTICA:
Procura tendenciosa na floresta, visitando apenas as soluções com mais potencial de serem boas, com base em algumas premissas previamente conhecidas.

A rapidez do processo aumenta, mas é possível que a melhor solução entre todas não tenha sido encontrada, pois eventualmente pode ter ficado fora do trajeto percorrido.

Como escolher o melhor atributo?

Problema: Como definir alguma característica sobre os dados que permita definir um critério para identificação do melhor atributo em cada nível da árvore ?

Abordagem baseada na Teoria da Informação

Boa subdivisão:

Produz grupos mais homogêneos com relação ao atributo categórico.

Idéia → Classificação evidencia as linhas gerais que fazem um elemento pertencer a uma determinada classe, o que é facilitado quando se produz agrupamentos mais organizados.

Melhor atributo "split"

Atributo mais informativo sobre a lógica dos dados num determinado contexto.

Como escolher o melhor atributo?

CASCA	COR	TAMANHO	POLPA	RISCO
aspera	marrom	grande	dura	baixo
aspera	verde	grande	dura	baixo
lisa	vermelho	grande	macia	alto
aspera	verde	grande	macia	baixo
aspera	vermelho	pequena	dura	baixo
lisa	vermelho	pequena	dura	baixo
lisa	marrom	pequena	dura	baixo
aspera	verde	pequena	macia	alto
lisa	verde	pequena	dura	alto
aspera	vermelho	grande	dura	baixo
lisa	marrom	grande	macia	baixo
lisa	verde	pequena	macia	alto
aspera	vermelho	pequena	macia	baixo
lisa	vermelho	grande	dura	alto
lisa	vermelho	pequena	dura	baixo
aspera	verde	pequena	dura	alto

Como escolher o melhor atributo?

Cálculo da Entropia ⟶ $- \log_2 p (c_i | a_j)$

"Quantidade de informação" que a_j tem a oferecer sobre a conclusão c_i

$$\text{Entropia} = -\sum_{i=1}^{n} p(c_i | a_j) \log_2 p(c_i | a_j)$$

Quanto menor a Entropia ⟶ Menor a "dúvida"

↓

Maior a informação que a_j traz sobre C

Melhor atributo "split" ⟶ Subconjuntos mais homogêneos (grupos menos "confusos" com relação à classe).

Conceito de Entropia (**Termodinâmica**):

Inversamente proporcional ao grau de informação (valor entre 0 e 1)

Como escolher o melhor atributo?

$$Entropia(S) = -\sum_{i=1}^{n} p_i \; Log_2(p_i)$$

Onde: ***S*** é a distribuição de probabilidade das ***n*** mensagens possíveis;
p_i é a probabilidade de ocorrência da *i*-ésima mensagem

• Para o caso de um classificador construído para um problema com 2 classes possíveis (A e B), um atributo ***x*** vai permitir dividir os dados em tantos subconjuntos ***S*** quantos forem os seus possíveis valores.

• A entropia de cada um desses subconjuntos S_k seria calculada por:

$$Entropia\,(S_k) = -p_A \; Log_2(p_A) - p_B \; Log_2(p_B)$$

P(A)	P(B)	Entropia
0,50	0,50	1,00
0,67	0,33	0,92
1,00	0,00	0,00

Quanto mais uniforme a distribuição, maior o grau de entropia.

Como escolher o melhor atributo?

Para o caso de um atributo ***x*** que possa assumir três valores (por exemplo, valores inteiros entre 1 e 3), três subconjuntos de ***S*** são formados, cada um com seu próprio grau de entropia.

Pode-se avaliar a entropia em ***S*** quando considerado o atributo ***x***, através da média ponderada dos graus de entropia dos subconjuntos gerados (***S_1***, ***S_2*** e ***S_3***, neste exemplo).

$$Entropia\ (x,S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot Entropia\ (S_i)$$

Ganho de Informação (*"**Information Gain**"*) devido a ***x*** na predição da classe é determinada pela redução da entropia original de ***S***.

$$Ganho\ de\ Informação\ (x,S) = Entropia\ (S) - Entropia\ (x,S)$$

Como escolher o melhor atributo?

