



# Aprendizado Supervisionado Parte 2

Árvores de Decisão

Silvia Moraes

Material elaborado pelo prof. **Duncan Ruiz**

# Roteiro

Relembrando

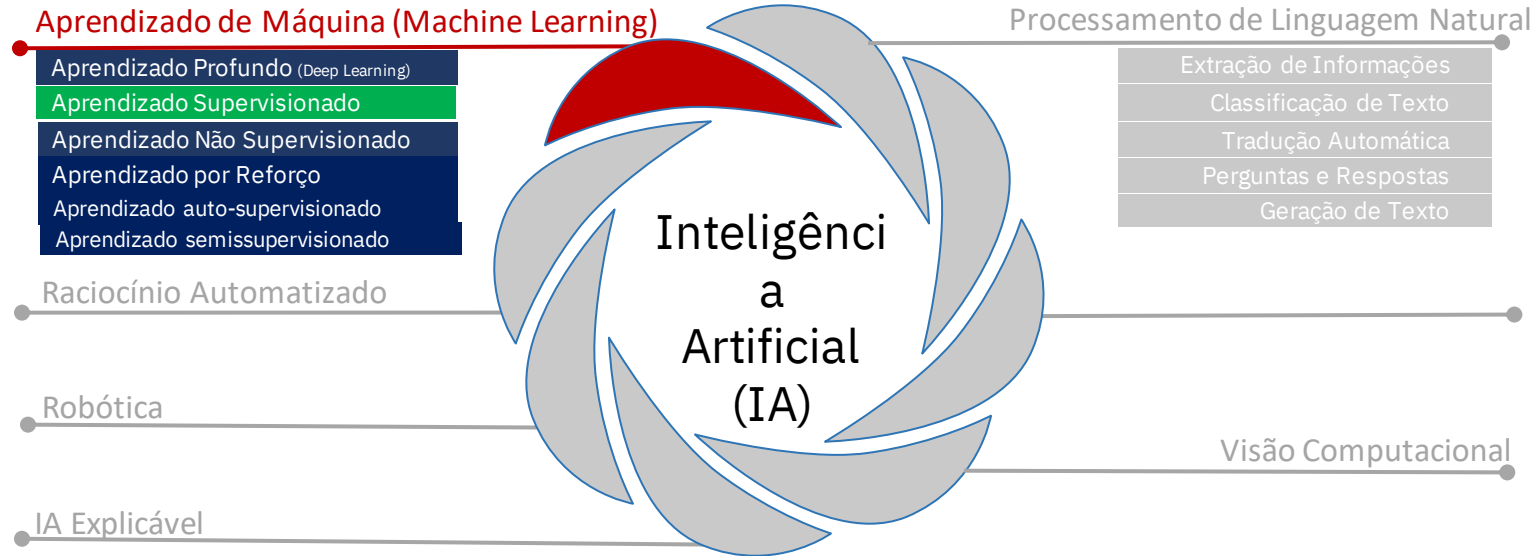
Machine Learning: Aprendizado Supervisionado

Tarefas Preditivas

Classificação & Regressão

Árvores de Decisão

# Subáreas da Inteligência Artificial



# Aprendizado Supervisionado

Exige que os **dados** estejam **rotulados** (anotados com suas respectivas classes/valores de saída)

Os algoritmos que seguem esse tipo de aprendizado recebem pares de valores:

- os dados de entrada (x) e
- os valores de saída (rótulos) correspondentes (y).



# Aprendizado Supervisionado

Em um conjunto de dados (exemplos) rotulado :

- Cada dado corresponde a um indivíduo do domínio e é formado por uma tupla contendo características (features).

**Atributo de entrada**  
(atributo previsor)

sepal length	sepal width	petal length	petal width	class
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
7,0	3,2	4,7	7,1	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
6,3	3,3	6,0	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica

**Atributo de saída**  
(atributo alvo ou meta)

**Rótulo**  
(Classes)

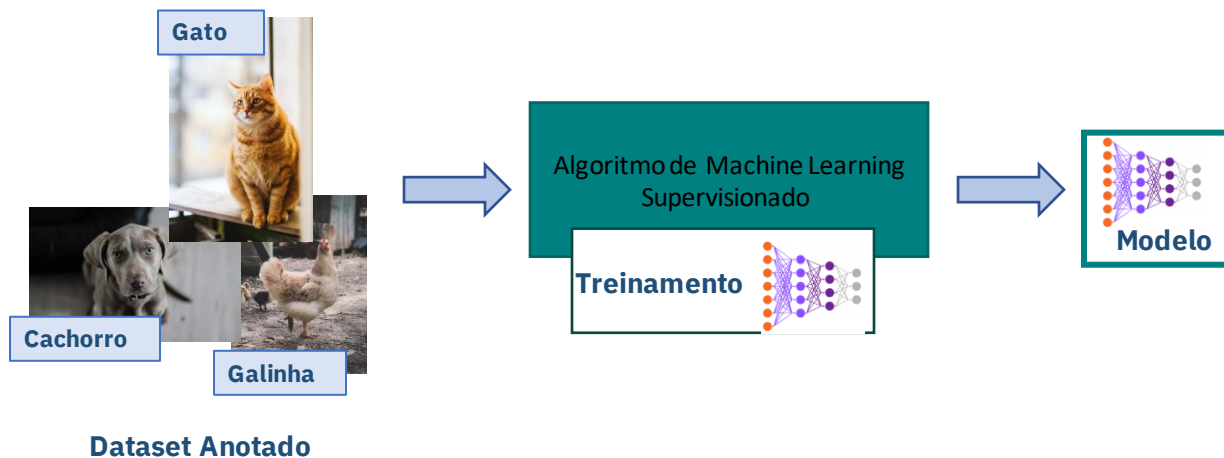


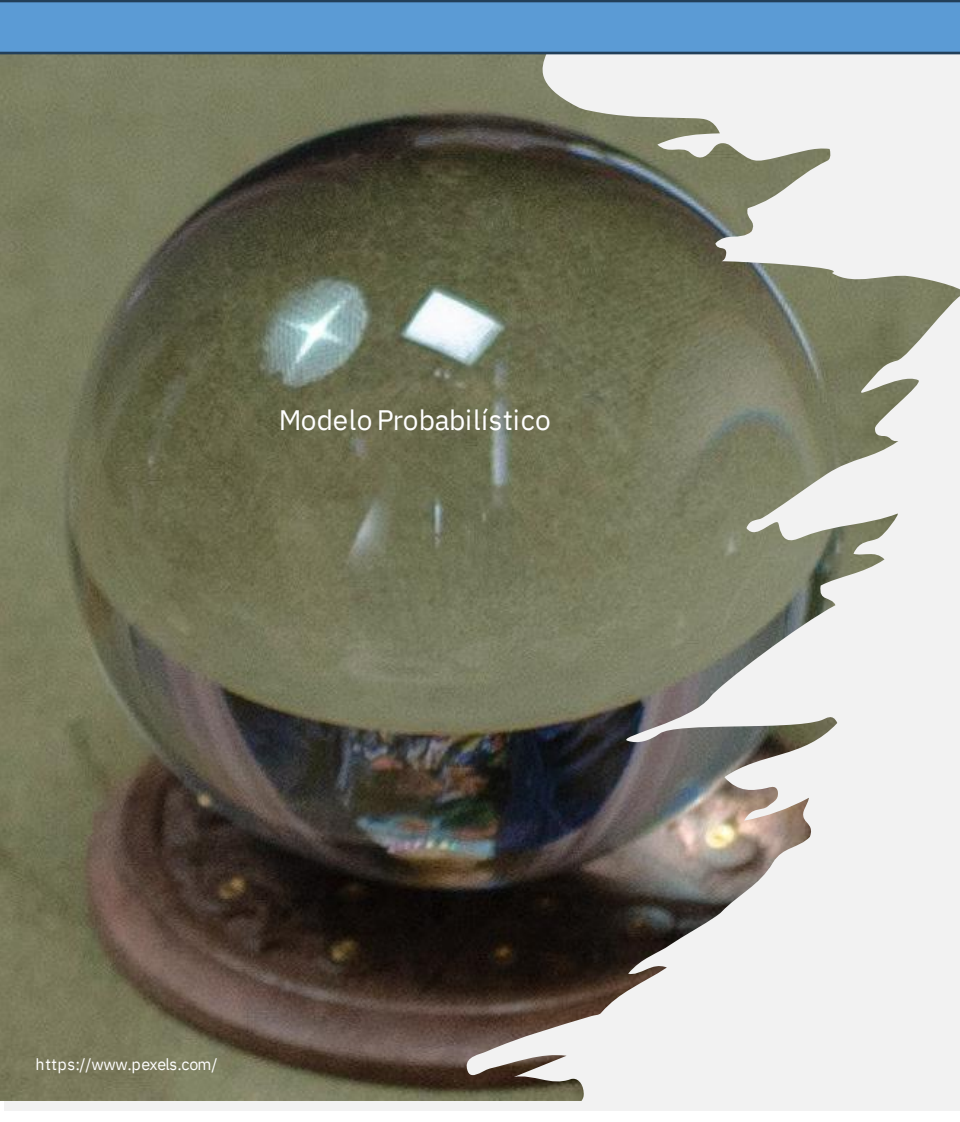
# Aprendizado Supervisionado

O objetivo é encontrar um modelo capaz de mapear os valores de entrada ( $x$ ) nos valores de saída  $y$ .

Em outras palavras, que aproxime  $f$ , tal que  $f(x) = y$ .

Supervisão: ajuste usando o erro em relação à saída esperada.





Modelo Probabilístico

# Aprendizado Supervisionado

**Tarefa preditiva:** encontra uma função (modelo) a partir dos dados de treino que possa ser usada para prever um rótulo (classe) ou valor de um novo exemplo.

Pode ser:

**classificação** (rótulos discretos)

**regressão** (rótulos contínuos)

# Classificação



É o processo de automaticamente atribuir rótulos a dados.

