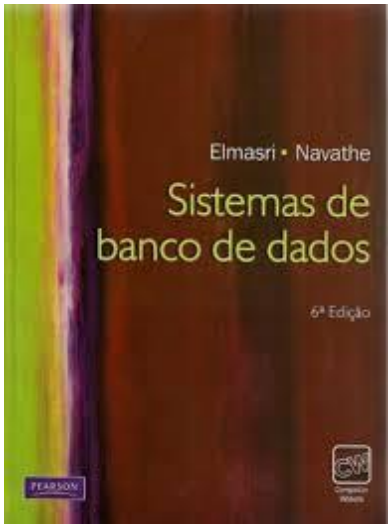
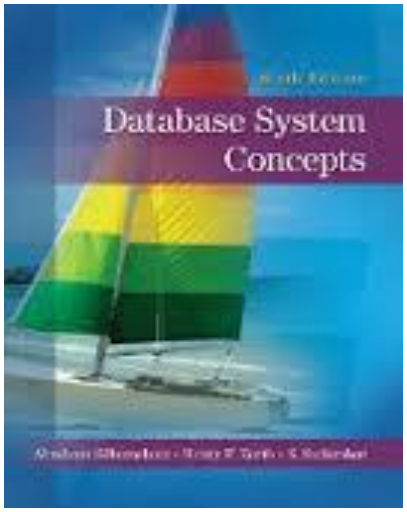
Unidade 23 – Conceitos de Mineração de Dados – Parte 3

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 Editon

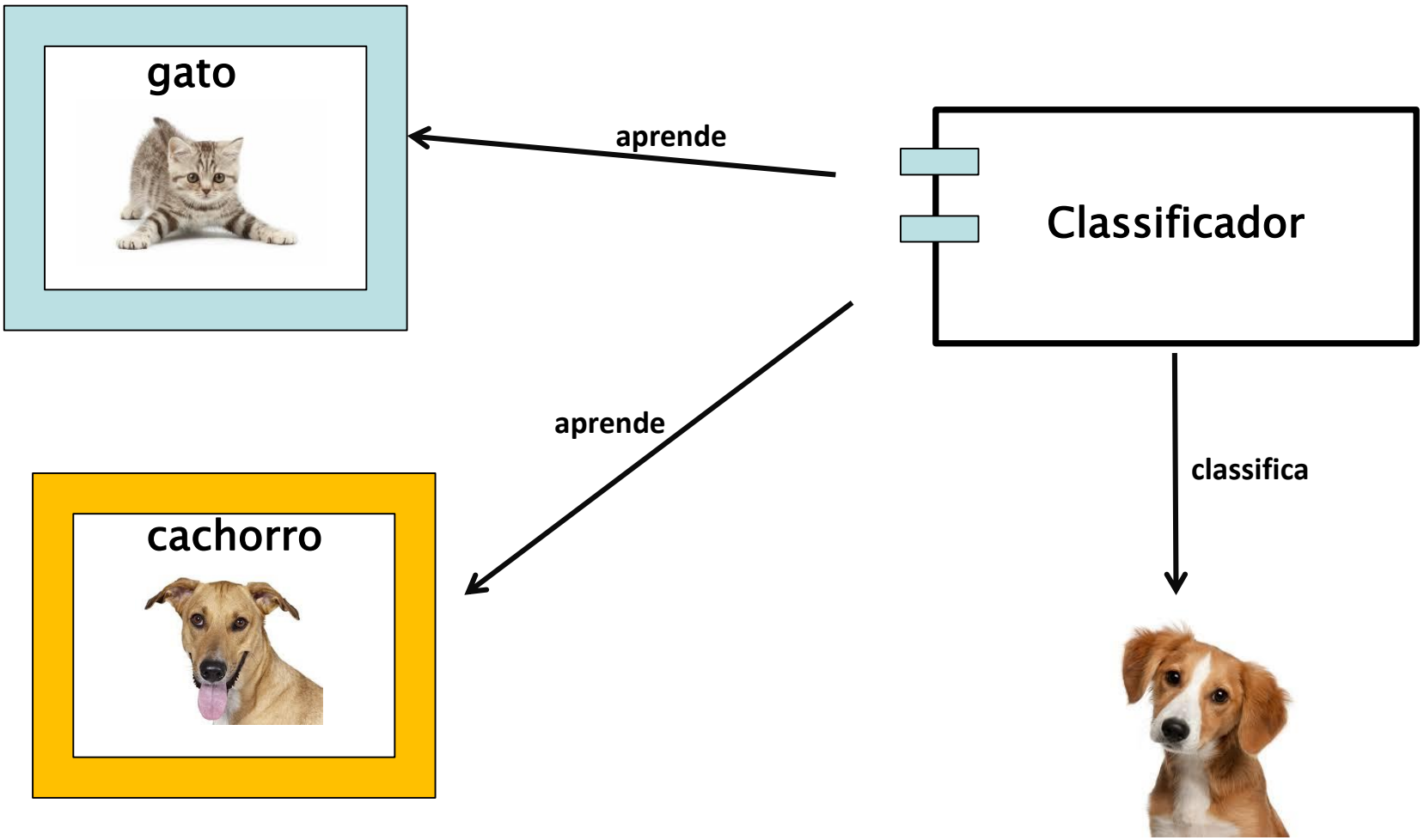
Tipos de Conhecimentos descobertos pela Mineração de Dados

