



Unidade 19 – Conceitos de Mineração de Dados – Parte 3



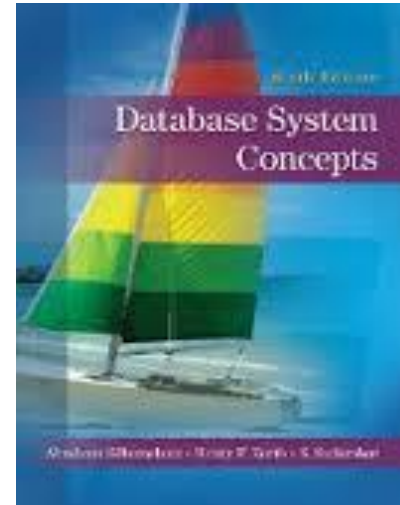
Prof. Aparecido V. de Freitas
Doutor em Engenharia
da Computação pela EPUSP



Bibliografia



Sistemas de Banco de Dados
Elmasri / Navathe 6ª edição



Sistema de Banco de Dados
Korth, Silberschatz – Sixth Edition



Tipos de Conhecimentos descobertos pela Mineração de Dados

- Ⓔ A mineração de dados enfoca o Conhecimento Indutivo, que descobre novas regras e padrões com base nos **dados fornecidos**;
- Ⓔ É comum descrever-se o conhecimento descoberto durante a Mineração de Dados por:
 - ✓ **Regras de Associação;**
 - ✓ **Classificação;**
 - ✓ **Padrões Sequenciais;**
 - ✓ **Padrões dentro de séries temporais;**
 - ✓ **Agrupamento.**



Classificação



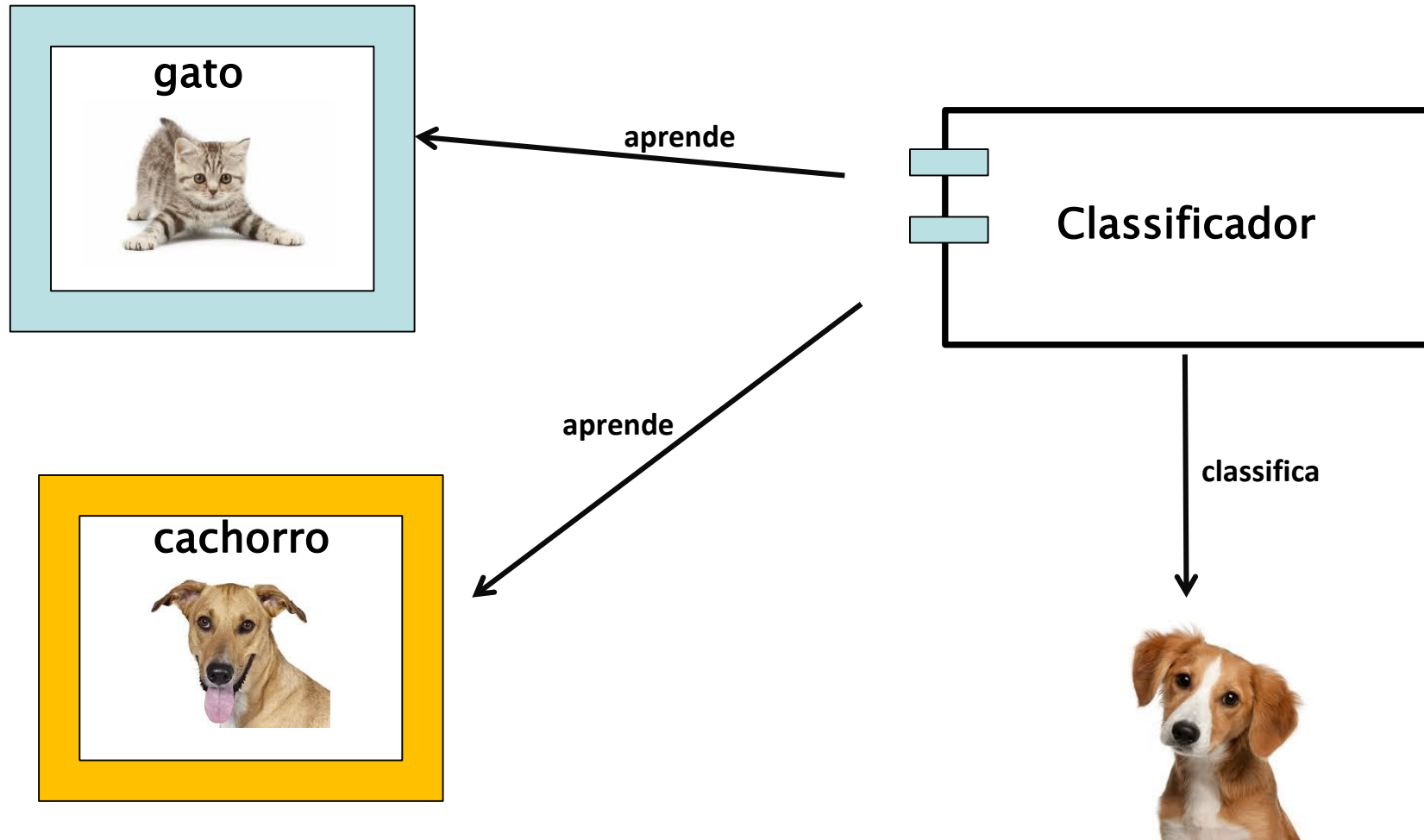
- Ⓜ É o processo que descreve diferentes classes de dados;
- Ⓜ Por exemplo, em uma aplicação bancária, os clientes que solicitam um cartão de crédito podem ser classificados como risco fraco, risco médio ou risco bom;
- Ⓜ É uma das técnicas mais utilizadas em mineração de dados;
- Ⓜ Classificar um objeto consiste em se determinar com que grupo de entidades, já classificadas anteriormente, esse objeto apresenta mais semelhança;
- Ⓜ Logo, esse tipo de atividade também é chamada Aprendizado Supervisionado.

pentaho
open source business intelligence



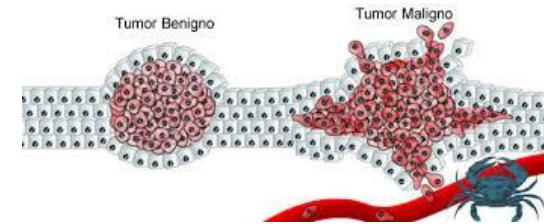


Aprendizado Supervisionado

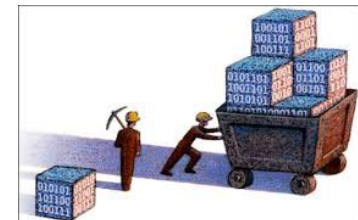




Exemplos – Classificação



- Ⓢ Predizer se um tumor é benigno ou maligno;
- Ⓢ Classificar transações de cartões de crédito como legítimas ou fraudulentas;
- Ⓢ Analisar concessão de empréstimos bancários em Instituições financeiras;
- Ⓢ Filtrar (marcar) e-mails que seriam spams em softwares de correio eletrônico.





Classificação – Procedimento



- Ⓢ O primeiro passo – aprendizado do modelo – é realizado com um conjunto de dados que já foram classificados;
- Ⓢ Cada registro nos dados de treinamento, contém um atributo chamado rótulo de classe (**label**) que indica a que classe o registro pertence;
- Ⓢ O modelo que é produzido costuma estar na forma de árvore de decisão;

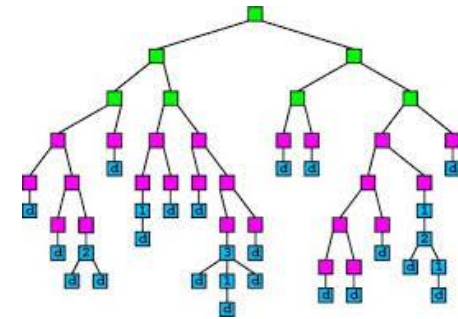




Árvore de Decisão



- © É uma representação gráfica da descrição de cada classe ou, em outras palavras, uma representação das regras de classificação





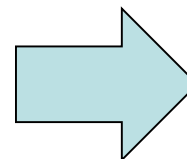
Exemplo – Árvore de Decisão

classe



Id	Casa própria	EstCivil	Rendim.	Mau Pagador
1	S	Solteiro	125K	NÃO
2	N	Casado	100K	NÃO
3	N	Solteiro	70K	NÃO
4	S	Casado	120K	NÃO
5	N	Divorc.	95K	SIM
6	N	Casado	60K	NÃO
7	S	Divorc.	220K	NÃO
8	N	Solteiro	85K	SIM
9	N	Casado	75K	NÃO
10	N	Solteiro	90K	SIM

Set Training



Learning
Algorithm



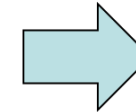
Algoritmo de Aprendizado



classe



	Casa			
Id	própria	EstCivil	Rendim.	Mau Pagador
1	S	Solteiro	125K	NÃO
2	N	Casado	100K	NÃO
3	N	Solteiro	70K	NÃO
4	S	Casado	120K	NÃO
5	N	Divorc.	95K	SIM
6	N	Casado	60K	NÃO
7	S	Divorc.	220K	NÃO
8	N	Solteiro	85K	SIM
9	N	Casado	75K	NÃO
10	N	Solteiro	90K	SIM



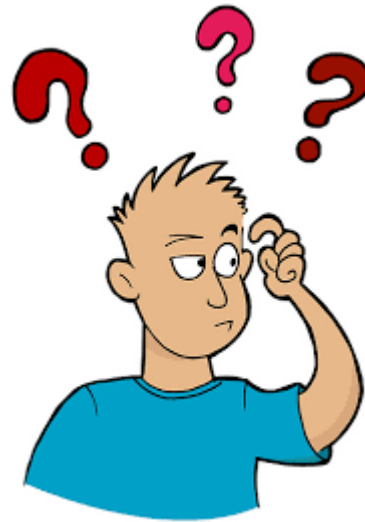
Learning
Algorithm

Set Training

- @ Uma coleção de registros (**set Training**), ou conjunto de dados de treinamento, é submetida ao algoritmo de aprendizado (**Learning Algorithm**);



O que o algoritmo de aprendizado faz com os dados de treinamento?