Pode ser do tipo

- Binária: possui apenas duas classes

- Multiclasse: possui mais de duas classes

Pode atribuir

- Um único rótulo (single label)

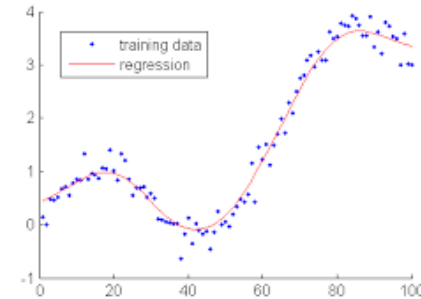
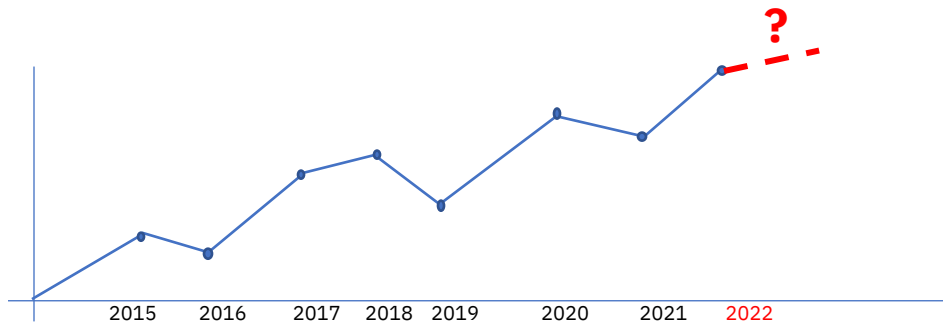
- Vários rótulos (multi-label)



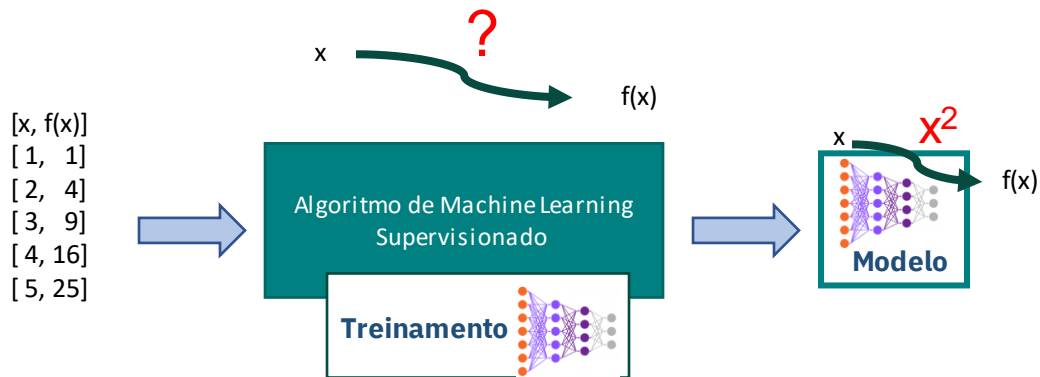


# Regressão

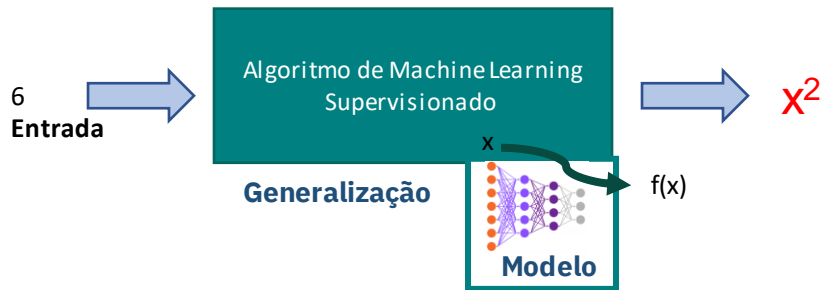
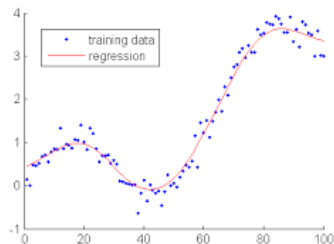
É o processo de automaticamente predizer novos valores  $y$ .  
Neste caso, os dados são anotados com valores.



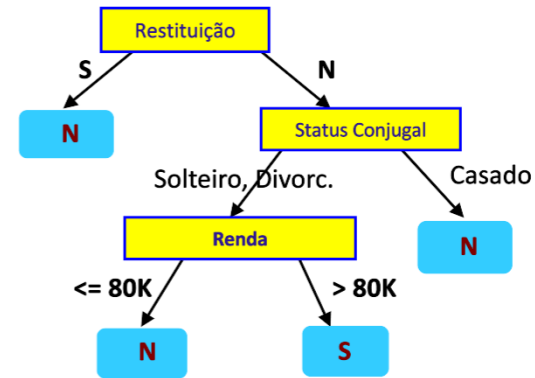
# Regressão



Dataset Anotado



# Classificação & Regressão com Árvores de Decisão

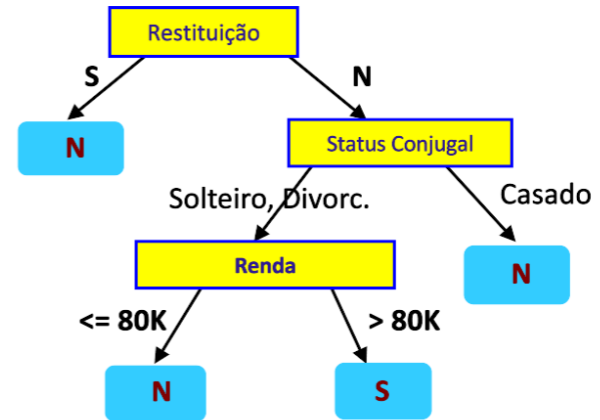


# Classificação



# Árvores de Decisão

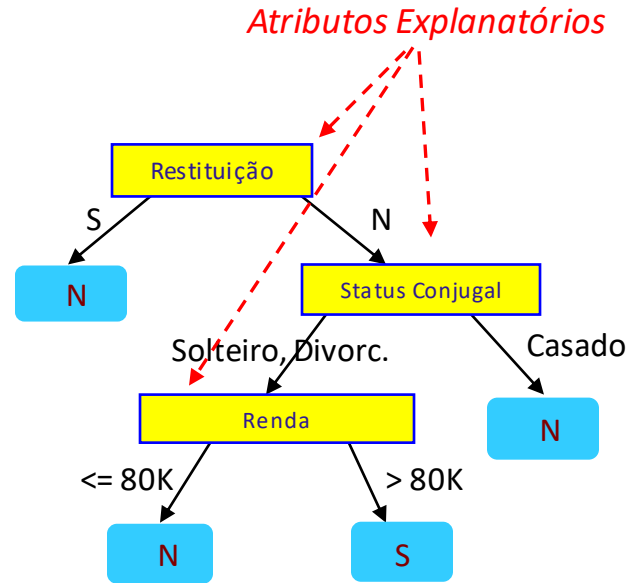
- Método para aproximar funções discretas ou contínuas, representadas por meio de um grafo acíclico direcionado, com vértice inicial único
- Tal grafo pode ser representado por um conjunto de regras “SE...ENTÃO” (Compreensibilidade)
- Amplamente utilizado em aplicações práticas, principalmente em problemas de classificação



# Exemplo de Árvore de Decisão

<i>Tid</i>	<i>Restituição</i>	<i>Status Conjugal</i>	<i>Renda</i>	<i>Calote ?</i>
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Conjunto de Treino



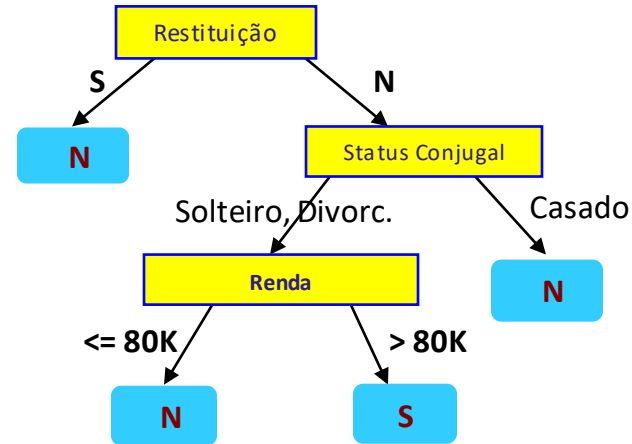
Modelo: Árvore de Decisão

# Exemplo de Árvore de Decisão

*categorico*  
*Categorico*  
*contínuo*  
*classe*

<i>Tid</i>	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Conjunto de Treino



Modelo: Árvore de Decisão

# Outro exemplo de Árvore de Decisão

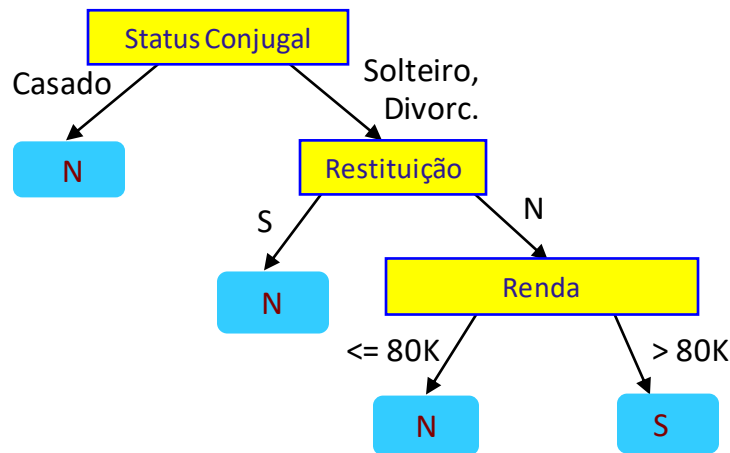
<i>Tid</i>	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

categorico

Categorico

contínuo

classe



Pode existir mais de uma árvore de decisão adequada para os mesmos dados!

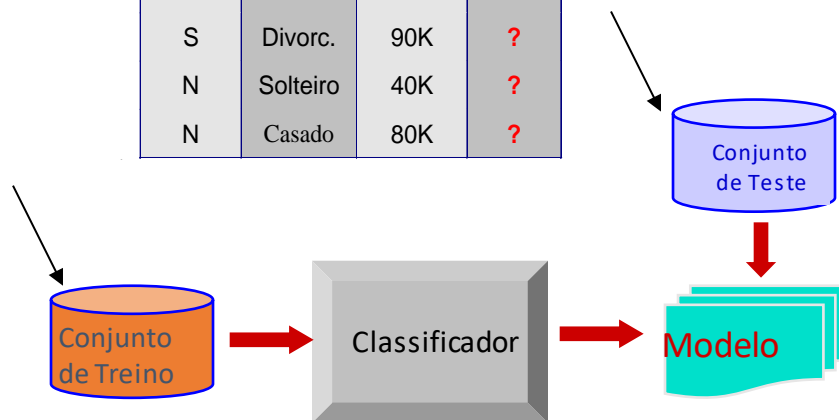


# Exemplo de Classificação

*categorico*  
*Categorico*  
*contínuo*  
*classe*

<i>Tid</i>	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Restituição	Status Conjugal	Renda	Calote ?
N	Solteiro	75K	?
S	Casado	50K	?
N	Casado	150K	?
S	Divorc.	90K	?
N	Solteiro	40K	?
N	Casado	80K	?



# Exemplo de Classificação

Tid	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Treino

Indução

Algoritmo de  
Indução de  
Árvore de  
Decisão

Aprendizado

Modelo

Aplicação

Dedução

Teste

Restituição	Status Conjugal	Renda	Calote ?
N	Solteiro	75K	?
S	Casado	50K	?
N	Casado	150K	?
S	Divorc.	90K	?
N	Solteiro	40K	?
N	Casado	80K	?

# Exemplo de Classificação

Tid	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Treino

Indução

Algoritmo de  
Indução de  
Árvore de  
Decisão

Aprendizado

Modelo

Aplicação

Dedução

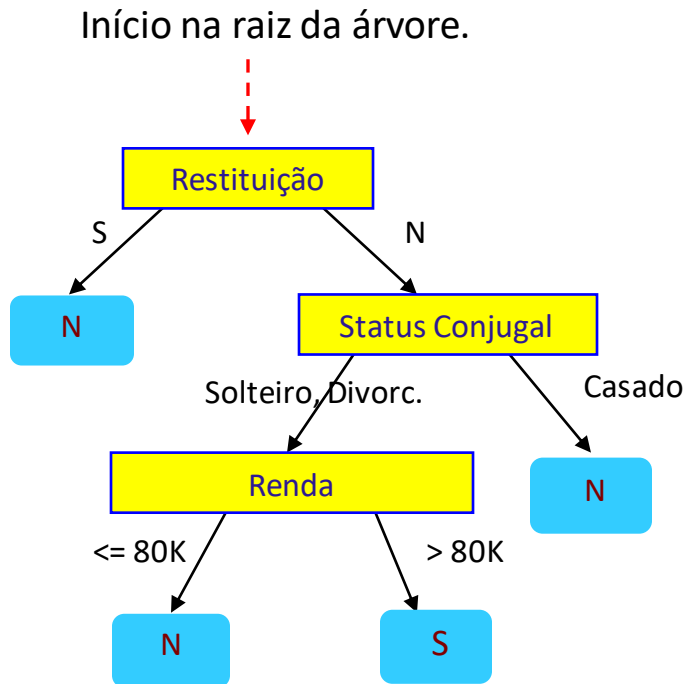
Teste

Restituição	Status Conjugal	Renda	Calote ?
N	Solteiro	75K	?
S	Casado	50K	?
N	Casado	150K	?
S	Divorc.	90K	?
N	Solteiro	40K	?
N	Casado	80K	?