- 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.**

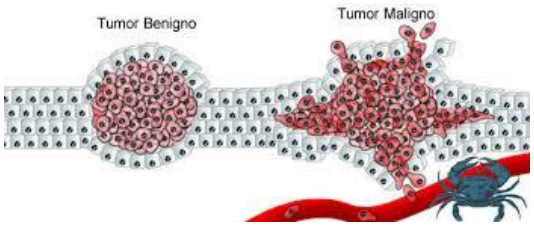
Classificação

- 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**.

Aprendizado Supervisionado



Exemplos - Classificação



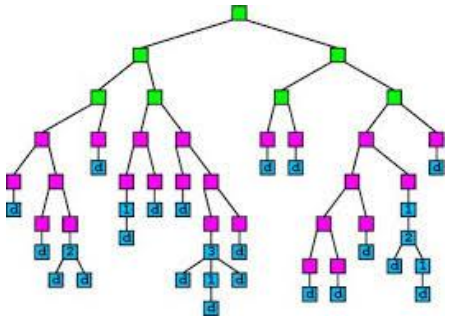
- Predizer se um tumor é benigno ou maligno;
- Classificar transações de cartões de crédito como legítimas ou fraudulentas;
- Analisar concessão de empréstimos bancários em Instituições financeiras;
- Filtrar (marcar) e-mails que seriam spams em softwares de correio eletrônico.

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

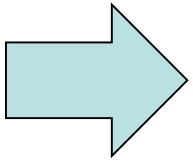


Exemplo – Árvore de Decisão

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



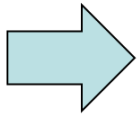
Learning Algorithm

Algoritmo de Aprendizado

classe

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

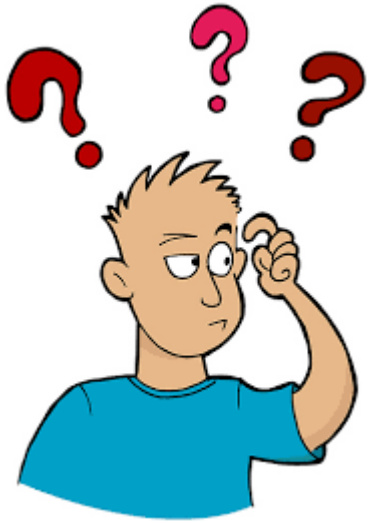
Set Training



Learning Algorithm

- 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?



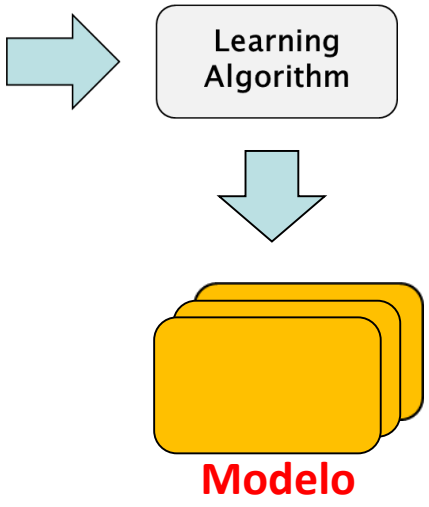
Algoritmo de Aprendizado

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



Para que serve o modelo gerado pelo algoritmo de aprendizado?

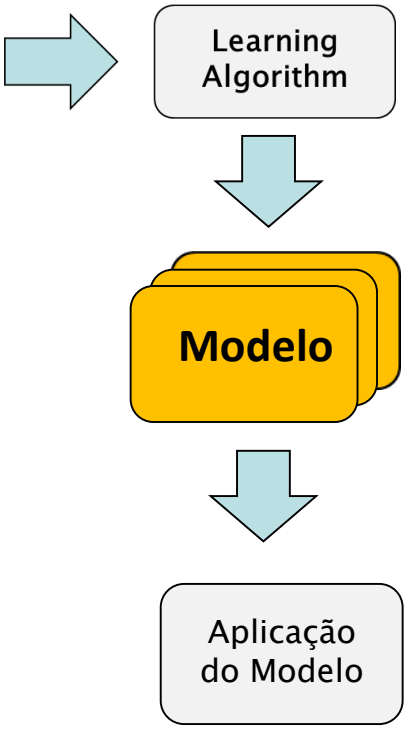
Uso do modelo

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

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



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	?