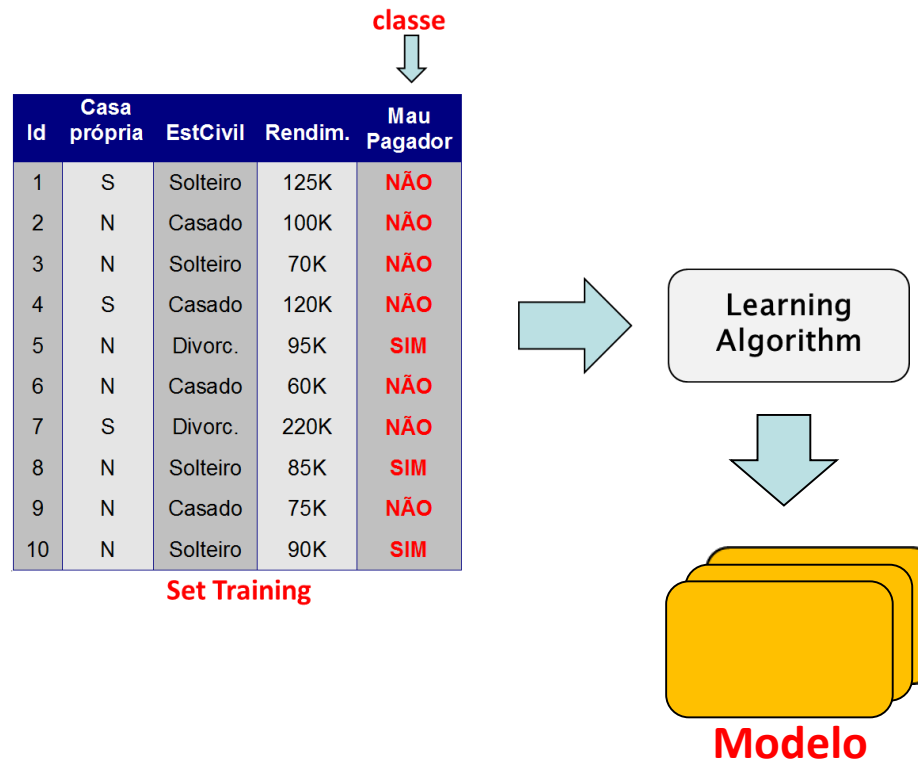




Algoritmo de Aprendizado



- ② O algoritmo de aprendizado procura encontrar um modelo para determinar o valor do atributo classe em função dos valores dos outros atributos;
- ② Esse modelo, costuma ser produzido na forma de uma árvore de decisão;





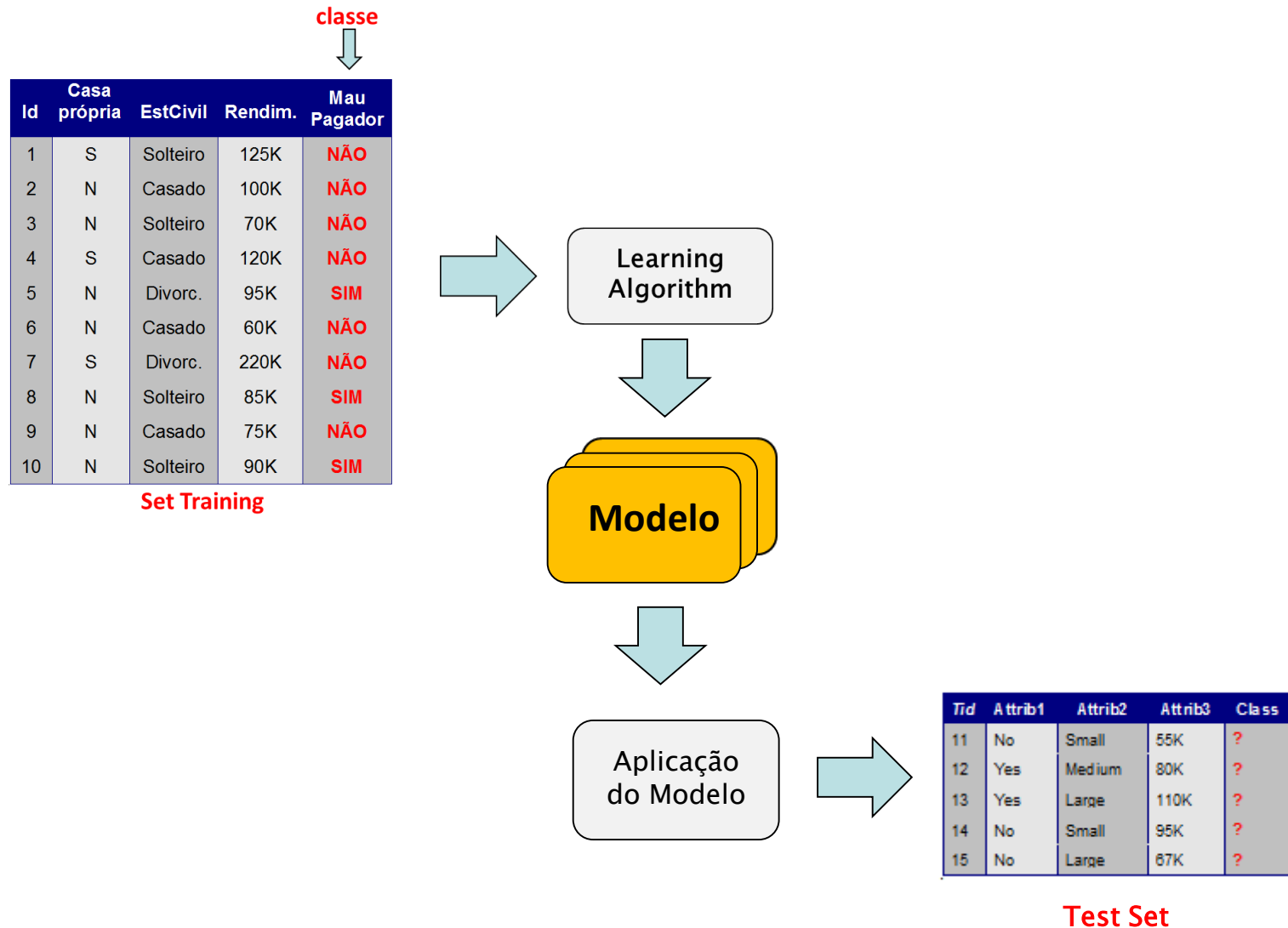
Para que serve o modelo gerado pelo algoritmo de aprendizado?





Uso do modelo

- Ⓐ A partir do modelo, pode-se usá-lo para se classificar novos dados.





As árvores de decisão são representações do modelo que consistem em:

-
- ```

graph TD
 A[is sex male?] -- yes --> B[is age > 9.5?]
 A -- no --> C[survived
0.73 36%]
 B -- died --> D[died
0.17 61%]
 B -- survived --> E[is sibsp > 2.5?]
 E -- died --> F[died
0.05 2%]
 E -- survived --> G[survived
0.89 2%]