# Aplicando o Modelo aos Dados de Teste

Dados de Teste

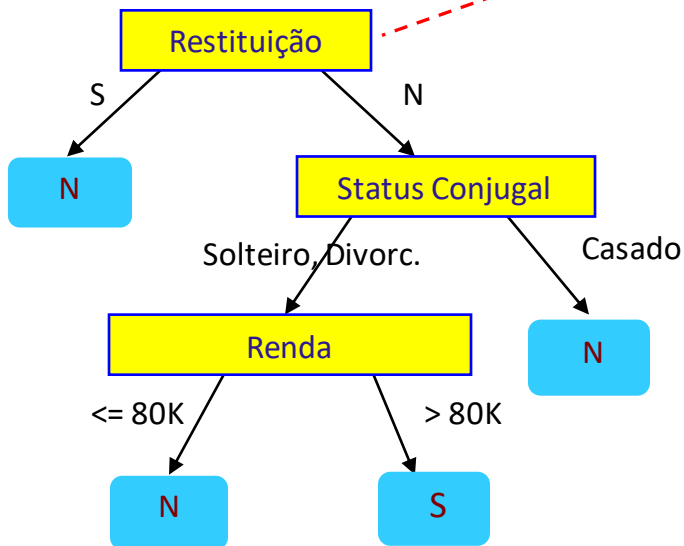
Restituição	Status Conjugal	Renda	Calote ?
N	Casado	80K	?



# Aplicando o Modelo aos Dados de Teste

Dados de Teste

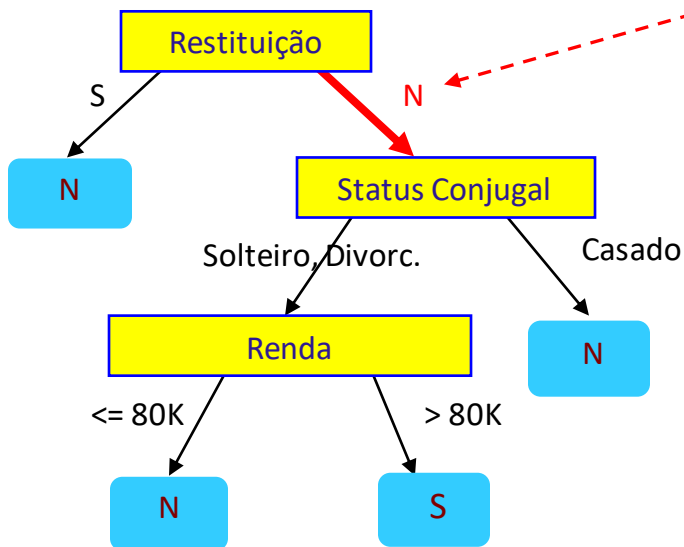
Restituição	Status Conjugal	Renda	Calote ?
N	Casado	80K	?



# Aplicando o Modelo aos Dados de Teste

Dados de Teste

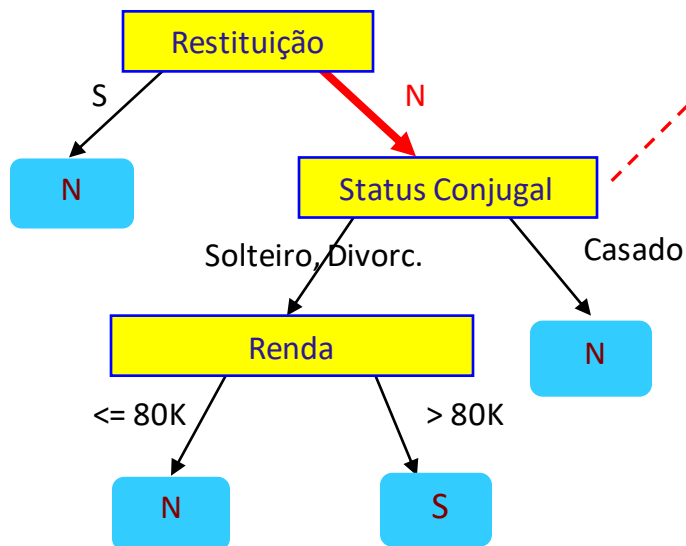
Restituição	Status Conjugal	Renda	Calote ?
N	Casado	80K	?



# Aplicando o Modelo aos Dados de Teste

Dados de Teste

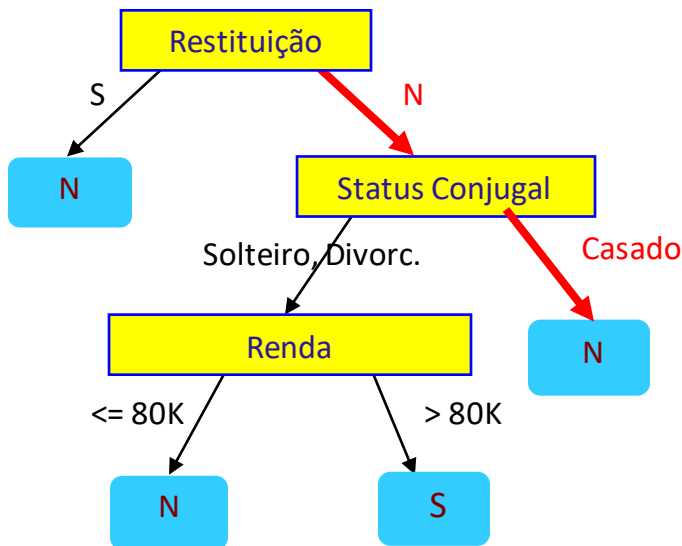
Restituição	Status Conjugal	Renda	Calote ?
N	Casado	80K	?



# Aplicando o Modelo aos Dados de Teste

Dados de Teste

Restituição	Status Conjugal	Renda	Calote ?
N	Casado	80K	?

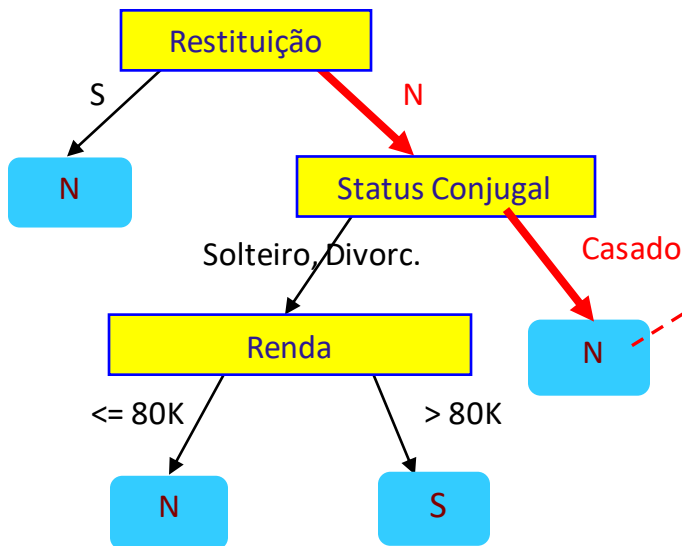




# Aplicando o Modelo aos Dados de Teste

Dados de Teste

Restituição	Status Conjugal	Renda	Calote ?
N	Casado	80K	?



Atribuir "N" para Calote

# Indução de Árvores de Decisão

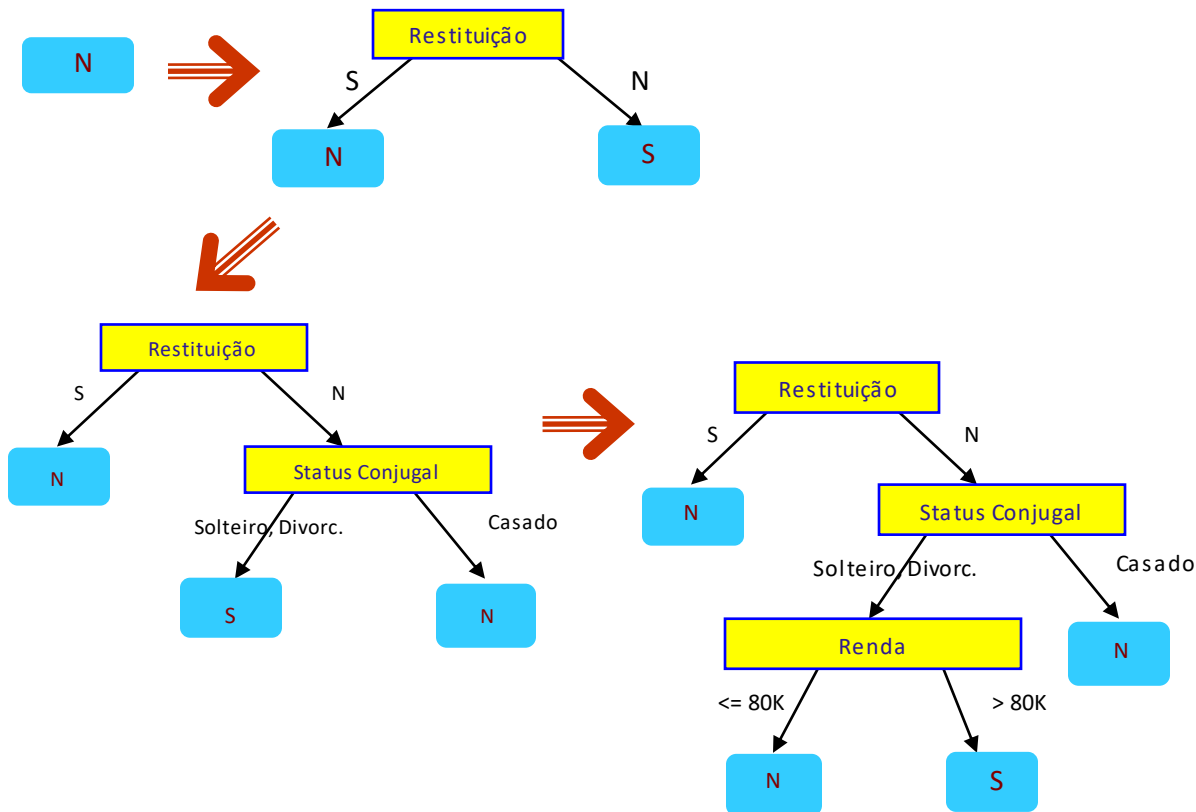
- Descobrir “árvore ótima” é problema NP-Difícil
- Muitas heurísticas para gerar árvores
  - Top-Down
  - Bottom-Up
  - Híbrida
  - Algoritmos Evolutivos
  - etc.