Test Set

Modelo – Árvore de Decisão

- As árvores de decisão são representações do modelo que consistem em:
  - **Nós internos**, que representam os atributos;
  - **Arestas** que correspondem aos valores dos atributos;
  - **Nós folha**, que designam uma classificação.

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;

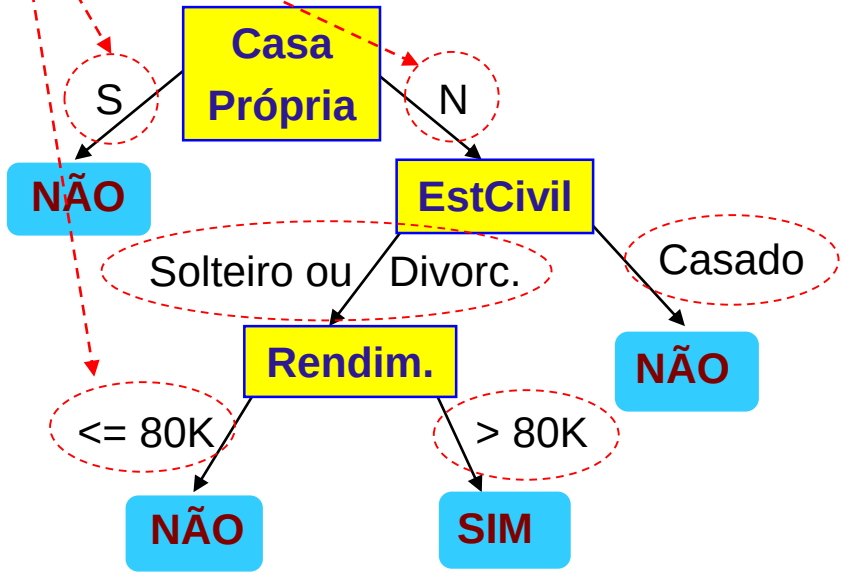
Modelo: árvore de decisão

Modelo – Árvore de Decisão

	categórico	categórico	continuo	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;



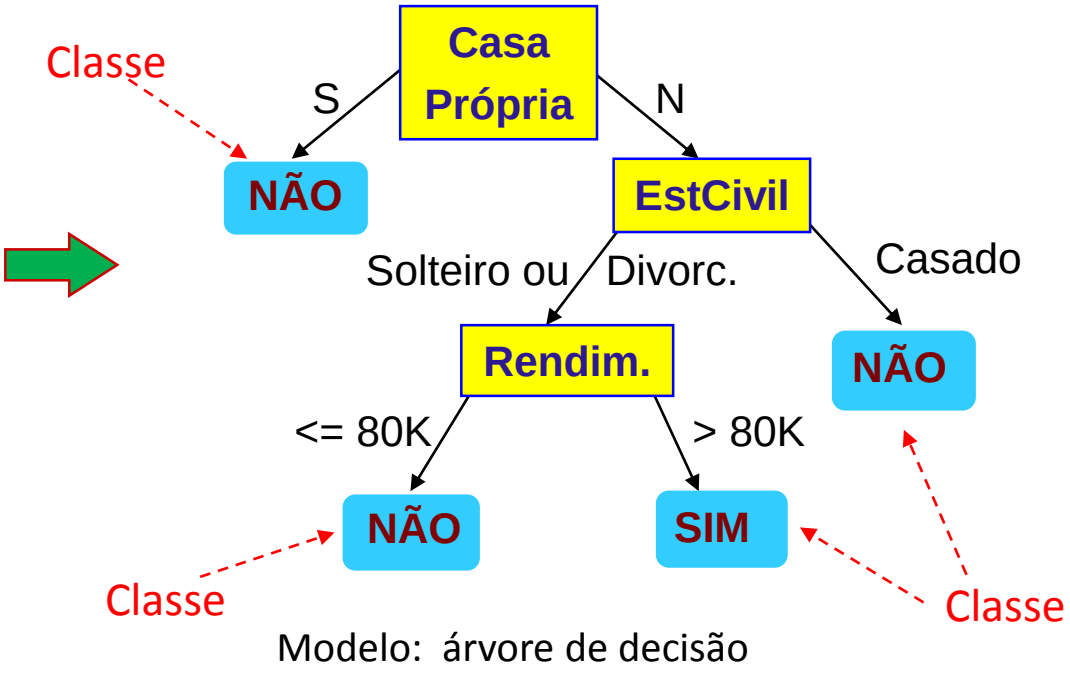
Modelo: árvore de decisão

Modelo – Árvore de Decisão

	categórico	categórico	continuo	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 folha** designam a classificação;