```



# Modelo – Árvore de Decisão

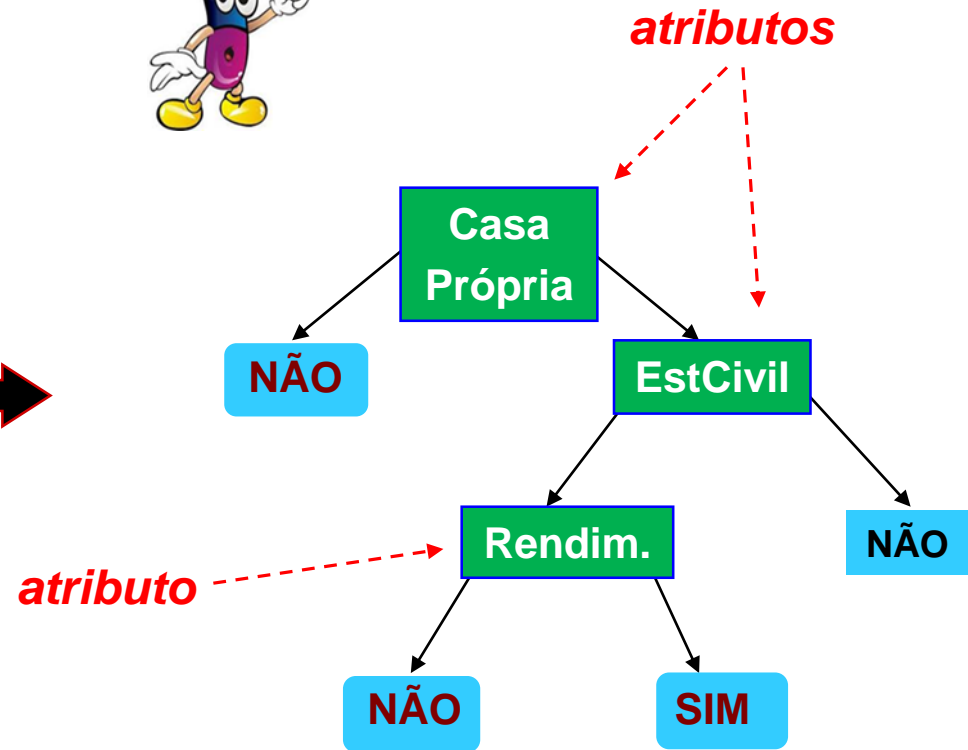


*classe*

| Id | Casa própria | EstCivil | Rendim. | Mau Pagador |
|----|--------------|----------|---------|-------------|
| 1  | S            | Solteiro | 125K    | NÃO         |
| 2  | N            | Casado   | 100K    | NÃO         |
| 3  | N            | Solteiro | 70K     | NÃO         |
| 4  | S            | Casado   | 120K    | NÃO         |
| 5  | N            | Divorc.  | 95K     | SIM         |
| 6  | N            | Casado   | 60K     | NÃO         |
| 7  | S            | Divorc.  | 220K    | NÃO         |
| 8  | N            | Solteiro | 85K     | SIM         |
| 9  | N            | Casado   | 75K     | NÃO         |
| 10 | N            | Solteiro | 90K     | SIM         |

Dados de treinamento

✓ Nós internos representam os atributos;



Modelo: árvore de decisão



# Modelo – Árvore de Decisão

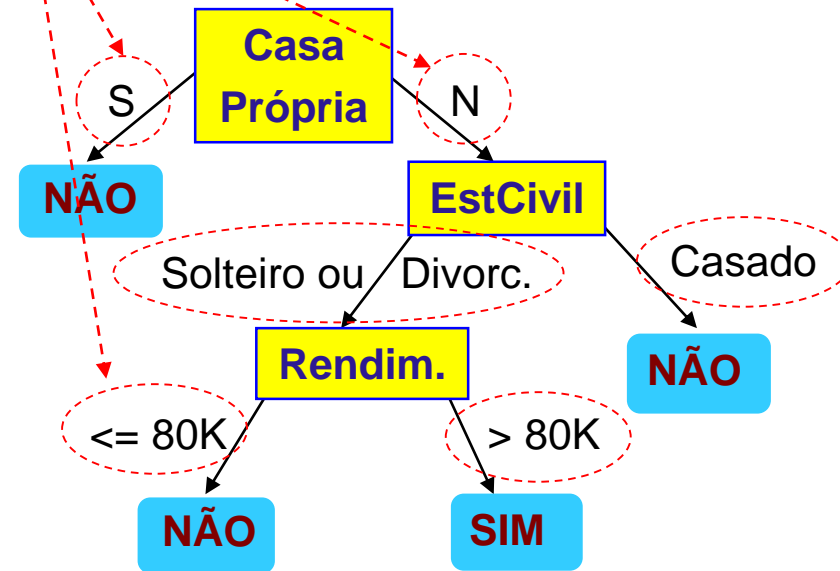


categorico categorico contínuo classe

| Id | Casa própria | EstCivil | Rendim. | Mau Pagador |
|----|--------------|----------|---------|-------------|
| 1  | S            | Solteiro | 125K    | NÃO         |
| 2  | N            | Casado   | 100K    | NÃO         |
| 3  | N            | Solteiro | 70K     | NÃO         |
| 4  | S            | Casado   | 120K    | NÃO         |
| 5  | N            | Divorc.  | 95K     | SIM         |
| 6  | N            | Casado   | 60K     | NÃO         |
| 7  | S            | Divorc.  | 220K    | NÃO         |
| 8  | N            | Solteiro | 85K     | SIM         |
| 9  | N            | Casado   | 75K     | NÃO         |
| 10 | N            | Solteiro | 90K     | SIM         |

Dados de treinamento

✓ Arestas representam os valores dos atributos;



Modelo: árvore de decisão



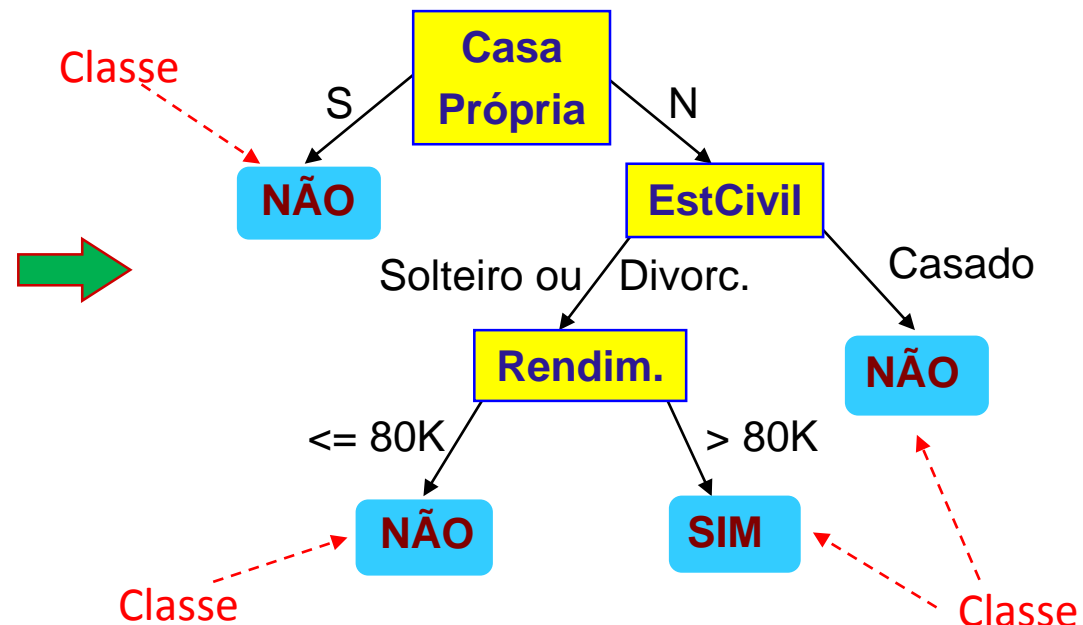
# Modelo – Árvore de Decisão



|    | categórico   | categórico | contínuo | classe      |
|----|--------------|------------|----------|-------------|
|    | Casa própria | EstCivil   | Rendim.  | Mau Pagador |
| 1  | S            | Solteiro   | 125K     | NÃO         |
| 2  | N            | Casado     | 100K     | NÃO         |
| 3  | N            | Solteiro   | 70K      | NÃO         |
| 4  | S            | Casado     | 120K     | NÃO         |
| 5  | N            | Divorc.    | 95K      | SIM         |
| 6  | N            | Casado     | 60K      | NÃO         |
| 7  | S            | Divorc.    | 220K     | NÃO         |
| 8  | N            | Solteiro   | 85K      | SIM         |
| 9  | N            | Casado     | 75K      | NÃO         |
| 10 | N            | Solteiro   | 90K      | SIM         |

Dados de treinamento

✓ Nós folha designam a classificação;



Modelo: árvore de decisão



Pode haver mais de uma árvore de decisão?

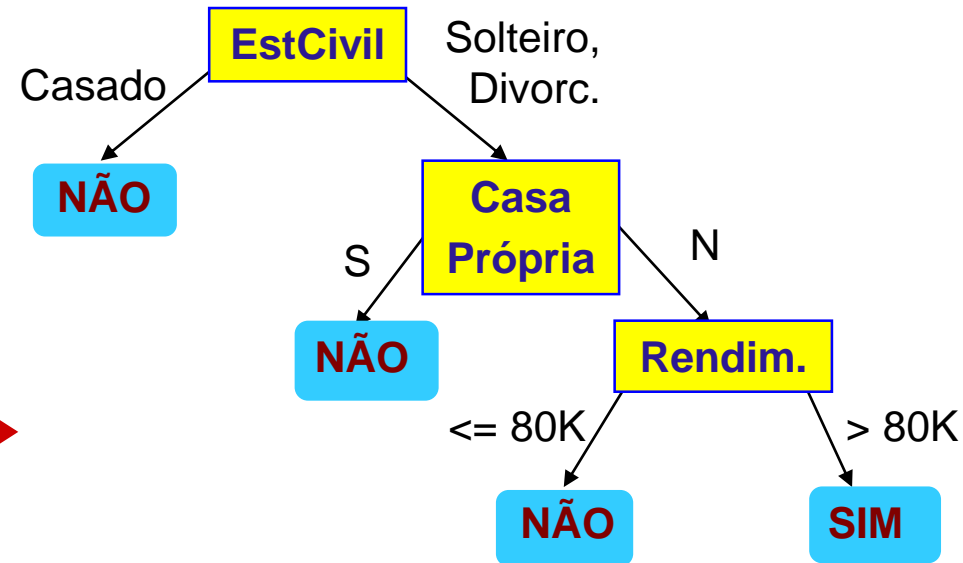




## Outra Árvore de Decisão

classe

| Id | Casa própria | EstCivil | Rendim. | Mau Pagador |
|----|--------------|----------|---------|-------------|
| 1  | S            | Solteiro | 125K    | NÃO         |
| 2  | N            | Casado   | 100K    | NÃO         |
| 3  | N            | Solteiro | 70K     | NÃO         |
| 4  | S            | Casado   | 120K    | NÃO         |
| 5  | N            | Divorc.  | 95K     | SIM         |
| 6  | N            | Casado   | 60K     | NÃO         |
| 7  | S            | Divorc.  | 220K    | NÃO         |
| 8  | N            | Solteiro | 85K     | SIM         |
| 9  | N            | Casado   | 75K     | NÃO         |
| 10 | N            | Solteiro | 90K     | SIM         |