# Indução Top-Down

- Algoritmo de Hunt

- Assuma que  $D_t$  é o conjunto de exemplos de treino que chega ao nó  $t$
- Assuma que  $y = \{y_1, \dots, y_c\}$  são os rótulos das classes
- Passo 1:
  - Se todas instâncias em  $D_t$  pertencem a mesma classe  $y_t$ , então  $t$  é um nó folha rotulado como  $y_t$
- Passo 2:
  - Se  $D_t$  contém instâncias de mais de uma classe, **um teste sobre determinado atributo** é selecionado para particionar os registros em sub-conjuntos menores. Um nó é criado para cada resultado do teste e as instâncias em  $D_t$  são distribuídas por estes nós de acordo com os resultados. Aplicar algoritmo **recursivamente para cada nó gerado**.

# Algoritmo de Hunt



Tid	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

# Indução Top-Down

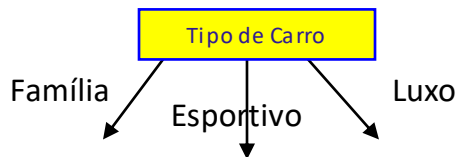
- Estratégia Recursiva
- Estratégia Gulosa (*greedy*)
  - Divide os registros com base em teste sobre atributo que otimiza localmente determinado critério
- Questões de Projeto
  - Determinar como particionar os dados
    - Como filtrar os dados com base em um atributo?
    - Como escolher o atributo a ser utilizado?
  - Determinar quando parar de particionar

# Como filtrar os dados com base em um atributo?

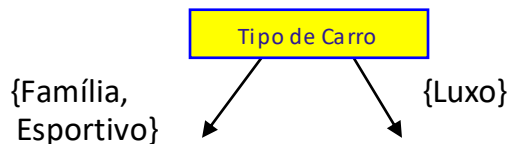
- Depende do tipo de atributo
  - Nominal
  - Ordinal
  - Contínuo
- Depende do número de divisões desejado
  - Binária
  - Múltipla

# Divisão para atributos categóricos nominais

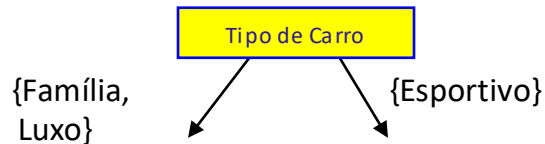
- **Múltipla:** dividir com base no número de categorias



- **Binária:** agregar categorias em dois sub-conjuntos. Necessário encontrar a divisão ótima.



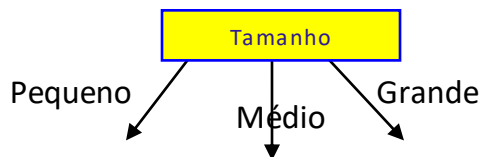
OU



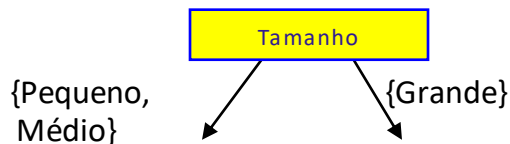
OU ....

# Divisão para atributos categóricos ordinais

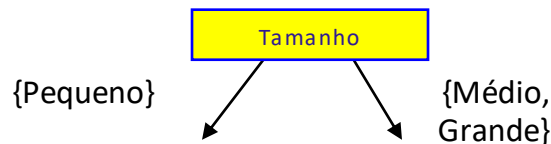
- **Múltipla**: dividir com base no número de categorias



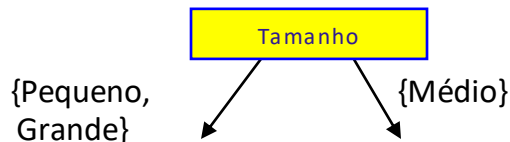
- **Binária**: agregar categorias em dois sub-conjuntos. Necessário encontrar a divisão ótima.



OU



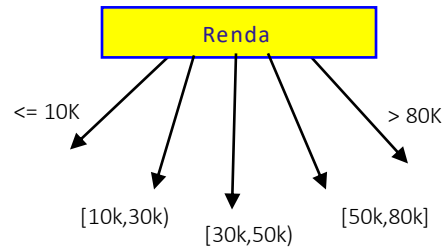
- E esta divisão?



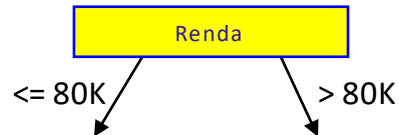


# Divisão para atributos contínuos

- **Múltipla:** discretizar os valores em intervalos



- **Binária:** definir ponto de divisão

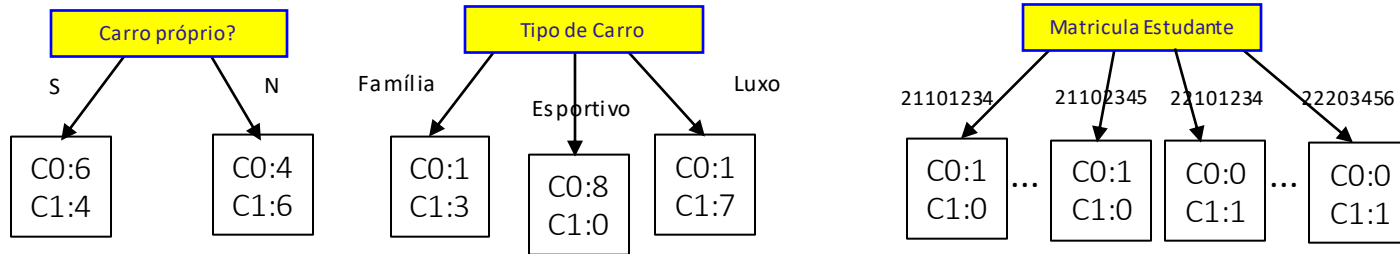


# Indução Top-Down

- Estratégia Recursiva
- Estratégia Gulosa (*greedy*)
  - Divide os registros com base em teste sobre atributo que otimiza localmente determinado critério
- Questões de Projeto
  - Determinar como particionar os dados
    - Como filtrar os dados com base em um atributo?
    - Como escolher o atributo a ser utilizado?
  - Determinar quando parar de particionar

# Como escolher o atributo?

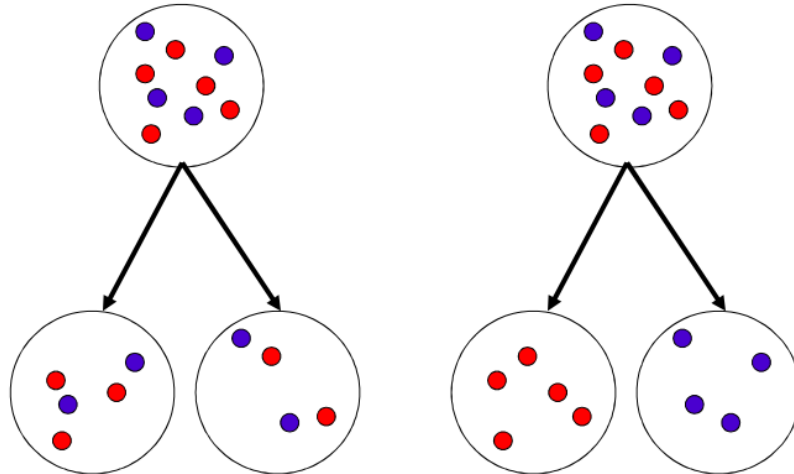
Antes da divisão: 10 exemplos da classe 0  
10 exemplos da classe 1



Qual atributo é melhor para dividir os dados?

## Como escolher o atributo?

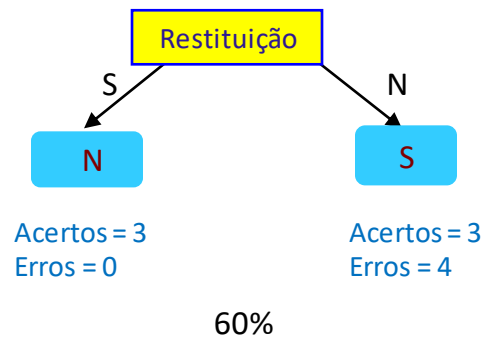
- Estratégia gulosa
  - Dar preferência a nós com distribuição de classe **homogênea**
  - Para tanto, precisamos de uma medida para quantificar **impureza**!



# Qual o melhor atributo?

*categorico*  
*Categorico*  
*contínuo*  
*classe*

<i>Tid</i>	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S



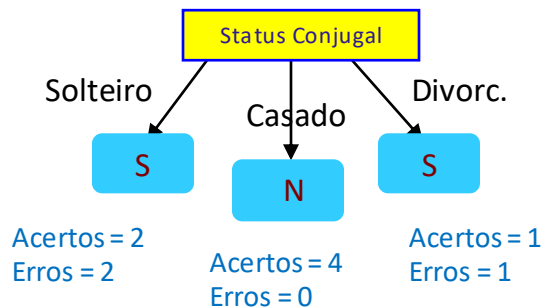
Conjunto de Treino

# Qual o melhor atributo?

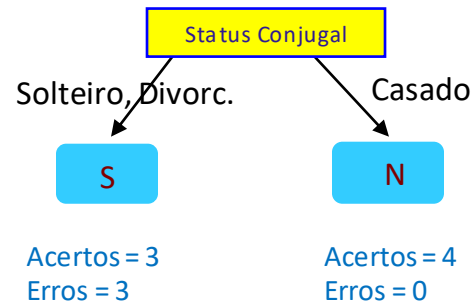
*categorico*  
*Categorico*  
*contínuo*  
*classe*

<i>Tid</i>	Restituição	Status Conjugal	Rendim. Tributáveis	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Conjunto de Treino



70%



70%

# Qual o melhor atributo?

*categorico*  
*Categorico*  
*contínuo*  
*classe*

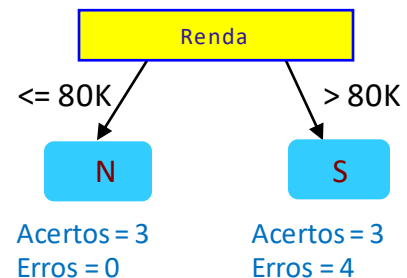
Tid	Restituição	Status Conjugal	Rendim. Tributáveis	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Conjunto de Treino

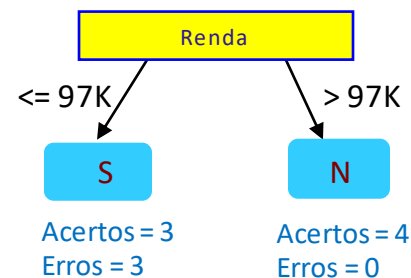
Rendim. Tributáveis	Calote ?
60K	N
70K	N
75K	N
85K	S
90K	S
95K	S
100K	N
120K	N
125K	N
220K	N

← 80K

← 97K



60%



70%

# Medidas para Impureza de Nodos

- Índice Gini

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

$$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

- Entropia

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

- Erros de classificação

$$Error(t) = 1 - \max_i P(i|t)$$



# Índice Gini

- Índice Gini para um nó  $t$ : 
$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

$p(j|t)$  é a frequência relativa da classe  $j$  no nó  $t$

- Valor **máximo**:  $1 - \frac{1}{c}$  (quando classes forem equiprováveis)
- Valor **mínimo**: **0** (quando todas instâncias pertencem à mesma classe)