Entropia de um Atributo A com relação à Classe C

$$-\sum_{j=1}^{m} p(a_j) \sum_{i=1}^{n} p(c_i \mid a_j) \log_2 p(c_i \mid a_j)$$

Atributo com **MENOR** entropia é o **MELHOR** para determinar a Classe

CASCA	COR	TAMANHO	POLPA	RISCO
aspera	marrom	grande	dura	baixo
aspera	verde	grande	dura	baixo
lisa	vermelho	grande	macia	alto
aspera	verde	grande	macia	baixo
aspera	vermelho	pequena	dura	baixo
lisa	vermelho	pequena	dura	baixo
lisa	marrom	pequena	dura	baixo
aspera	verde	pequena	macia	alto
lisa	verde	pequena	dura	alto
aspera	vermelho	grande	dura	baixo
lisa	marrom	grande	macia	baixo
lisa	verde	pequena	macia	alto
aspera	vermelho	pequena	macia	baixo
lisa	vermelho	grande	dura	alto
lisa	vermelho	pequena	dura	baixo
aspera	verde	pequena	dura	alto

Consideremos o Atributo "**Casca**"

p(baixo | aspera) = 6 / 8
p(alto | aspera) = 2 / 8
p(aspera) = 8 / 16

p(baixo | lisa) = 4 / 8
p(alto | lisa) = 4 / 8
p(lisa) = 8 / 16

Entropia para o Atributo "**Casca**"

$$-\frac{8}{16}\left(\frac{6}{8}\log\left(\frac{6}{8}\right)+\frac{2}{8}\log\left(\frac{2}{8}\right)\right)+\frac{8}{16}\left(\frac{4}{8}\log\left(\frac{4}{8}\right)+\frac{4}{8}\log\left(\frac{4}{8}\right)\right)$$

0.90564

CASCA	COR	TAMANHO	POLPA	RISCO
aspera	marrom	grande	dura	baixo
aspera	verde	grande	dura	baixo
lisa	vermelho	grande	macia	alto
aspera	verde	grande	macia	baixo
aspera	vermelho	pequena	dura	baixo
lisa	vermelho	pequena	dura	baixo
lisa	marrom	pequena	dura	baixo
aspera	verde	pequena	macia	alto
lisa	verde	pequena	dura	alto
aspera	vermelho	grande	dura	baixo
lisa	marrom	grande	macia	baixo
lisa	verde	pequena	macia	alto
aspera	vermelho	pequena	macia	baixo
lisa	vermelho	grande	dura	alto
lisa	vermelho	pequena	dura	baixo
aspera	verde	pequena	dura	alto

Consideremos o Atributo "**Cor**"

p(baixo | marrom) = 3 / 3
p(alto | marrom) = 0 / 3
p(marrom) = 3 / 16

p(baixo | verde) = 2 / 6
p(alto | verde) = 4 / 6
p(verde) = 6 / 16

p(baixo | vermelho) = 5 / 7
p(alto | vermelho) = 2 / 7
p(vermelho) = 7 / 16

Entropia para o Atributo "**Cor**"

$$\frac{3}{16}\left(\frac{3}{3}\log\left(\frac{3}{3}\right)+\frac{0}{3}\log\left(\frac{0}{3}\right)\right)+\frac{6}{16}\left(\frac{2}{6}\log\left(\frac{2}{6}\right)+\frac{4}{6}\log\left(\frac{4}{6}\right)\right)+\frac{7}{16}\left(\frac{5}{7}\log\left(\frac{5}{7}\right)+\frac{2}{7}\log\left(\frac{2}{7}\right)\right)$$

=

0.721976

CASCA	COR	TAMANHO	POLPA	RISCO
aspera	marrom	grande	dura	baixo
aspera	verde	grande	dura	baixo
lisa	vermelho	grande	macia	alto
aspera	verde	grande	macia	baixo
aspera	vermelho	pequena	dura	baixo
lisa	vermelho	pequena	dura	baixo
lisa	marrom	pequena	dura	baixo
aspera	verde	pequena	macia	alto
lisa	verde	pequena	dura	alto
aspera	vermelho	grande	dura	baixo
lisa	marrom	grande	macia	baixo
lisa	verde	pequena	macia	alto
aspera	vermelho	pequena	macia	baixo
lisa	vermelho	grande	dura	alto
lisa	vermelho	pequena	dura	baixo
aspera	verde	pequena	dura	alto