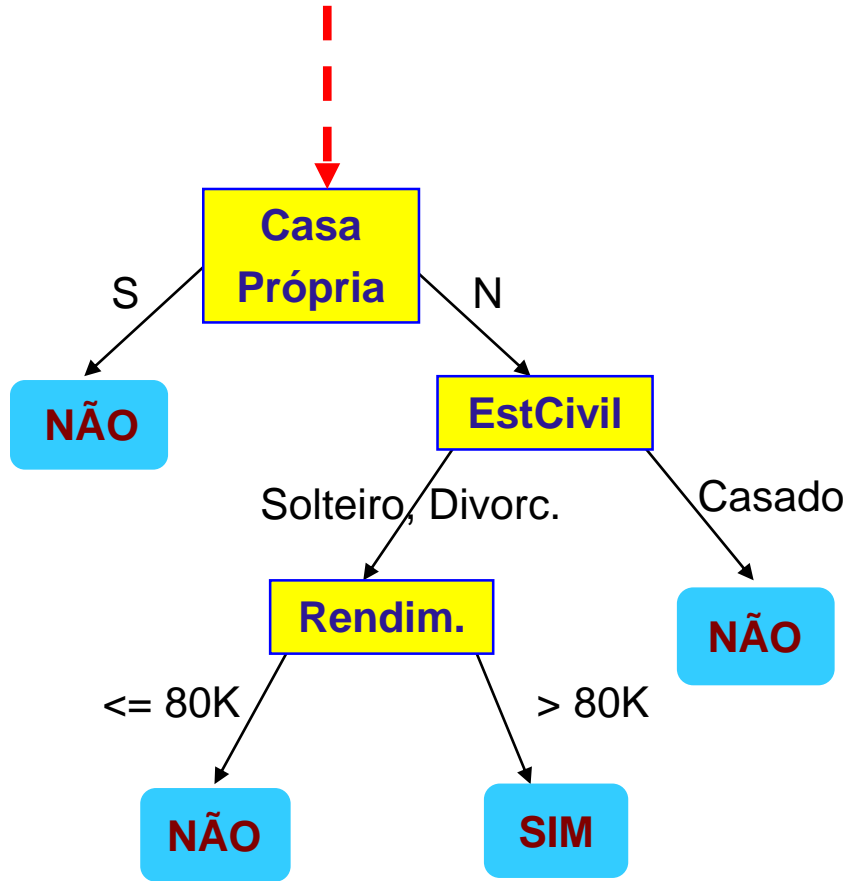
Pode haver mais de uma árvore para o mesmo conjunto de dados!!!





## Aplicando o modelo nos dados de teste

Início pela raiz da árvore !



Dado para teste

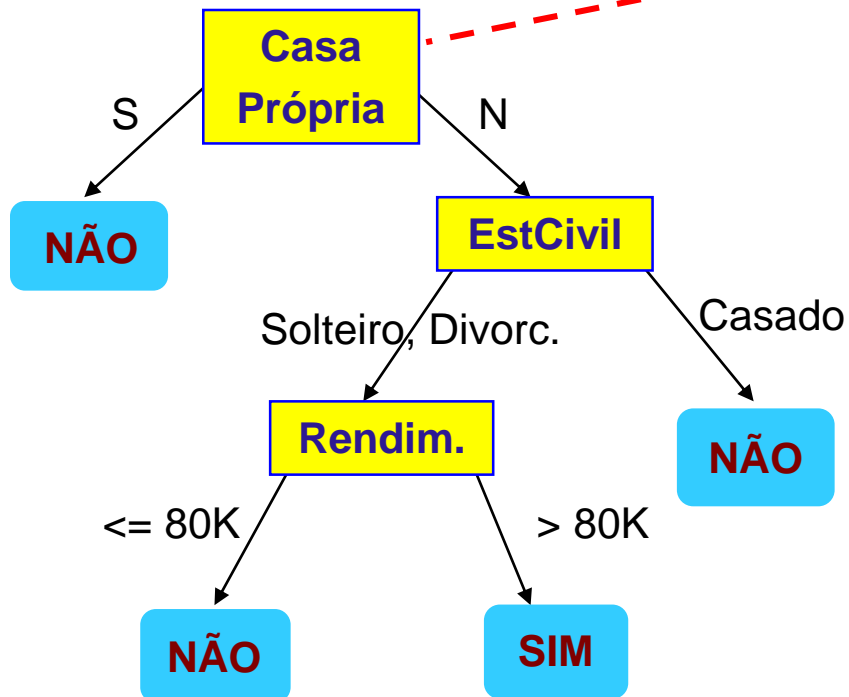
| Casa Própria | Estado Civil | Rendim. | Mau pagador |
|--------------|--------------|---------|-------------|
| N            | Casado       | 80K     | ?           |



## Aplicando o modelo nos dados de teste

Dado para teste

| Casa Própria | Estado Civil | Rendim. | Mau pagador |
|--------------|--------------|---------|-------------|
| N            | Casado       | 80K     | ?           |