C1	<b>0</b>
C2	<b>6</b>
<b>Gini=0.000</b>	

C1	<b>1</b>
C2	<b>5</b>
<b>Gini=0.278</b>	

C1	<b>2</b>
C2	<b>4</b>
<b>Gini=0.444</b>	

C1	<b>3</b>
C2	<b>3</b>
<b>Gini=0.500</b>	

# Índice Gini

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

C1	<b>0</b>
C2	<b>6</b>

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$Gini = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$$

C1	<b>1</b>
C2	<b>5</b>

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278$$

C1	<b>2</b>
C2	<b>4</b>

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$Gini = 1 - (2/6)^2 - (4/6)^2 = 0.444$$

## Computando uma divisão com o Índice Gini

- Quando um nó  $p$  é dividido em  $k$  partições (filhos), a qualidade dessa divisão é dada por:

$$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

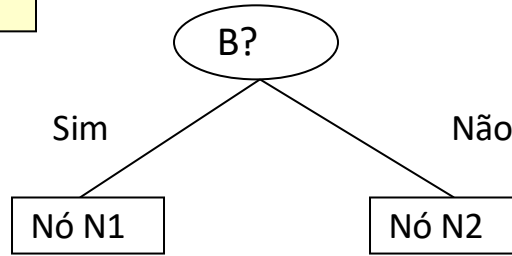
onde,

$n_i$  = número de exemplos no filho  $i$

$n$  = número de exemplos no nó pai  $p$

# Computando Índice Gini para Atributos Binários

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$



	<b>Pai</b>
C1	<b>6</b>
C2	<b>6</b>
<b>Gini = 0.500</b>	

$$Gini(N1) = 1 - [(5/7)^2 + (2/7)^2] = 0.4082$$

	<b>N1</b>	<b>N2</b>
C1	<b>5</b>	<b>1</b>
C2	<b>2</b>	<b>4</b>

$$Gini(N2) = 1 - [(1/5)^2 + (4/5)^2] = 0.32$$

$$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

$$Gini(divisão) = [(7/12) * 0.4082] + [(5/12) * 0.32] = 0.37145$$

# Árvore elementar: Calculando o Índice GINI

<i>Tid</i>	<i>Restituição</i>	<i>Status Conjugal</i>	<i>Renda</i>	<i>Calote ?</i>
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

N

Acertos = 7

Erros = 3

70%

$$\text{Gini} = 1 - (7/10)^2 - (3/10)^2$$

$$\text{Gini} = 1 - 49/100 - 9/100$$

$$\text{Gini} = (100 - 49 - 9)/100$$

$$\text{Gini} = 0,42$$

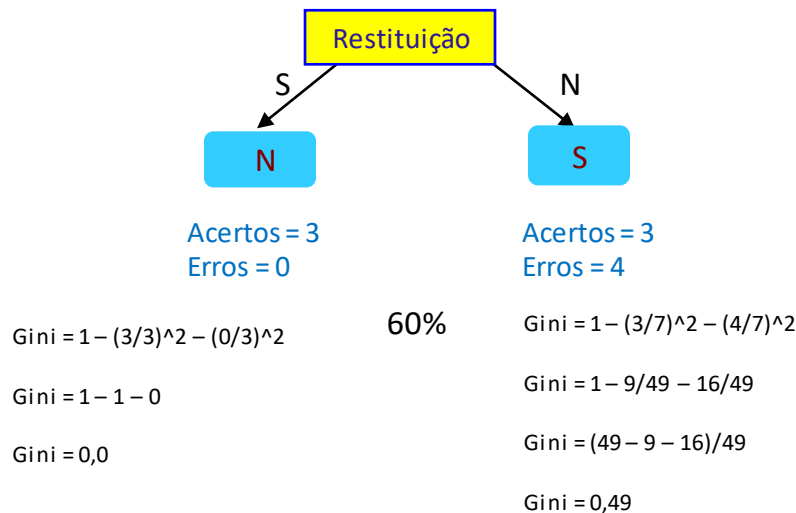
Conjunto de Treino

# Atributos Categóricos: Calculando o Índice GINI

*categórico*  
*Categórico*  
*contínuo*  
*classe*

Tid	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Conjunto de Treino



$$Gini_{split} = (3/10) * 0,0 + (7/10) * 0,49$$

$$Gini_{split} = 0 + 0,34$$

$$Gini_{split} = 0,34$$

# Atributos Categóricos: Calculando Índice GINI

- Para cada valor distinto, apurar população para cada classe do conjunto de dados
- Usar a matriz com populações para tomar a decisão

Particionamento em n ramos

	TipoVeículo		
	Familiar	Esportivo	Luxo
C1	1	2	1
C2	4	1	1
Gini	0.393		

Particionamento em 2 ramos  
(busca pela melhor divisão de valores)

	TipoVeículo			TipoVeículo	
	{Esportivo, Luxo}	{Familiar}		{Esportivo}	{Familiar, Luxo}
C1	3	1	C1	2	2
C2	2	4	C2	1	5
Gini	0.400		Gini	0.419	

# Atributos Categóricos: Calculando o Índice GINI

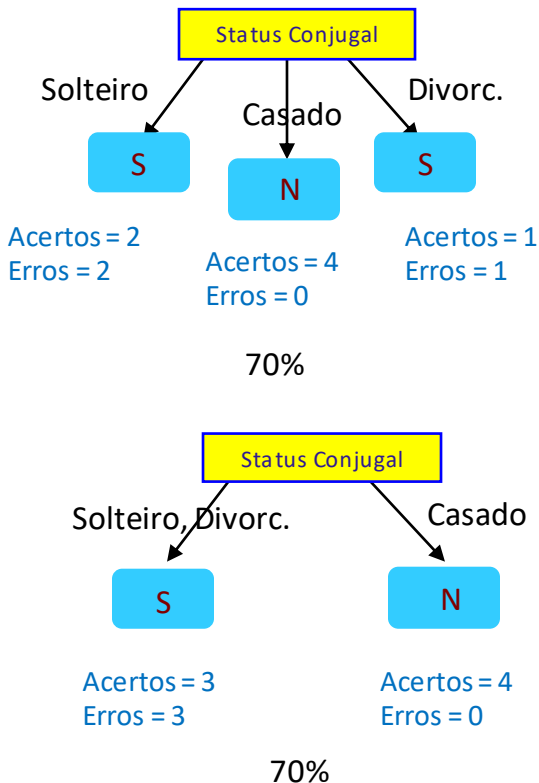
*categórico*  
*Categórico*  
*contínuo*  
*classe*

<i>Tid</i>	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Conjunto de Treino

$Gini_{split} = 0,3$

$Gini_{split} = 0,3$





# Atributos Contínuos: Calculando o Índice GINI

- Para a eficiência computacional: para cada atributo,
  - Classificar valores existentes
  - Pesquisar linearmente estes valores, apurando a população envolvida, e calculando o índice GINI
  - Escolher a posição de particionamento que apresenta o menor índice GINI

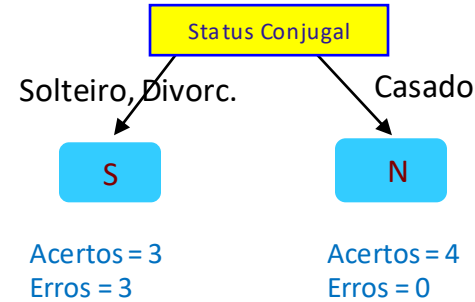
Calote		N		N		N		S		S		S		N		N		N		N	
		Renda																			
Valores Ordenados		60		70		75		85		90		95		100		120		125		220	
Posições de		55		65		72		80		87		92		97		110		122		172	
Particionamento		<= >		<= >		<= >		<= >		<= >		<= >		<= >		<= >		<= >		<= >	
S		0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0
N		0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1
Gini		0.420		0.400		0.375		0.343		0.417		0.400		<u>0.300</u>		0.343		0.375		0.400	

## Induzindo o 2o. Nível da árvore de decisão

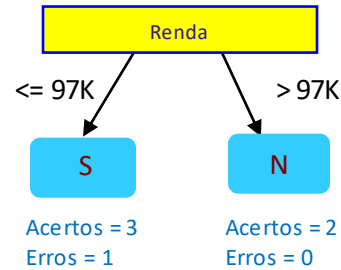
*categorico*  
*Categorico*  
*contínuo*  
*classe*

Tid	Restituição	Status Conjugal	Renda	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	80K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

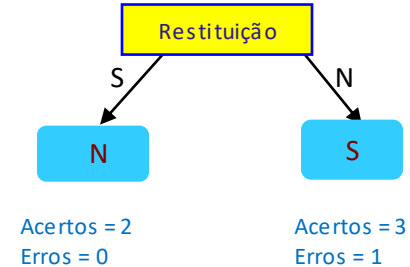
Conjunto de Treino



Aqui parou.



Ginisplit = 0,25



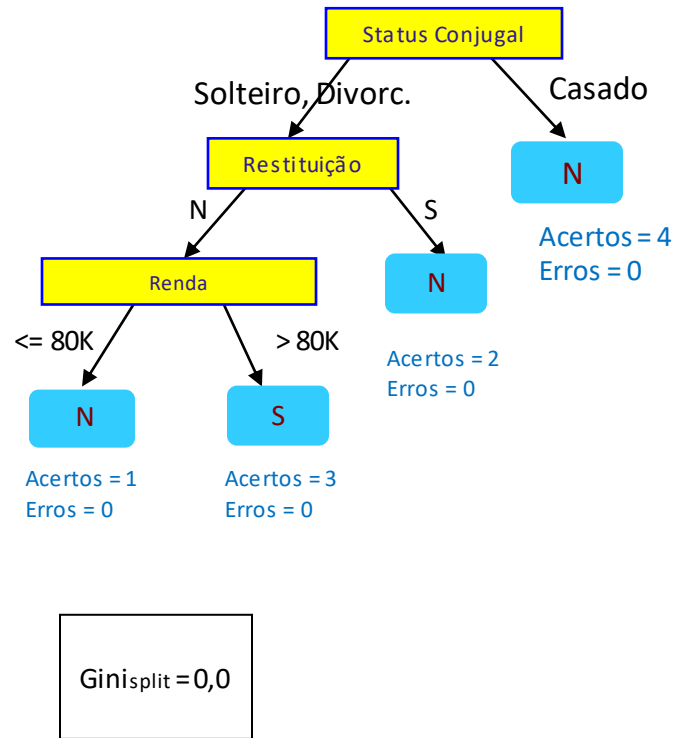
Ginisplit = 0,25

## Induzindo o 3o. Nível da árvore de decisão

*categorico*  
*Categorico*  
*contínuo*  
*classe*

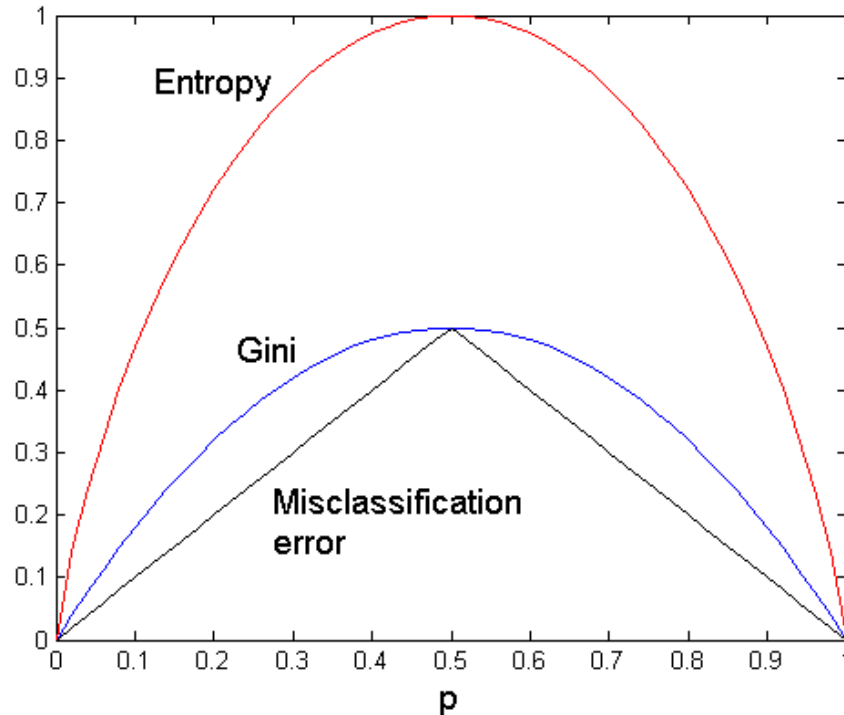
Tid	Restituição	Status Conjugal	Rendim. Tributáveis	Calote ?
1	S	Solteiro	125K	N
2	N	Casado	100K	N
3	N	Solteiro	70K	N
4	S	Casado	120K	N
5	N	Divorc.	95K	S
6	N	Casado	60K	N
7	S	Divorc.	220K	N
8	N	Solteiro	85K	S
9	N	Casado	75K	N
10	N	Solteiro	90K	S