Consideremos o Atributo "**Tamanho**"

p(baixo | grande) = 5 / 7
p(alto | grande) = 2 / 7
p(grande) = 7 / 16

p(baixo | pequeno) = 5 / 9
p(alto | pequeno) = 4 / 9
p(pequeno) = 9 / 16

Entropia para o Atributo "**Tamanho**"

$$\frac{7}{16}\left(\frac{5}{7}\log\left(\frac{5}{7}\right)+\frac{2}{7}\log\left(\frac{2}{7}\right)\right)+\frac{9}{16}\left(\frac{5}{9}\log\left(\frac{5}{9}\right)+\frac{4}{9}\log\left(\frac{4}{9}\right)\right)$$

=

0.9350955

CASCA	COR	TAMANHO	POLPA	RISCO
aspera	marrom	grande	dura	baixo
aspera	verde	grande	dura	baixo
lisa	vermelho	grande	macia	alto
aspera	verde	grande	macia	baixo
aspera	vermelho	pequena	dura	baixo
lisa	vermelho	pequena	dura	baixo
lisa	marrom	pequena	dura	baixo
aspera	verde	pequena	macia	alto
lisa	verde	pequena	dura	alto
aspera	vermelho	grande	dura	baixo
lisa	marrom	grande	macia	baixo
lisa	verde	pequena	macia	alto
aspera	vermelho	pequena	macia	baixo
lisa	vermelho	grande	dura	alto
lisa	vermelho	pequena	dura	baixo
aspera	verde	pequena	dura	alto

Consideremos o Atributo "**Polpa**"

p(baixo | dura) = 7 / 10
p(alto | dura) = 3 / 10
p(dura) = 10 / 16

p(baixo | macia) = 3 / 6
p(alto | macia) = 3 / 6
p(macia) = 6 / 16

Entropia para o Atributo "**Polpa**"

$$\frac{10}{16}\left(\frac{7}{10}\log\left(\frac{7}{10}\right)+\frac{3}{10}\log\left(\frac{3}{10}\right)\right)+\frac{6}{16}\left(\frac{3}{6}\log\left(\frac{3}{6}\right)+\frac{3}{6}\log\left(\frac{3}{6}\right)\right)$$

=

0.92581

Resultados do cálculo da entropia

Atributo	Entropia
Casca	0.90564
Cor	0.721976
Tamanho	0.9350955
Polpa	0.92581

Como **"Cor"** tem a **menor entropia**, pode-se afirmar que também tem o maior ganho de informação. Logo deve ser usado como **"atributo split"**.

Análise dos Resultados

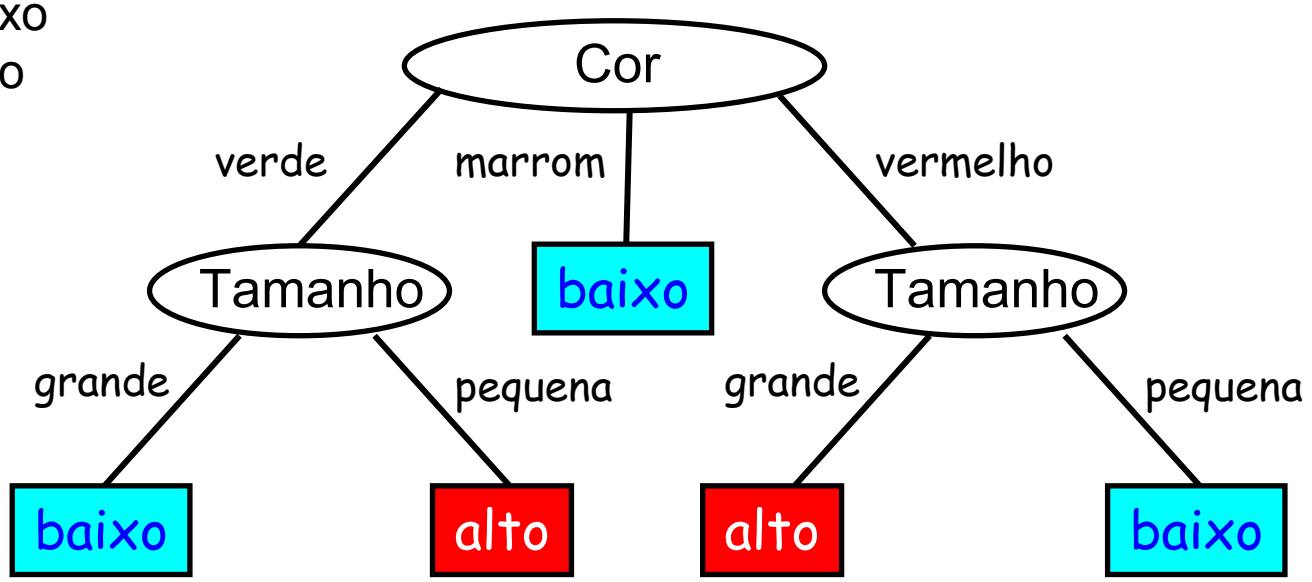
Resultado WEKA

```
=== Confusion Matrix ===
 a b   <-- classified as
 9 1 | a = baixo
 0 6 | b = alto
```