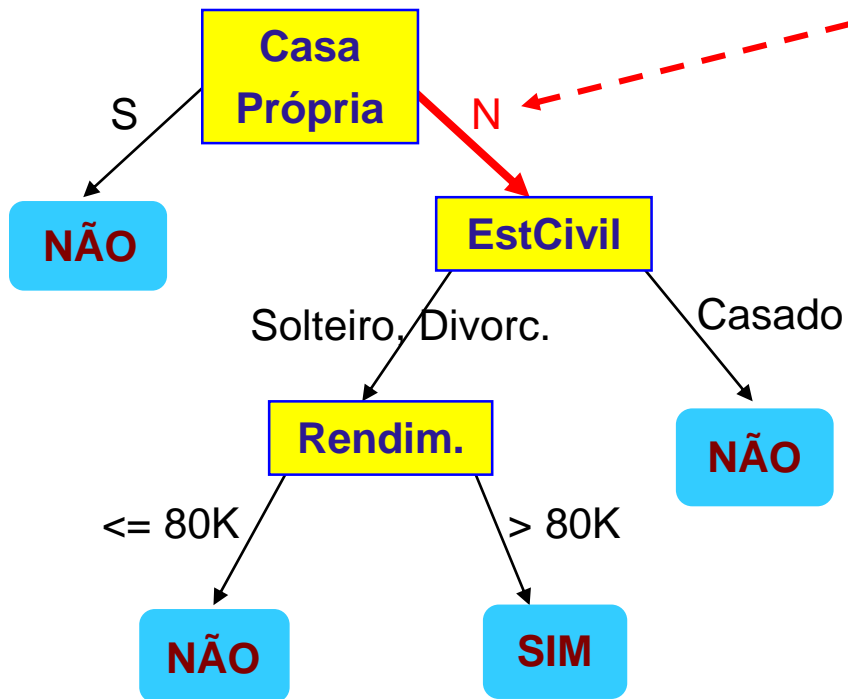




## Aplicando o modelo nos dados de teste

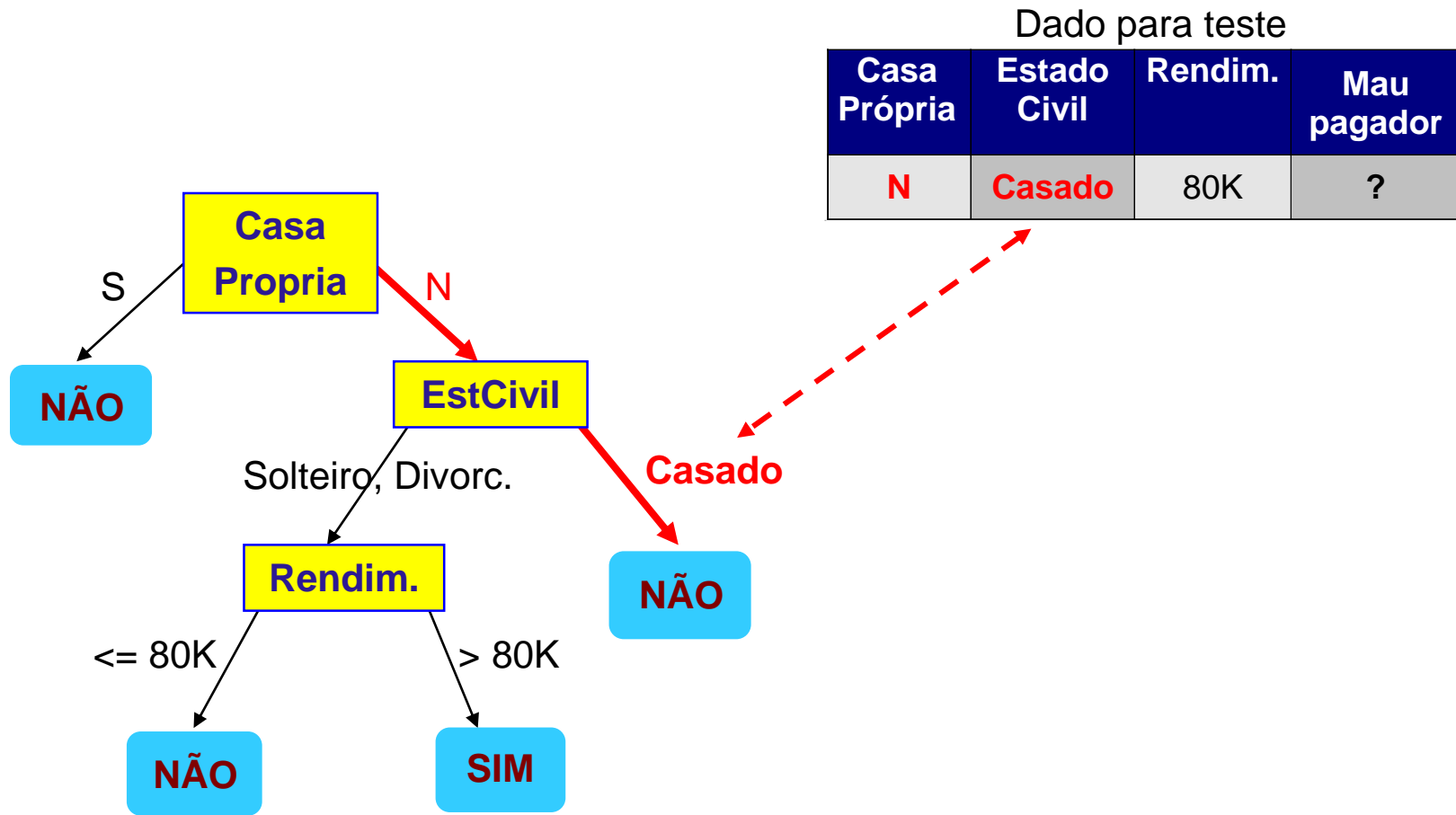
Dado para teste

| Casa Própria | Estado Civil | Rendim. | Mau pagador |
|--------------|--------------|---------|-------------|
| N            | Casado       | 80K     | ?           |





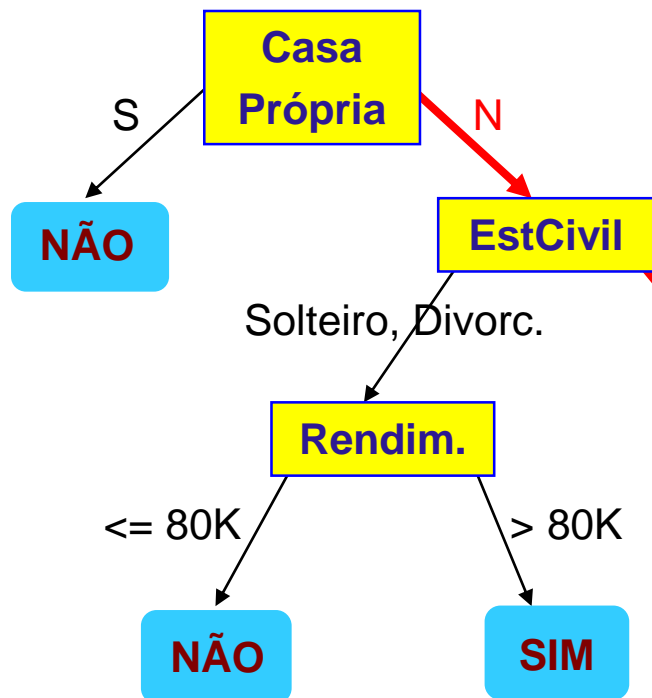
## Aplicando o modelo nos dados de teste







## Aplicando o modelo nos dados de teste



Dado para teste

| Casa Própria | Estado Civil | Rendim. | Mau pagador |
|--------------|--------------|---------|-------------|
| N            | Casado       | 80K     | ?           |

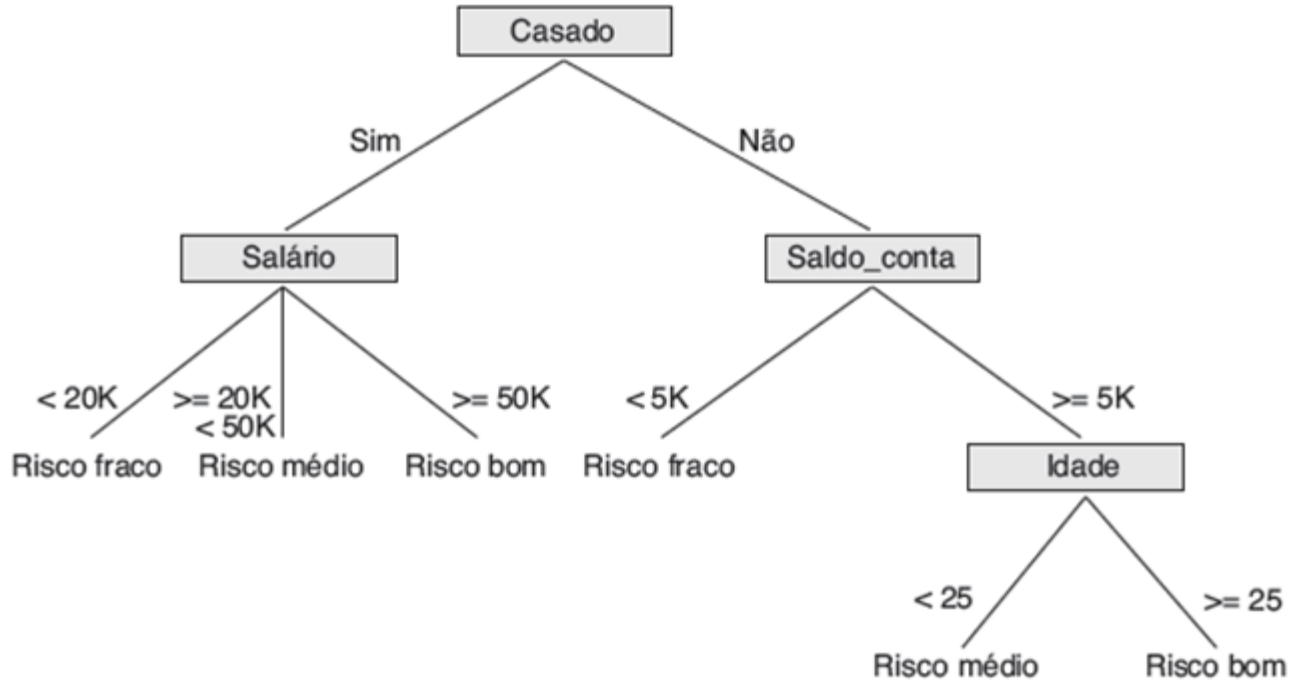
**NÃO**

Será atribuído à classe Mau pagador o valor **NÃO**.





## Exemplo – Árvore de Decisão – Elmasri

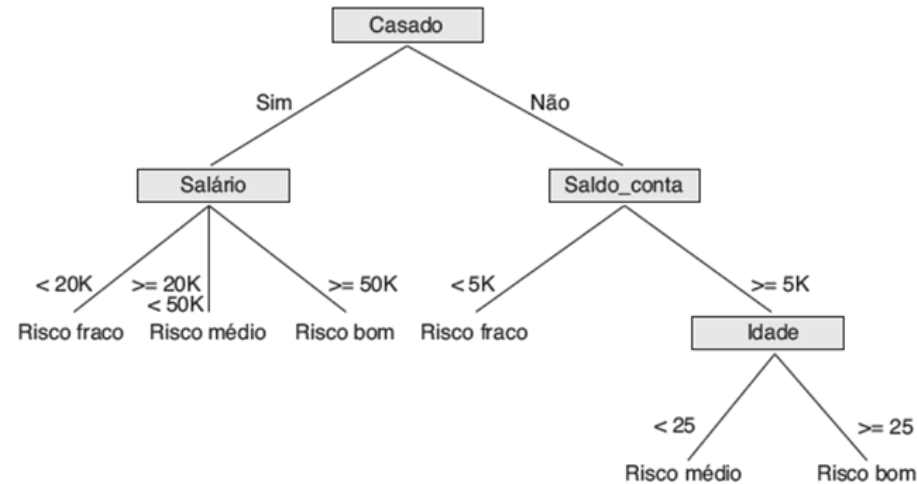


Árvore de decisão da amostra para aplicações de cartão de crédito





## Exemplo – Árvore de Decisão – Elmasri



- ✓ A árvore mostra que se um **cliente** for **casado** e se **salário**  **$\geq 50K$** , então ele tem um **risco bom** para um cartão de crédito bancário;
- ✓ Essa é uma das **regras** que descrevem o **risco bom**;
- ✓ A travessia da árvore a partir da raiz para algum nó folha forma outras regras para essa e outras classes.





Como uma árvore de decisão é criada?