Conjunto de Treino



# Comparação entre os critérios de divisão

Para um problema de 2 classes:



# Indução Top-Down

- Estratégia Recursiva
- Estratégia Gulosa (*greedy*)
  - Divide os registros com base em teste sobre atributo que otimiza localmente determinado critério
- Questões de Projeto
  - Determinar como particionar os dados
    - Como filtrar os dados com base em um atributo?
    - Como escolher o atributo a ser utilizado?
  - **Determinar quando parar de particionar**

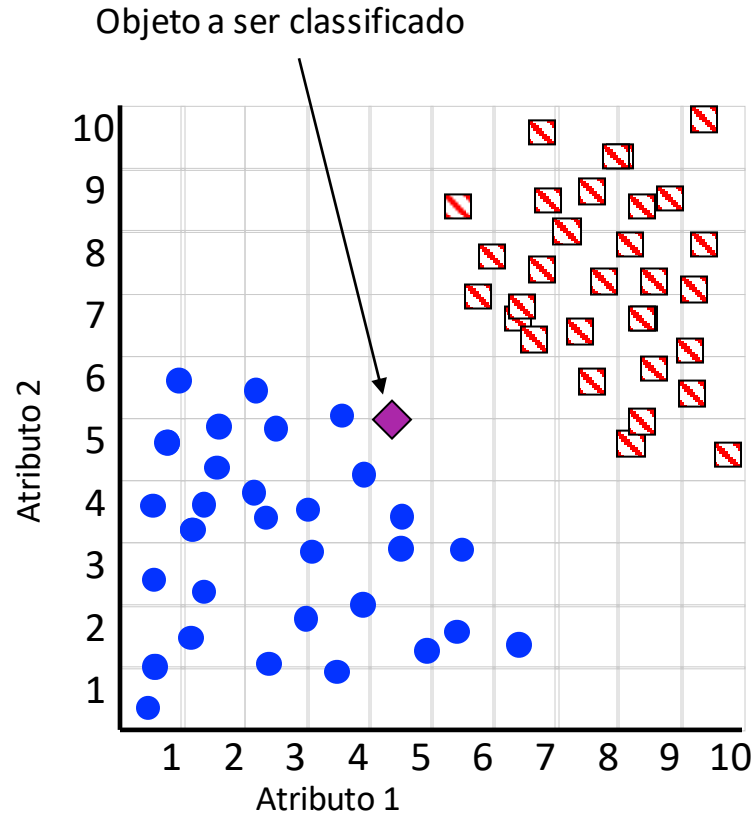
# Critérios de Parada para Indução Top-Down

- Parar de expandir nós quando:
  - Todas instâncias forem da mesma classe (**homogeneidade de classe**)
  - Todos valores de atributos forem iguais (**homogeneidade de instâncias**)
  - Attingir **valor satisfatório do critério de divisão** (parâmetro)
  - Attingir **profundidade máxima** (parâmetro)
  - ...

# Vantagens e Desvantagens de Árvores de Decisão

- Vantagens:
  - Fácil de compreender (muito utilizadas por médicos!)
  - Possível gerar regras com base nas árvores
  - Custo baixo de geração do modelo:  $O(m \cdot N \log N)$
  - Extremamente rápida para classificar novas instâncias
- Desvantagens:
  - Podem tornar-se **muito grandes**
  - Sujeitas a **overfitting** (super-ajuste aos dados)
  - Geram apenas hiperplanos paralelos aos eixos
    - Logo, não lidam bem com atributos correlacionados (por quê?)
  - Solução localmente ótima pode estar longe do ótimo global

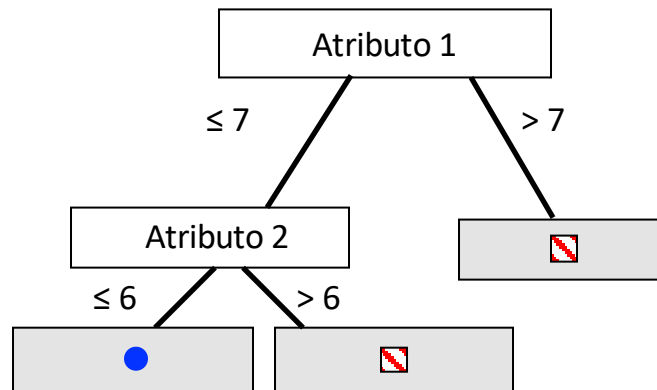
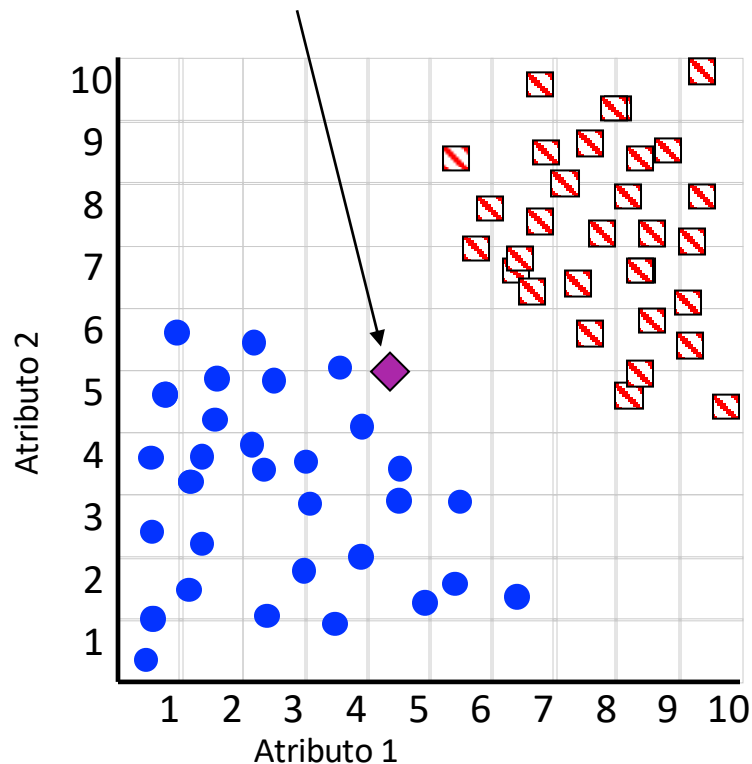
# Visualização Geométrica de uma Árvore de Decisão





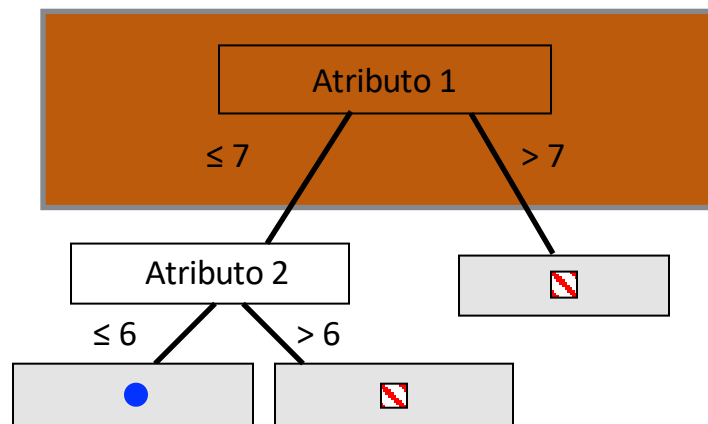
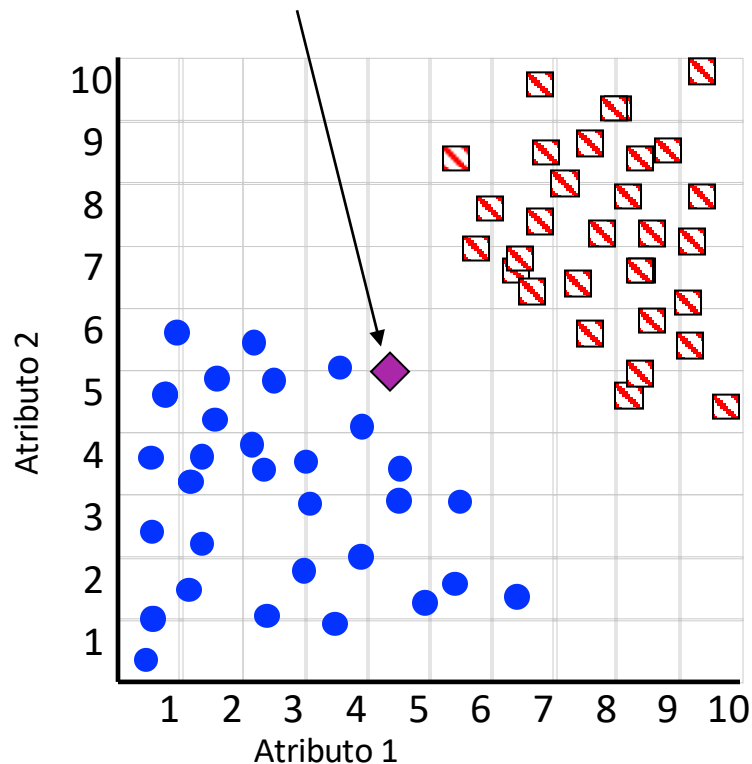
# Visualização Geométrica de uma Árvore de Decisão