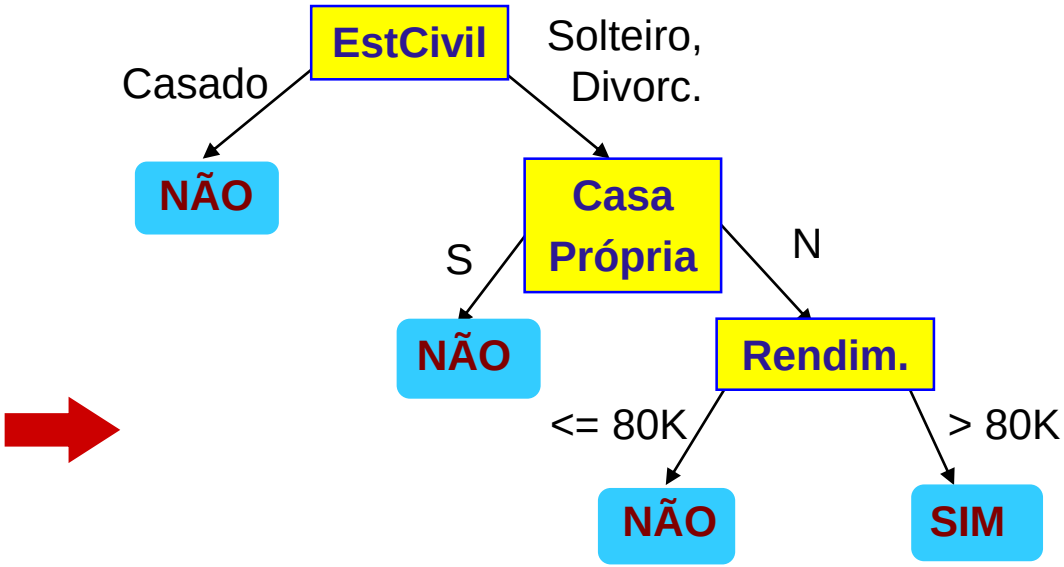


Modelo: árvore de decisão

Pode haver mais de uma árvore de decisão?

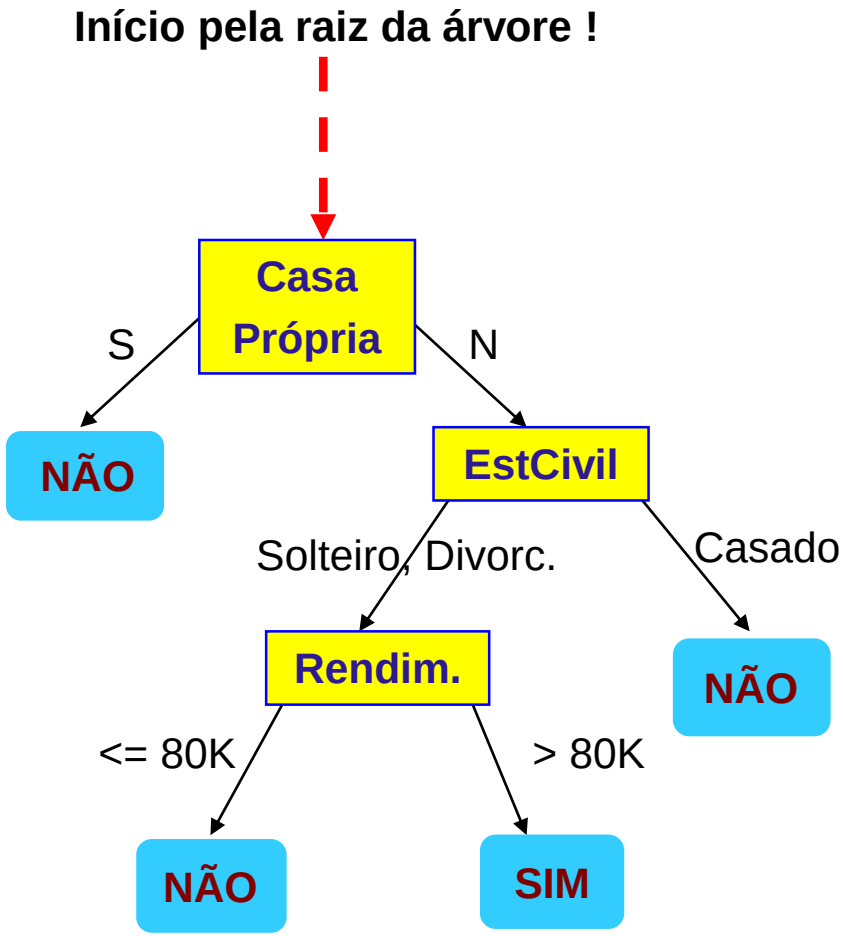
Outra Árvore de Decisão

Id	Casa própria	EstCivil	Rendim.	Mau Pagador (classe)
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