## Geração da Árvore de Decisão



- ⊕ Os algoritmos mais conhecidos para a geração da árvore de decisão são **ID3** (Quinlan, 1986) e **C4.5** (Quinlan, 1993);
- ⊕ **John Ross Quinlan** é um pesquisador na área de Data Mining e Teoria da Decisão;





## Algoritmo ID3

- ⊕ O algoritmo recebe como entrada um conjunto de dados para treinamento;
- ⊕ Por meio do algoritmo, constrói-se a árvore de decisão em uma abordagem top-down considerando a questão: “Qual atributo é o mais importante e, portanto, que deve ser colocado na raiz da árvore?”;



- ⊕ Para isso cada atributo é testado e sua capacidade para se tornar nó raiz é avaliada;
- ⊕ Cria-se tantos nós filhos da raiz quantos valores possíveis esse atributo puder assumir;
- ⊕ Repete-se o processo para cada nó filho da raiz e assim, sucessivamente.

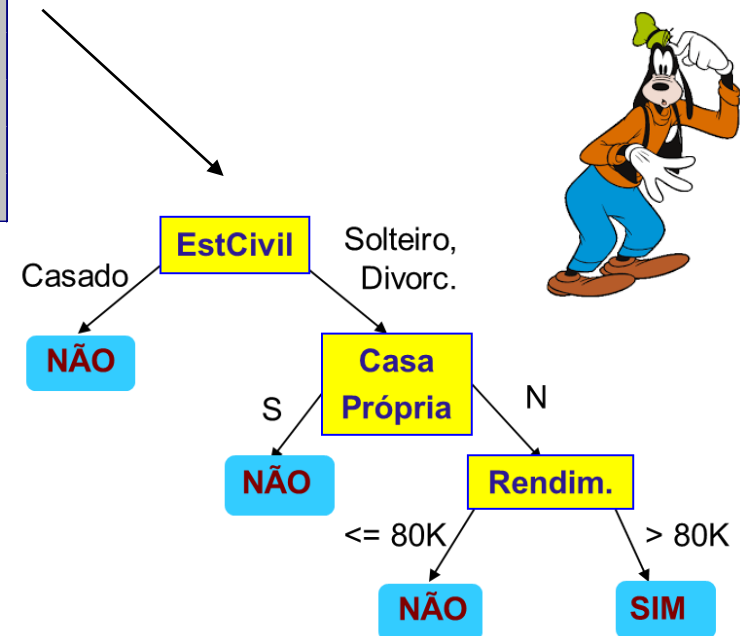
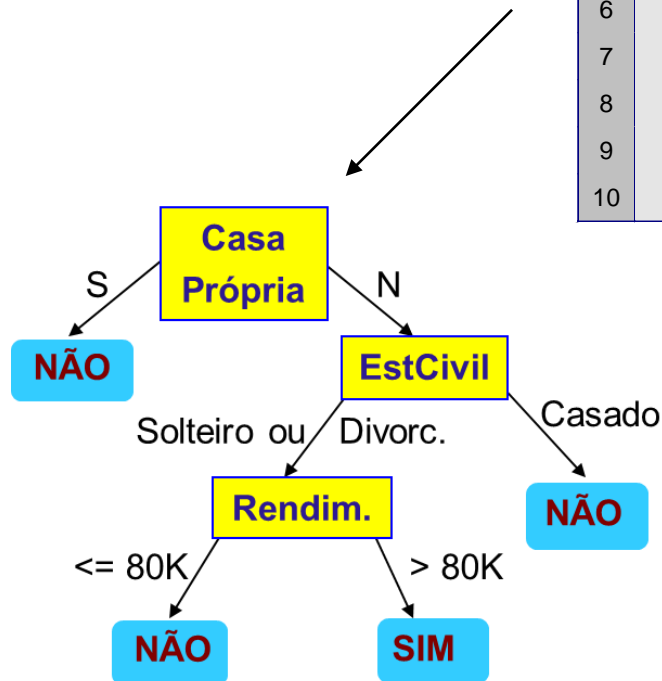


# Algoritmo ID3 - Observação

⊕ “O atributo mais importante deverá ser colocado na raiz da árvore”.



| Id | Casa própria | EstCivil | Rendim. | Mau Pagador |
|----|--------------|----------|---------|-------------|
| 1  | S            | Solteiro | 125K    | NÃO         |
| 2  | N            | Casado   | 100K    | NÃO         |
| 3  | N            | Solteiro | 70K     | NÃO         |
| 4  | S            | Casado   | 120K    | NÃO         |
| 5  | N            | Divorc.  | 95K     | SIM         |
| 6  | N            | Casado   | 60K     | NÃO         |
| 7  | S            | Divorc.  | 220K    | NÃO         |
| 8  | N            | Solteiro | 85K     | SIM         |
| 9  | N            | Casado   | 75K     | NÃO         |
| 10 | N            | Solteiro | 90K     | SIM         |



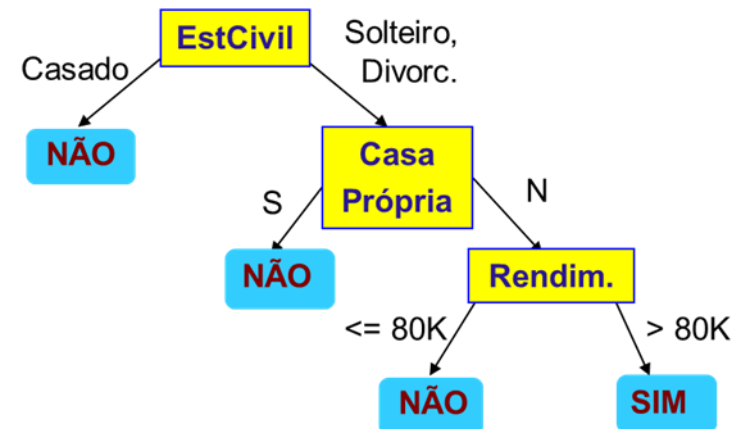
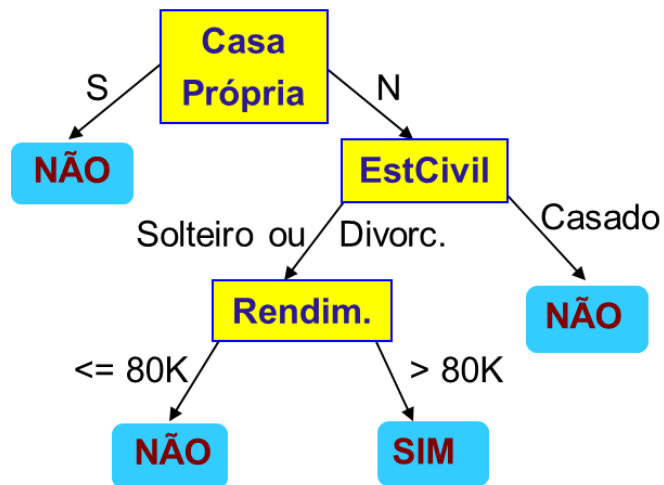
A raiz será Casa Própria ou EstCivil?





Como definir o atributo que será usado como raiz da Árvore de Decisão?

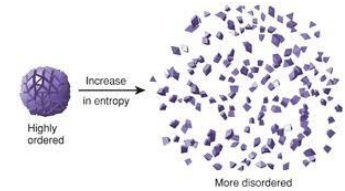
No exemplo, escolho CasaPrópria ou EstCivil?







## Como definir o atributo mais adequado?



- ⊕ O algoritmo ID3 utiliza a medida de Ganho de Informação;
- ⊕ Para definir **Ganho de Informação**, será necessário estudar-se o conceito de **Entropia**.
- ⊕ Entropia é um conceito da Termodinâmica usado para se determinar a quantidade de energia útil de um determinado sistema;
- ⊕ Entropia está associada à medida da desordem das partículas de um sistema físico;
- ⊕ De acordo com a Lei da Termodinâmica, quanto maior for a **desordem** de um sistema, maior será a sua entropia (medida da incerteza).



## Entropia no Algoritmo ID3

- ⊕ O conceito de Entropia foi incorporado no algoritmo ID3;
- ⊕ Considere-se uma coleção  $S$  de instâncias, com duas classes distintas (por exemplo, maupagador = **sim** e maupagador = **não**, a Entropia será calculada por:

$$E(S) = - p_{\text{sim}} \times \log_2 p_{\text{sim}} - p_{\text{nao}} \times \log_2 p_{\text{nao}}$$

- ⊕  $E(S)$  – Entropia corresponde à informação necessária para classificar os dados de treinamento de  $S$  amostras (instância);
- ⊕  $p_{\text{sim}}$  é a probabilidade de que uma amostra aleatória pertença à classe maupagador = **sim**;
- ⊕  $p_{\text{nao}}$  é a probabilidade de que uma amostra aleatória pertença à classe maupagador = **não**.