Objeto a ser classificado



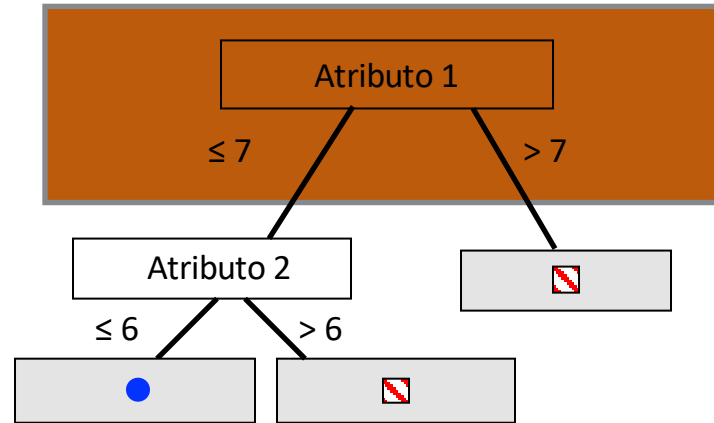
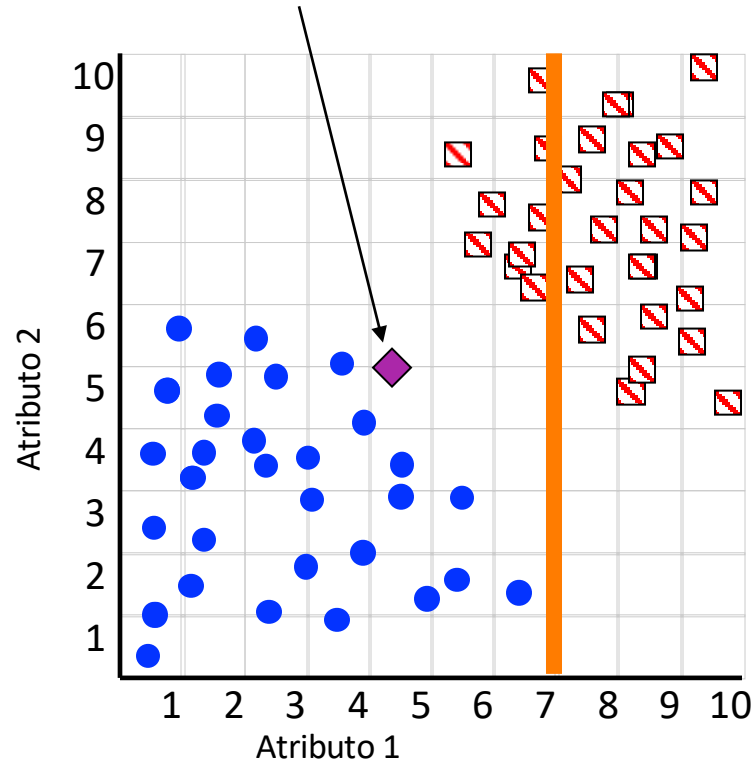
# Visualização Geométrica de uma Árvore de Decisão

Objeto a ser classificado



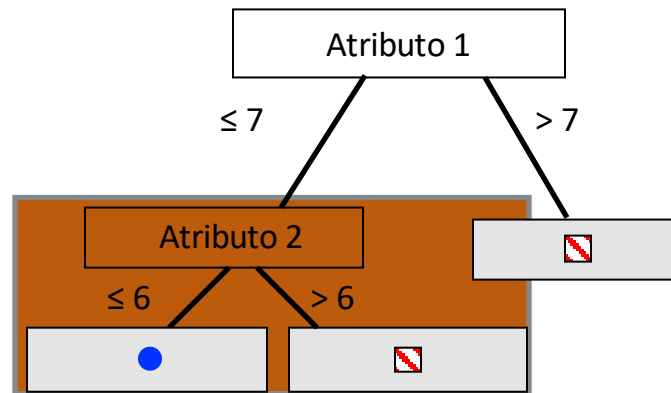
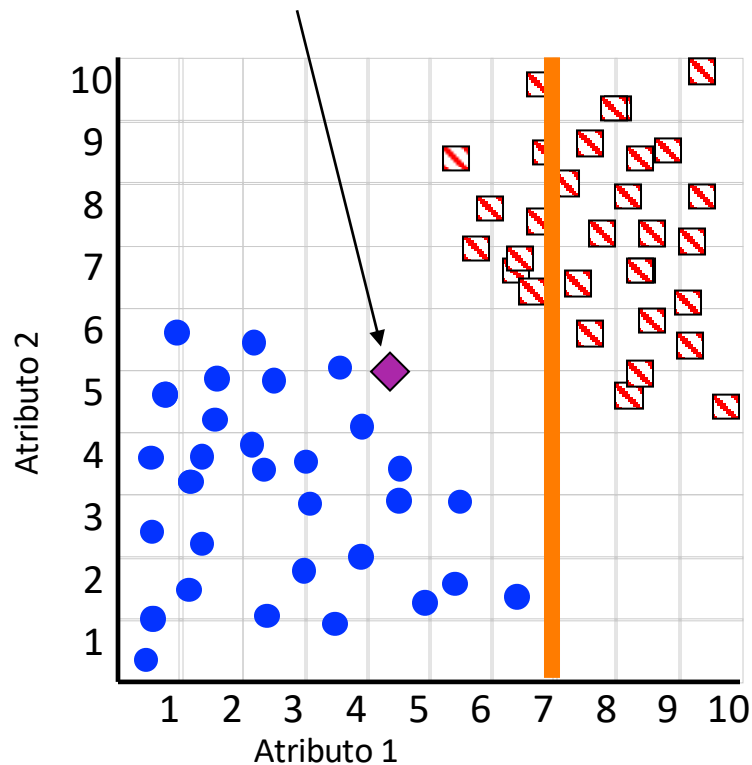
# Visualização Geométrica de uma Árvore de Decisão

Objeto a ser classificado

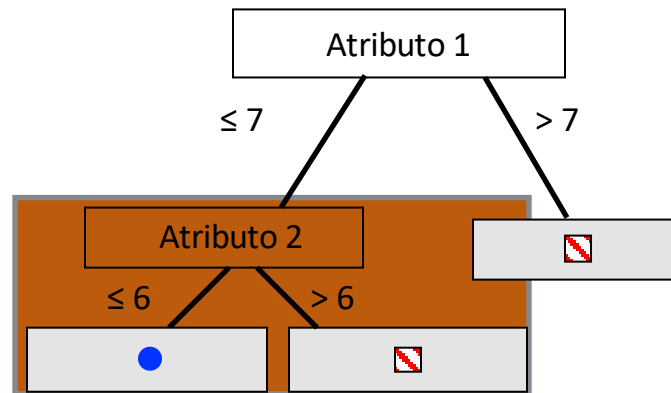
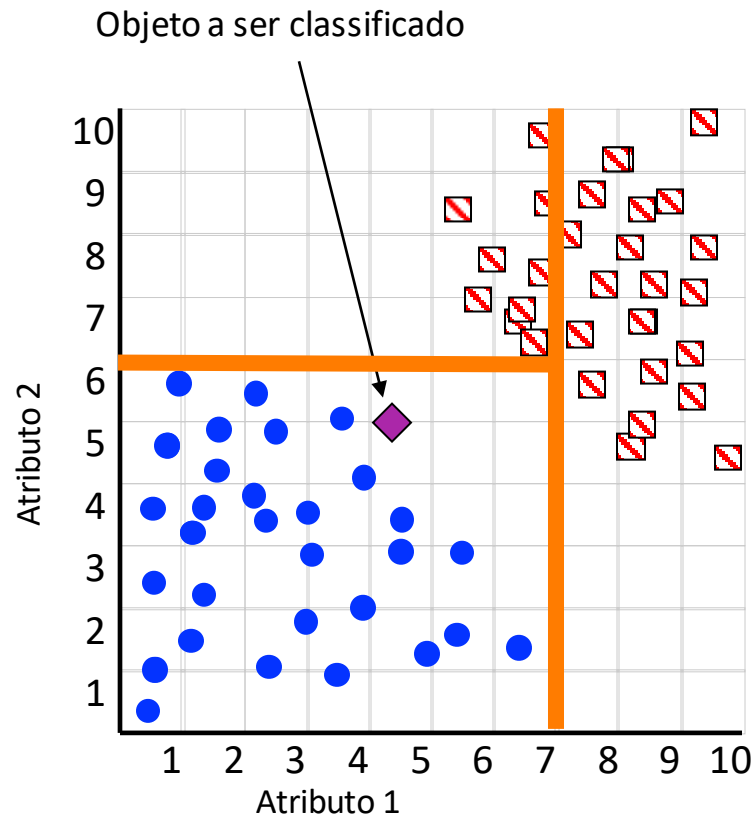


# Visualização Geométrica de uma Árvore de Decisão

Objeto a ser classificado



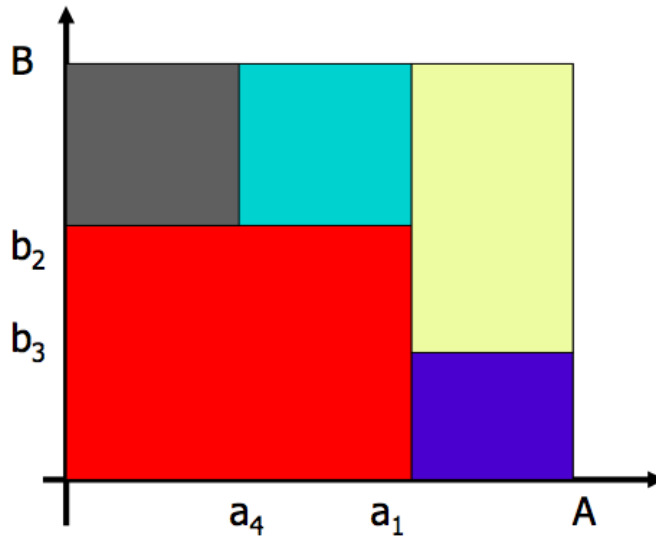
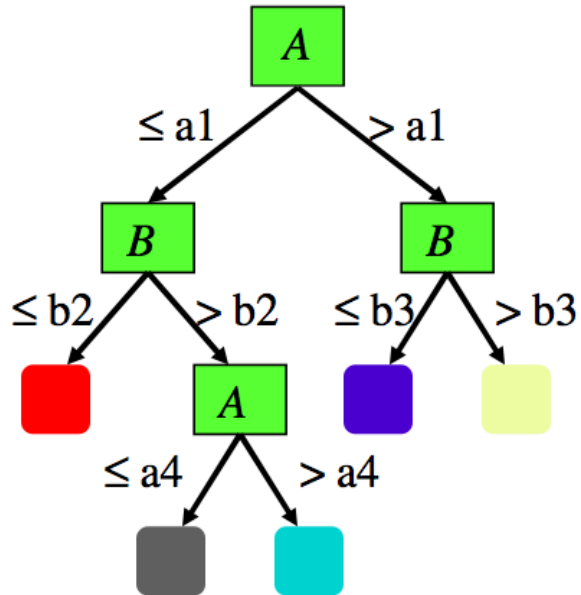
# Visualização Geométrica de uma Árvore de Decisão



# Espaço de Hipóteses

- Cada percurso da raiz até o nó folha representa uma **regra de classificação**
- Cada nó folha
  - Está associado a uma classe
  - Corresponde a **uma região do domínio dos atributos**
    - **Hiper-retângulo**
    - Intersecção de hiper-retângulos é vazia
    - União é o espaço total

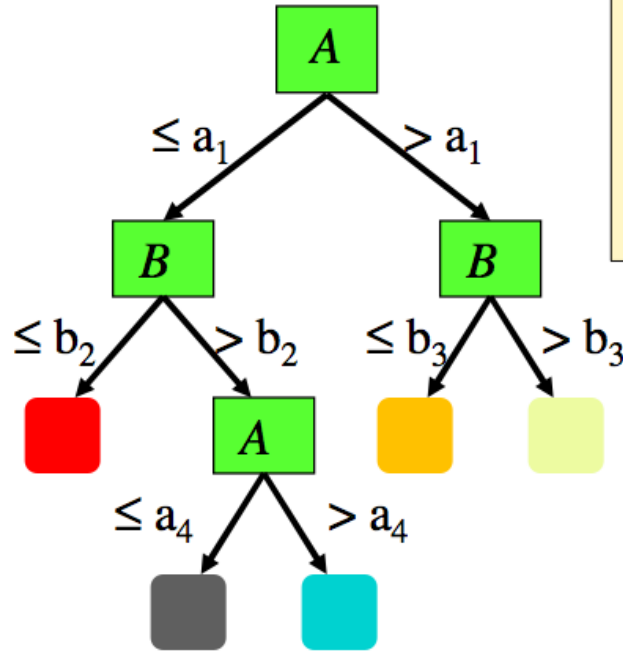
# Espaço de Hipóteses



# De árvores para regras

Regras: disjunções de conjunções lógicas

1. **Se**  $A \leq a_1$  **E**  $B \leq b_2$  **Então** Classe = Vermelha  
**OU**
2. **Se**  $A > a_1$  **E**  $B \leq b_3$  **Então** Classe = Laranja  
**OU**
- ...