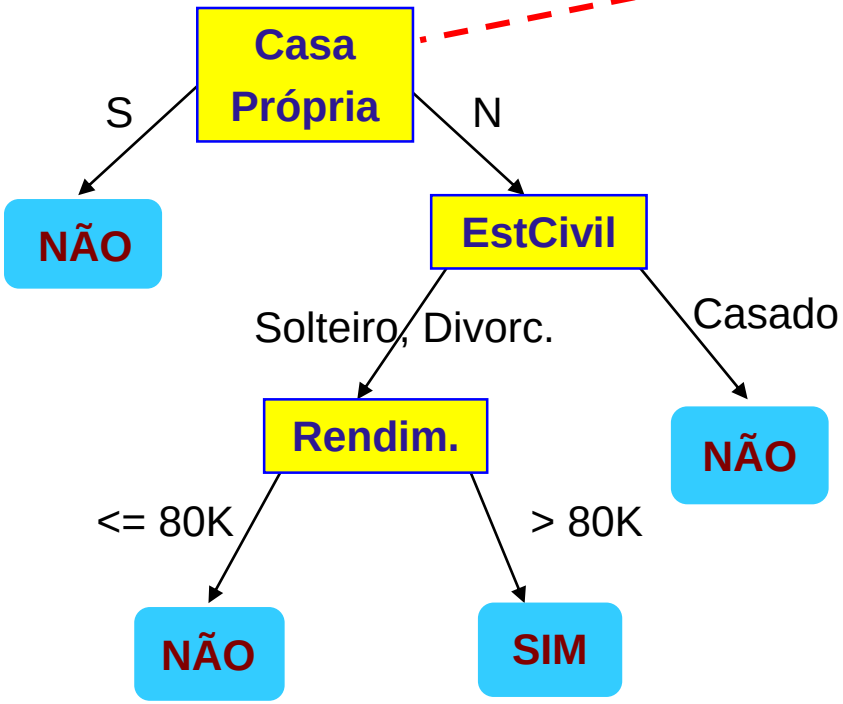
Dado para teste

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

Aplicando o modelo nos dados de teste

Dado para teste

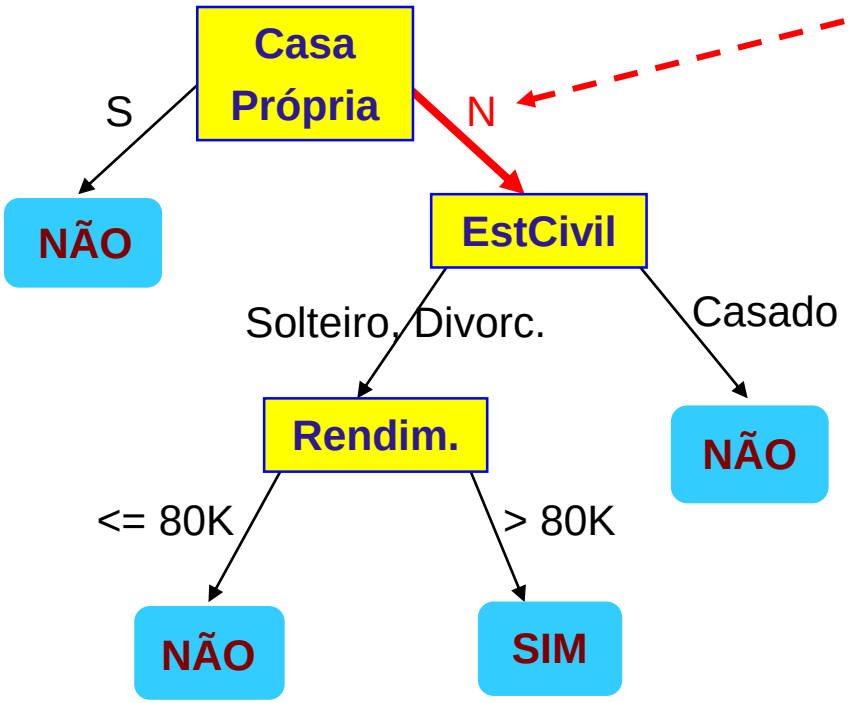
Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?