## Entropia no Algoritmo ID3

⊕ Para ilustrar, considere um conjunto  $S$  com 14 registros de algum conceito booleano:

- ✓ 4 positivos
- ✓ 9 positivos

⊕ Logo, a Entropia desse conjunto é dada por:

$$E(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$



## Entropia no Algoritmo ID3

- Em outros casos, note:

- Para [7+, 7-]

$$E(S) = -\frac{7}{14} \log_2 \frac{7}{14} - \frac{7}{14} \log_2 \frac{7}{14} = 0.99 \dots \approx 1 \quad \leftarrow \text{Incerteza !}$$

- Para [0+, 14-] ou [14+, 0-]

$$E(S) = -\frac{14}{14} \log_2 \frac{14}{14} = 0 \quad \leftarrow \text{Certeza !}$$

- Entropia mede o nível de certeza que temos sobre um evento



## Cálculo da Entropia – Exemplo

- ⊕ Considere a coleção de instâncias (amostra de treinamento) definida por 6 registros;
- ⊕ Cada instância tem 5 atributos (**Casado**, **Salário**, **Saldo\_conta**, **Idade** e **Emprestar**);
- ⊕ **RID** é um identificador de registros e, portanto, **não** fará parte da árvore de decisão;
- ⊕ O atributo **Emprestar** será o designador da classificação (rótulo de classe) e conterà duas classes: Emprestar = **não** e Emprestar = **sim** (2 Classes).

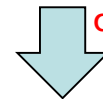
| RID | Casado | Salário     | Saldo_conta | Idade | Emprestar |
|-----|--------|-------------|-------------|-------|-----------|
| 1   | não    | >=50K       | <5K         | >=25  | sim       |
| 2   | sim    | >=50K       | >=5K        | >=25  | sim       |
| 3   | sim    | 20K. . .50K | <5K         | <25   | não       |
| 4   | não    | <20K        | >=5K        | <25   | não       |
| 5   | não    | <20K        | <5K         | >=25  | não       |
| 6   | sim    | 20K. . .50K | >=5K        | >=25  | sim       |



## Cálculo da Entropia da Amostra – $E(S)$

- ⊕ O atributo **Emprestar** será o designador da classificação (rótulo de classe) e conterà duas classes: Emprestar = **não** e Emprestar = **sim** (2 Classes).
- ⊕  $P(\text{Emprestar} = \text{não}) = 3/6 = 0,5$
- ⊕  $P(\text{Emprestar} = \text{sim}) = 3/6 = 0,5$
- ⊕  $E(S) = -0,5 \log_2 0,5 - 0,5 \log_2 0,5 = -0,5 (-1) - 0,5 (-1) = 0,5 + 0,5 = \mathbf{1}$

Rótulo de  
classe



| RID | Casado | Salário    | Saldo_conta | Idade     | Emprestar |
|-----|--------|------------|-------------|-----------|-----------|
| 1   | não    | $\geq 50K$ | $< 5K$      | $\geq 25$ | sim       |
| 2   | sim    | $\geq 50K$ | $\geq 5K$   | $\geq 25$ | sim       |
| 3   | sim    | 20K. . 50K | $< 5K$      | $< 25$    | não       |
| 4   | não    | $< 20K$    | $\geq 5K$   | $< 25$    | não       |
| 5   | não    | $< 20K$    | $< 5K$      | $\geq 25$ | não       |
| 6   | sim    | 20K. . 50K | $\geq 5K$   | $\geq 25$ | sim       |



## Cálculo da Entropia do Atributo Casado = Sim

$$\oplus \quad E(\text{Casado} = \text{Sim}) = -2/3 \log_2(2/3) - 1/3 \log_2(1/3) = -2/3 \log_2(0,666) - 1/3 \log_2(0,333)$$

$$\oplus \quad E(\text{Casado} = \text{Sim}) = \mathbf{0,92}$$

| RID | Casado | Salário     | Saldo_conta | Idade | Emprestar |
|-----|--------|-------------|-------------|-------|-----------|
| 1   | não    | >=50K       | <5K         | >=25  | sim       |
| 2   | sim    | >=50K       | >=5K        | >=25  | sim       |
| 3   | sim    | 20K. . .50K | <5K         | <25   | não       |
| 4   | não    | <20K        | >=5K        | <25   | não       |
| 5   | não    | <20K        | <5K         | >=25  | não       |
| 6   | sim    | 20K. . .50K | >=5K        | >=25  | sim       |



## Cálculo da Entropia do Atributo Casado = Não

$$\oplus E(\text{Casado} = \text{Sim}) = -2/3 \log_2(2/3) - 1/3 \log_2(1/3) = -2/3 \log_2(0,666) - 1/3 \log_2(0,333)$$

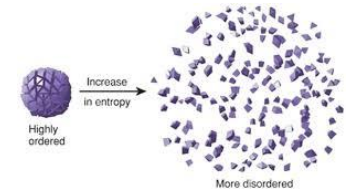
$$\oplus E(\text{Casado} = \text{Sim}) = \mathbf{0,92}$$

| RID | Casado | Salário     | Saldo_conta | Idade | Emprestar |
|-----|--------|-------------|-------------|-------|-----------|
| 1   | não    | >=50K       | <5K         | >=25  | sim       |
| 2   | sim    | >=50K       | >=5K        | >=25  | sim       |
| 3   | sim    | 20K. . .50K | <5K         | <25   | não       |
| 4   | não    | <20K        | >=5K        | <25   | não       |
| 5   | não    | <20K        | <5K         | >=25  | não       |
| 6   | sim    | 20K. . .50K | >=5K        | >=25  | sim       |





## Cálculo do Ganho de Informação



- ⊕ Ganho de informação é a redução esperada da entropia ao utilizarmos um determinado atributo da árvore;
- ⊕ Ganho de informação mede a redução na Entropia, ao se selecionar um determinado atributo;
- ⊕ O atributo com o maior ganho é escolhido como atributo de particionamento.

O ganho de informação é dado por:

$$\text{Ganho}(S, A) = \text{Entropia}(S) - \sum ((|S_v| / |S|) * \text{Entropia}(S_v))$$

Onde:

Ganho (S, A) é o ganho do atributo A sobre o conjunto S

S<sub>v</sub> = subconjunto de S para um valor do atributo A

|S<sub>v</sub>| = número de elementos de S<sub>v</sub>

|S| = número de elementos de S



## Exemplo de dados para concessão de Empréstimo Bancário

| <b>caso</b> | <b>montante</b> | <b>idade</b> | <b>salário</b> | <b>conta</b> | <b>empréstimo</b> |
|-------------|-----------------|--------------|----------------|--------------|-------------------|
| 1           | médio           | sênior       | baixo          | sim          | <b><i>não</i></b> |
| 2           | médio           | sênior       | baixo          | não          | <b><i>não</i></b> |
| 3           | baixo           | sênior       | baixo          | sim          | <b><i>sim</i></b> |
| 4           | alto            | média        | baixo          | sim          | <b><i>sim</i></b> |
| 5           | alto            | jovem        | alto           | sim          | <b><i>sim</i></b> |
| 6           | alto            | jovem        | alto           | não          | <b><i>não</i></b> |
| 7           | baixo           | jovem        | alto           | não          | <b><i>sim</i></b> |
| 8           | médio           | média        | baixo          | sim          | <b><i>não</i></b> |
| 9           | médio           | jovem        | alto           | sim          | <b><i>sim</i></b> |
| 10          | alto            | média        | alto           | sim          | <b><i>sim</i></b> |
| 11          | médio           | média        | alto           | não          | <b><i>sim</i></b> |
| 12          | baixo           | jovem        | baixo          | não          | <b><i>sim</i></b> |
| 13          | baixo           | sênior       | alto           | sim          | <b><i>sim</i></b> |
| 14          | alto            | média        | baixo          | não          | <b><i>não</i></b> |



## Selecionando o melhor atributo:

$$\text{Entropia}(S) = - 9/14 \log_2 (9/14) - 5/14 \log_2 (5/14) = 0,940$$

| caso | montante | idade  | salário | conta | empréstimo        |
|------|----------|--------|---------|-------|-------------------|
| 1    | médio    | sênior | baixo   | sim   | <i><b>não</b></i> |
| 2    | médio    | sênior | baixo   | não   | <i><b>não</b></i> |
| 3    | baixo    | sênior | baixo   | sim   | <i><b>sim</b></i> |
| 4    | alto     | média  | baixo   | sim   | <i><b>sim</b></i> |
| 5    | alto     | jovem  | alto    | sim   | <i><b>sim</b></i> |
| 6    | alto     | jovem  | alto    | não   | <i><b>não</b></i> |
| 7    | baixo    | jovem  | alto    | não   | <i><b>sim</b></i> |
| 8    | médio    | média  | baixo   | sim   | <i><b>não</b></i> |
| 9    | médio    | jovem  | alto    | sim   | <i><b>sim</b></i> |
| 10   | alto     | média  | alto    | sim   | <i><b>sim</b></i> |
| 11   | médio    | média  | alto    | não   | <i><b>sim</b></i> |
| 12   | baixo    | jovem  | baixo   | não   | <i><b>sim</b></i> |
| 13   | baixo    | sênior | alto    | sim   | <i><b>sim</b></i> |
| 14   | alto     | média  | baixo   | não   | <i><b>não</b></i> |