Acurácia: 93.75 %

TP Rate	Class
0.9	baixo
1	alto

```
Cor = marrom: baixo  (3.0)
Cor = verde
|   Tamanho = grande: baixo  (2.0)
|   Tamanho = pequena: alto  (4.0)
Cor = vermelho
|   Tamanho = grande: alto  (3.0/1.0)
|   Tamanho = pequena: baixo  (4.0)
```



Árvores de decisão: prós e contras

- **Vantagens**
  - Custo computacional é baixo.
  - Muito rápido para classificar amostras desconhecidas.
  - Fácil de interpretar árvores de tamanho pequeno.
  - Precisão é semelhante a de outros métodos de classificação, para muitos datasets simples.
- **Desvantagens**
  - "**Overfitting**" resulta em árvores de decisão que são mais complexas do que necessárias.
  - O treinamento do erro nem sempre produz uma boa estimativa com relação à execução da árvore para amostras desconhecidas.
  - Necessita de novas maneiras para estimar erros.

Exercício: Construir árvores de decisão usando J48, PART, JRIP, etc.

Estágio	Deficiência	Astigmatismo	Produção_Lágrima	Recomenda_Lente
Inicial	Miopia	Não	Baixa	Nenhuma
Inicial	Miopia	Não	Normal	Macia
Inicial	Miopia	Sim	Baixa	Nenhuma
Inicial	Miopia	Sim	Normal	Dura
Inicial	Hipermetropia	Não	Baixa	Nenhuma
Inicial	Hipermetropia	Não	Normal	Macia
Inicial	Hipermetropia	Sim	Baixa	Nenhuma
Inicial	Hipermetropia	Sim	Normal	Dura
Pré-Presbiopia	Miopia	Não	Baixa	Nenhuma
Pré-Presbiopia	Miopia	Não	Normal	Macia
Pré-Presbiopia	Miopia	Sim	Baixa	Nenhuma
Pré-Presbiopia	Miopia	Sim	Normal	Dura
Pré-Presbiopia	Hipermetropia	Não	Baixa	Nenhuma
Pré-Presbiopia	Hipermetropia	Não	Normal	Macia
Pré-Presbiopia	Hipermetropia	Sim	Baixa	Nenhuma
Pré-Presbiopia	Hipermetropia	Sim	Normal	Nenhuma
Presbiopia	Miopia	Não	Baixa	Nenhuma
Presbiopia	Miopia	Não	Normal	Nenhuma
Presbiopia	Miopia	Sim	Baixa	Nenhuma
Presbiopia	Miopia	Sim	Normal	Dura
Presbiopia	Hipermetropia	Não	Baixa	Nenhuma
Presbiopia	Hipermetropia	Não	Normal	Macia
Presbiopia	Hipermetropia	Sim	Baixa	Nenhuma
Presbiopia	Hipermetropia	Sim	Normal	Nenhuma