Aplicando o modelo nos dados de teste

Dado para teste

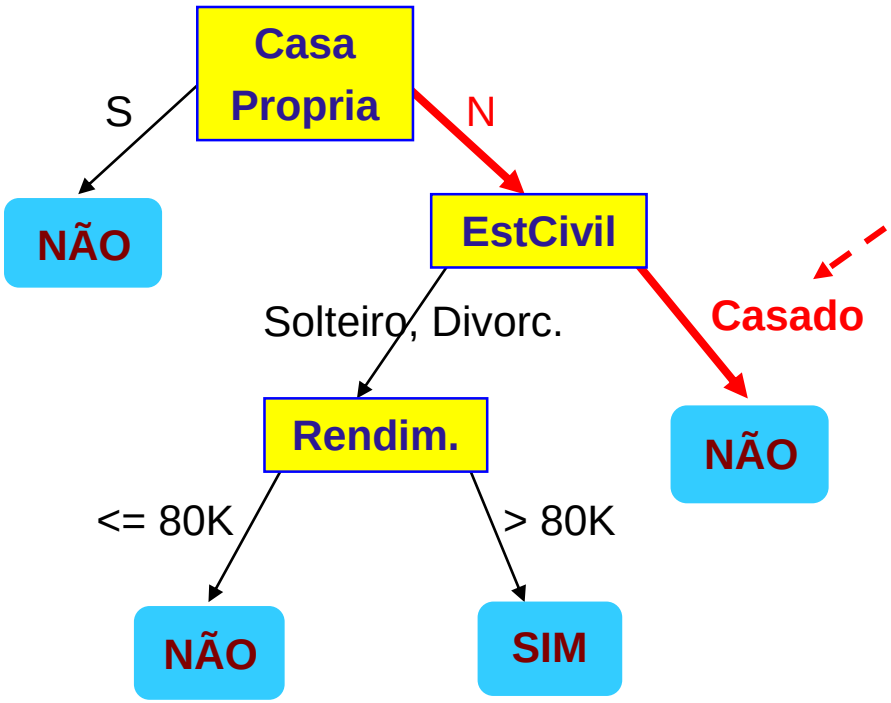
Casa Própria	Estado Civil	Rendim.	Mau pagador
N	Casado	80K	?