Amarelo = classe ***não***

Verde = classe ***sim***



## Selecionando o melhor atributo

$$\text{Entropia}(\text{montante}=\text{médio}) = - 2/5 \log_2 (2/5) - 3/5 \log_2 (3/5) = 0,971$$

| caso | montante | idade  | salário | conta | empréstimo |
|------|----------|--------|---------|-------|------------|
| 1    | médio    | sênior | baixo   | sim   | <i>não</i> |
| 2    | médio    | sênior | baixo   | não   | <i>não</i> |
| 3    | baixo    | sênior | baixo   | sim   | <i>sim</i> |
| 4    | alto     | média  | baixo   | sim   | <i>sim</i> |
| 5    | alto     | jovem  | alto    | sim   | <i>sim</i> |
| 6    | alto     | jovem  | alto    | não   | <i>não</i> |
| 7    | baixo    | jovem  | alto    | não   | <i>sim</i> |
| 8    | médio    | média  | baixo   | sim   | <i>não</i> |
| 9    | médio    | jovem  | alto    | sim   | <i>sim</i> |
| 10   | alto     | média  | alto    | sim   | <i>sim</i> |
| 11   | médio    | média  | alto    | não   | <i>sim</i> |
| 12   | baixo    | jovem  | baixo   | não   | <i>sim</i> |
| 13   | baixo    | sênior | alto    | sim   | <i>sim</i> |
| 14   | alto     | média  | baixo   | não   | <i>não</i> |



## Selecionando o melhor atributo

Entropia(montante=médio) =  $- 2/5 \log_2 (2/5) - 3/5 \log_2 (3/5) = 0,971$

Entropia(montante=baixo) =  $- 4/4 \log_2 (4/4) - 0/4 \log_2 (0/4) = 0$

| caso | montante | idade  | salário | conta | empréstimo        |
|------|----------|--------|---------|-------|-------------------|
| 1    | médio    | sênior | baixo   | sim   | <i><b>não</b></i> |
| 2    | médio    | sênior | baixo   | não   | <i><b>não</b></i> |
| 3    | baixo    | sênior | baixo   | sim   | <i><b>sim</b></i> |
| 4    | alto     | média  | baixo   | sim   | <i><b>sim</b></i> |
| 5    | alto     | jovem  | alto    | sim   | <i><b>sim</b></i> |
| 6    | alto     | jovem  | alto    | não   | <i><b>não</b></i> |
| 7    | baixo    | jovem  | alto    | não   | <i><b>sim</b></i> |
| 8    | médio    | média  | baixo   | sim   | <i><b>não</b></i> |
| 9    | médio    | jovem  | alto    | sim   | <i><b>sim</b></i> |
| 10   | alto     | média  | alto    | sim   | <i><b>sim</b></i> |
| 11   | médio    | média  | alto    | não   | <i><b>sim</b></i> |
| 12   | baixo    | jovem  | baixo   | não   | <i><b>sim</b></i> |
| 13   | baixo    | sênior | alto    | sim   | <i><b>sim</b></i> |
| 14   | alto     | média  | baixo   | não   | <i><b>não</b></i> |



## Selecionando o melhor atributo

Entropia(montante=médio) =  $- 2/5 \log_2 (2/5) - 3/5 \log_2 (3/5) = 0,971$

Entropia(montante=baixo) =  $- 4/4 \log_2 (4/4) - 0/4 \log_2 (0/4) = 0$

Entropia(montante=alto) =  $- 3/5 \log_2 (3/5) - 2/5 \log_2 (2/5) = 0,971$

| caso | montante | idade  | salário | conta | empréstimo |
|------|----------|--------|---------|-------|------------|
| 1    | médio    | sênior | baixo   | sim   | <i>não</i> |
| 2    | médio    | sênior | baixo   | não   | <i>não</i> |
| 3    | baixo    | sênior | baixo   | sim   | <i>sim</i> |
| 4    | alto     | média  | baixo   | sim   | <i>sim</i> |
| 5    | alto     | jovem  | alto    | sim   | <i>sim</i> |
| 6    | alto     | jovem  | alto    | não   | <i>não</i> |
| 7    | baixo    | jovem  | alto    | não   | <i>sim</i> |
| 8    | médio    | média  | baixo   | sim   | <i>não</i> |
| 9    | médio    | jovem  | alto    | sim   | <i>sim</i> |
| 10   | alto     | média  | alto    | sim   | <i>sim</i> |
| 11   | médio    | média  | alto    | não   | <i>sim</i> |
| 12   | baixo    | jovem  | baixo   | não   | <i>sim</i> |
| 13   | baixo    | sênior | alto    | sim   | <i>sim</i> |
| 14   | alto     | média  | baixo   | não   | <i>não</i> |



## Selecionando o melhor atributo

$$\text{Entropia}(S) = - 9/14 \log_2 (9/14) - 5/14 \log_2 (5/14) = 0,940$$

$$\text{Entropia}(\text{montante}=\text{médio}) = - 2/5 \log_2 (2/5) - 3/5 \log_2 (3/5) = 0,971$$

$$\text{Entropia}(\text{montante}=\text{baixo}) = - 4/4 \log_2 (4/4) - 0/4 \log_2 (0/4) = 0$$

$$\text{Entropia}(\text{montante}=\text{alto}) = - 3/5 \log_2 (3/5) - 2/5 \log_2 (2/5) = 0,971$$

$$\text{Entropia}(\text{idade} = \text{senior}) = - 2/4 \log_2 (2/4) - 2/4 \log_2 (2/4) = 1$$

$$\text{Entropia}(\text{idade} = \text{média}) = - 3/5 \log_2 (3/5) - 2/5 \log_2 (2/5) = 0,971$$

$$\text{Entropia}(\text{idade} = \text{jovem}) = - 4/5 \log_2 (4/5) - 1/5 \log_2 (1/5) = 0,722$$

.....

$$\text{Ganho}(S, \text{montante}) = 0,940 - (5/14) \cdot 0,971 - (4/14) \cdot 0 - (5/14) \cdot 0,971 = 0,246$$

$$\text{Ganho}(S, \text{idade}) = 0,940 - (4/14) \cdot 1 - (5/14) \cdot 0,971 - (5/14) \cdot 0,722 = 0,049$$

$$\text{Ganho}(S, \text{salário}) = 0,940 - (7/14) \cdot 0,592 - (7/14) \cdot 0,985 = 0,151$$

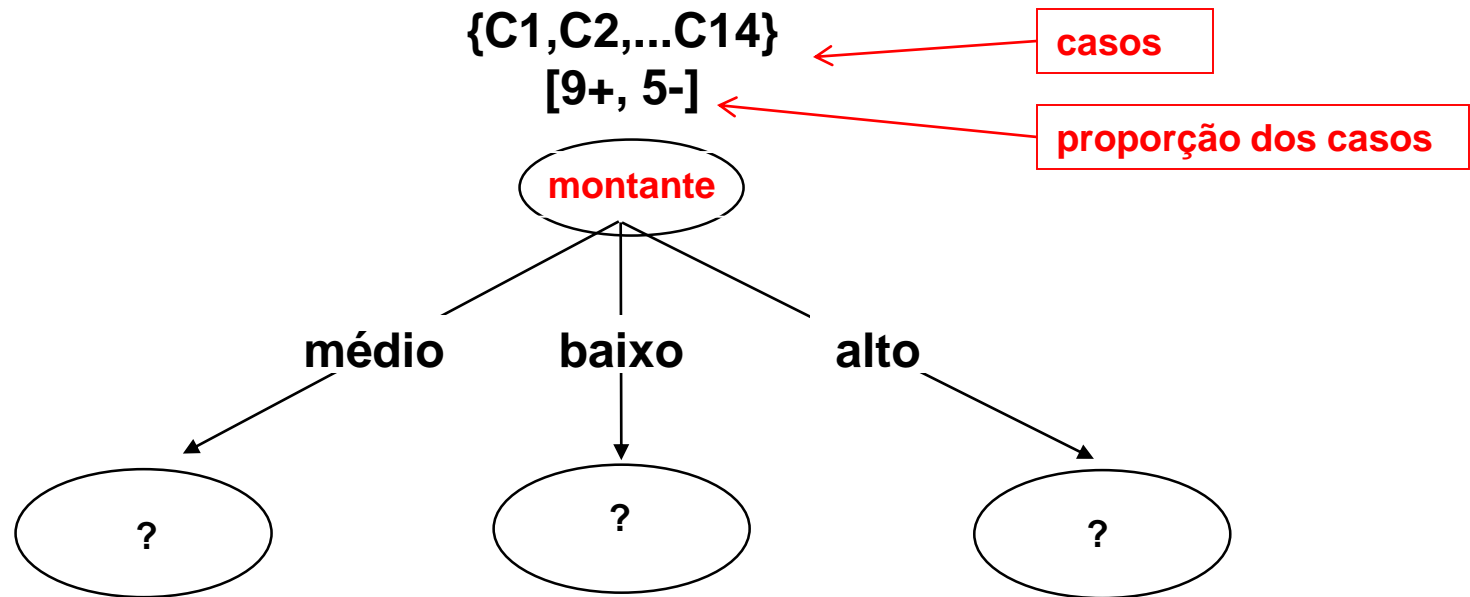
$$\text{Ganho}(S, \text{conta}) = 0,940 - (8/14) \cdot 0,811 - (6/14) \cdot 1 = 0,047$$

Maior  
Ganho





## Escolha do Próximo Atributo







## Ferramentas para Mineração de Dados

- ⊕ Kate
- ⊕ Knowledge SEEKER
- ⊕ Business Miner
- ⊕ QueryObject
- ⊕ Data Surveyour
- ⊕ DBMiner
- ⊕ Intellingent Miner
- ⊕ Enterprise Miner