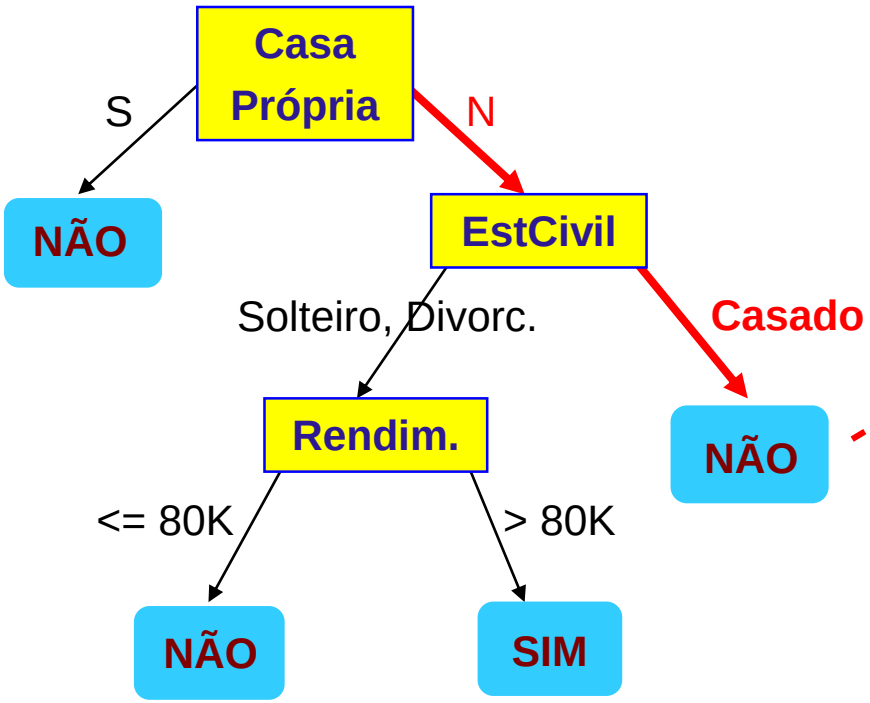
Aplicando o modelo nos dados de teste

Dado para teste

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



Aplicando o modelo nos dados de teste



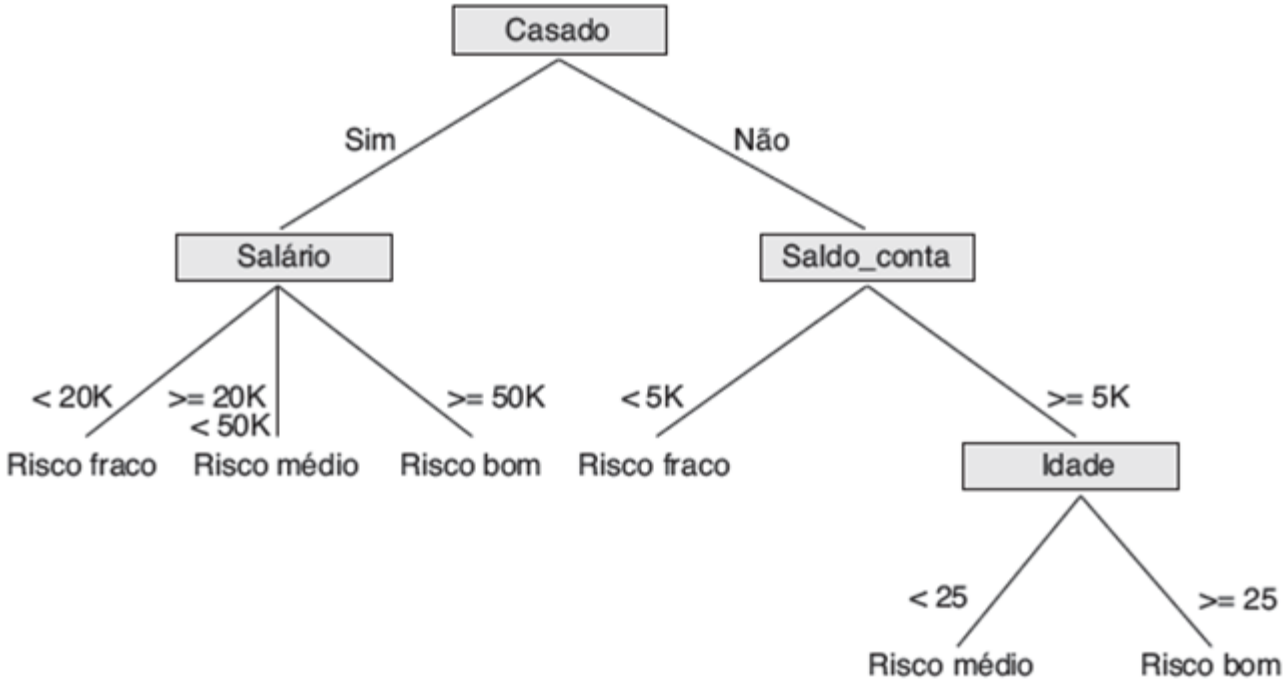
Dado para teste

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

NÃO

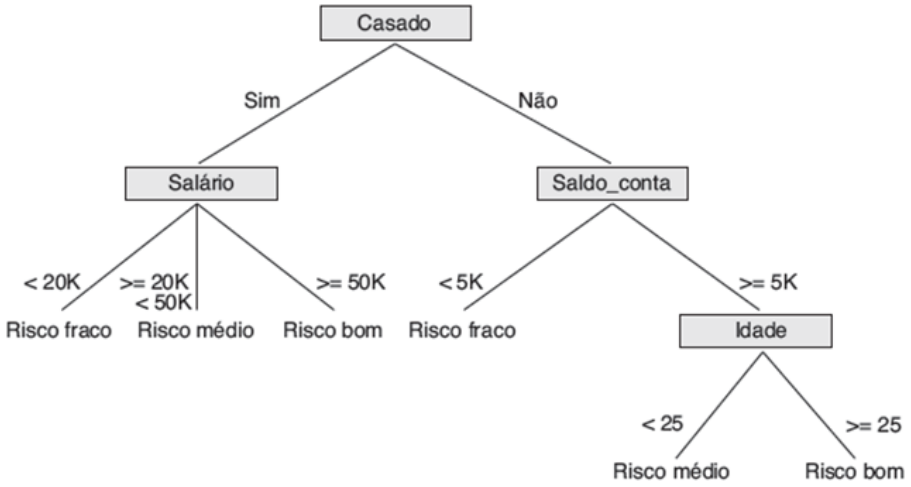
Será atribuído à classe Mau pagador o valor **NÃO**.

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 >= 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 deve ser colocado na raiz da árvore**?";

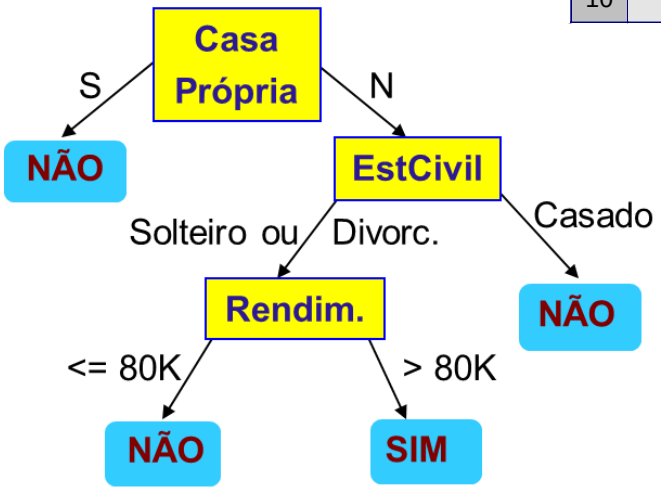
- Para isso cada atributo é testado e sua capacidade para se tornar nó raiz é avaliada;
- Cria-se tantos nós filhos da raiz quantos valores possíveis esse atributo puder assumir;
- Repete-se o processo para cada nó filho da raiz e assim, sucessivamente.

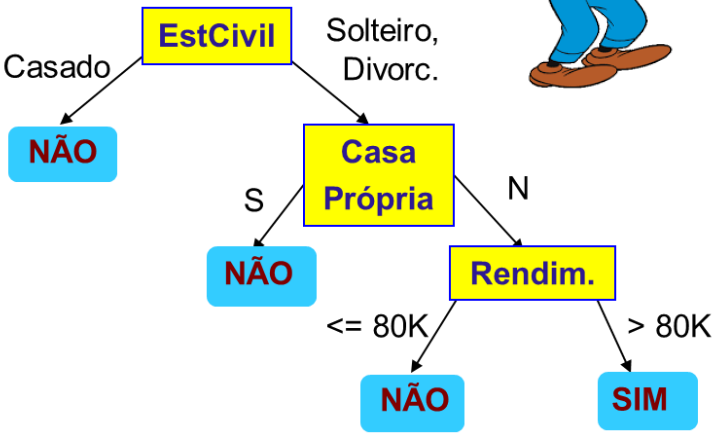
Algoritmo ID3 - Observação

- **"Qual atributo 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

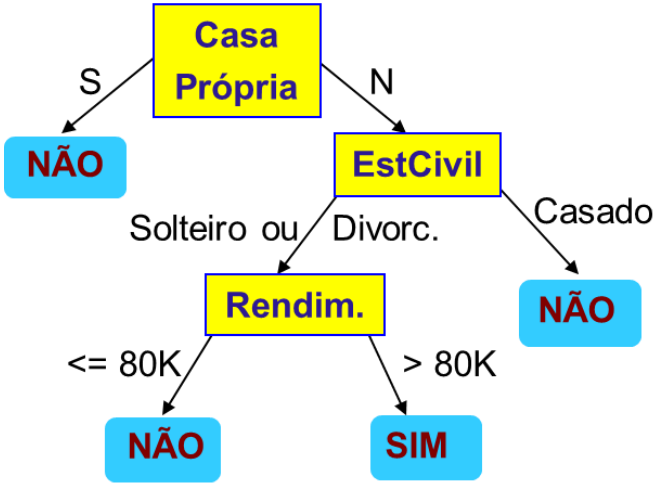
A raiz será CasaPróprio ou EstCivil?

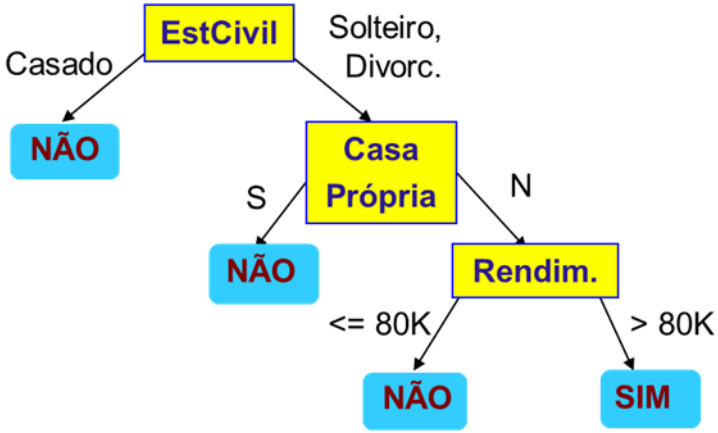




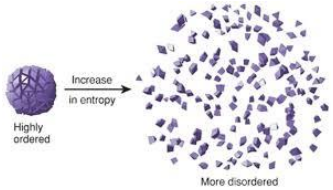
Como definir o atributo que será usado como raiz da Árvore de Decisão?

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:

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

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

$$E(S) = -\frac{7}{14}\log_2\frac{7}{14} - \frac{7}{14}\log_2\frac{7}{14} = 0.99\ldots \approx 1$$

Incerteza !

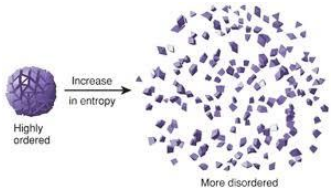
  - Para [0+, 14-] ou [14+, 0-]

$$E(S) = -\frac{14}{14}\log_2\frac{14}{14} = 0$$

Certeza !

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

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;
- Detalhamento do Algoritmo descrito no livro do **Navathe**.

Ferramentas para Mineração de Dados

- Kate
- Knowledege SEEKER
- Business Miner
- QueryObject
- Data Surveyour
- DBMiner
- Intellingent Miner
- Enterprise Miner