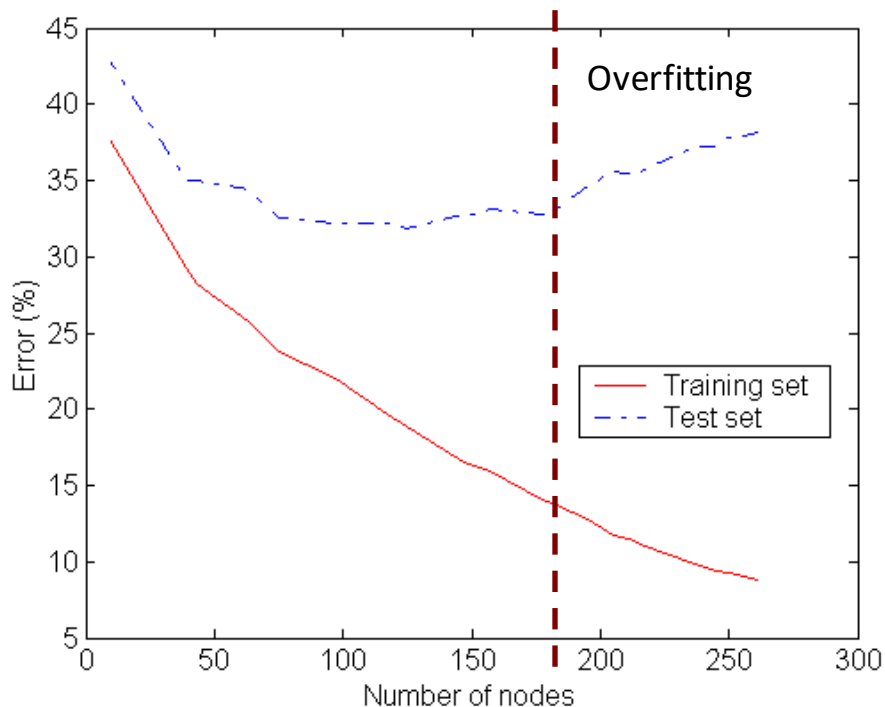
Exercício: complete as regras !



# Busca no Espaço de Hipóteses

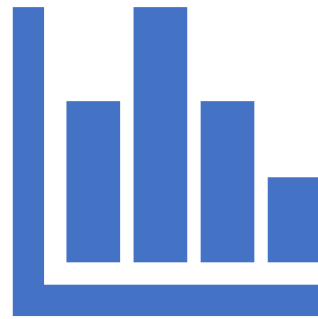
- Não há **backtracking**
  - Impureza é minimizada localmente em cada nó!
    - Suposição: soma dos ótimos locais aproxima bem o ótimo global
- Espaço de hipóteses completo
  - A função objetivo certamente está contida nele
  - Sem *bias* de restrição
    - Proporcionando chances de *overfitting*
  - Com ***bias de busca*** (preferência)
    - Árvores com atributos que geram maior redução de impureza estão acima na árvore
    - Tal *bias* implica em tendência para árvores mais curtas

# Underfitting and Overfitting



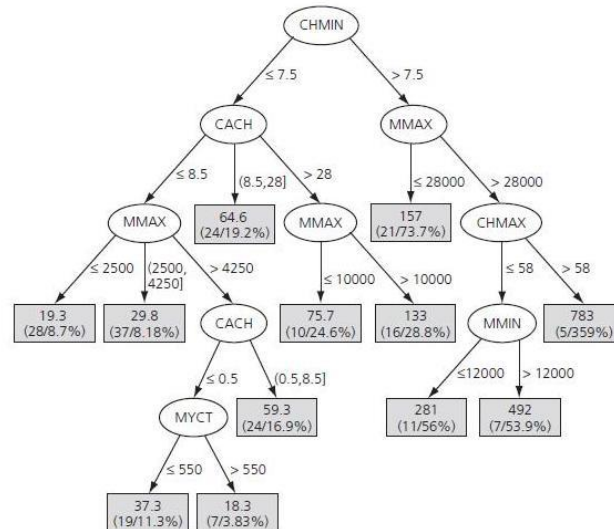
Underfitting: quando o modelo é simples demais, ambos erros, de treino e de teste, são grandes.

# Regressão



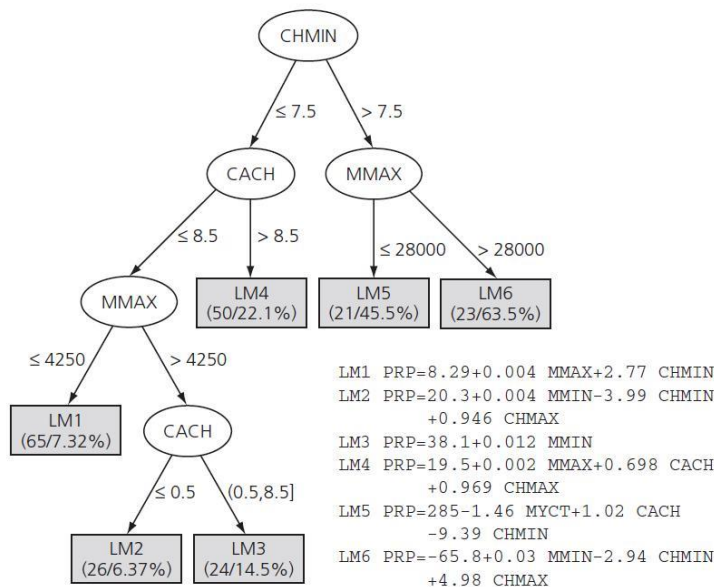
# Árvores de Decisão para Problemas de Regressão

- Árvores de Regressão
  - Folha contém **média dos valores** do atributo alvo dos exemplos de treino que chegam até lá



# Árvores de Decisão para Problemas de Regressão

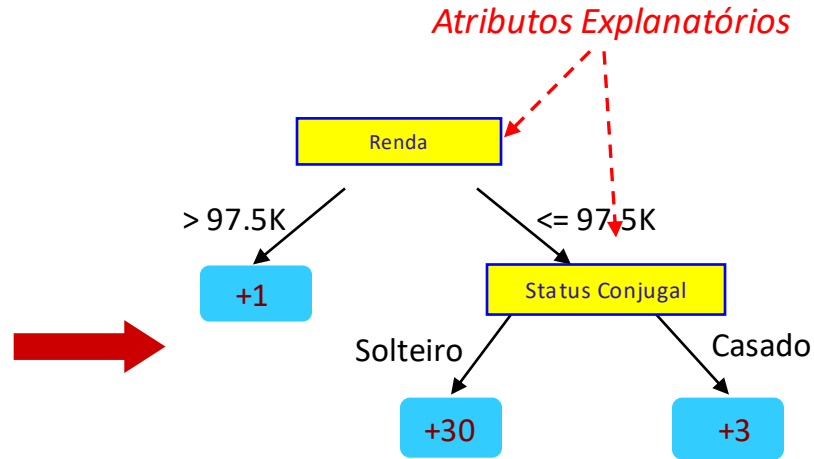
- Árvores de Modelos
  - Folha contém **função de regressão (não-)linear** calculada sobre as instâncias que chegam até lá



# Exemplo de Árvore de Regressão

*categorico*  
*Categorico*  
*contínuo*  
*alvo*

<i>Tid</i>	<i>Restituição</i>	<i>Status Conjugal</i>	<i>Renda</i>	<i>Atraso</i>
1	S	Solteiro	125K	0
2	N	Casado	100K	1
3	N	Solteiro	70K	30
4	S	Casado	120K	2
5	N	Solteiro	95K	24
6	N	Casado	60K	3
7	S	Solteiro	220K	1
8	N	Solteiro	85K	36
9	N	Casado	75K	3
10	N	Solteiro	90K	30



Conjunto de Treino

Modelo: Árvore de Regressão

# Exemplo de Árvore de Regressão

<i>Tid</i>	Restitui ção	Status Conjugal	Renda	Atraso	Atraso Predito	Diferen ça
1	S	Solteiro	125K	0	1	1
2	N	Casado	100K	1	1	0
3	N	Solteiro	70K	30	30	0
4	S	Casado	120K	2	1	1
5	N	Solteiro	95K	24	30	6
6	N	Casado	60K	3	3	0
7	S	Solteiro	220K	1	1	0
8	N	Solteiro	85K	36	30	6
9	N	Casado	75K	3	3	0
10	N	Solteiro	90K	30	30	0

Erro médio absoluto:

$$(1 + 1 + 6 + 6)/10 = 1,4$$

Raiz do erro médio quadrático:

$$\text{SQRT}((1+1+36+36)/10) = 2,72$$

# Exemplo de Árvore Modelo

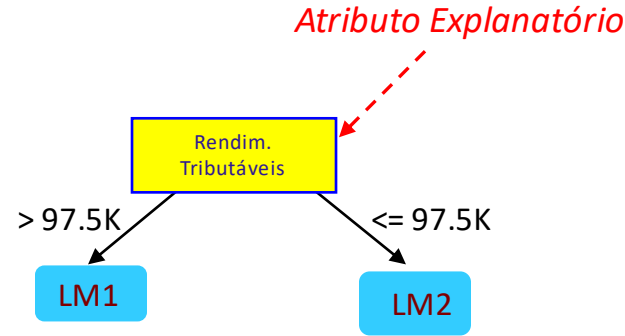
<i>Tid</i>	<i>Restituição</i>	<i>Status Conjugual</i>	<i>Rendim. Tributáveis</i>	<i>Atraso</i>
1	S	Solteiro	125K	0
2	N	Casado	100K	1
3	N	Solteiro	70K	30
4	S	Casado	120K	2
5	N	Solteiro	95K	24
6	N	Casado	60K	3
7	S	Solteiro	220K	1
8	N	Solteiro	85K	36
9	N	Casado	75K	3
10	N	Solteiro	90K	30

categorico

Categorico

contínuo

alvo



LM num: 1

Delay =

$$18.2396 * \text{Status Conjugual}=\text{Solteiro} \\ - 0.1611 * \text{Rendim. Tributáveis} \\ + 16.2853$$

LM num: 2

Delay =

$$24.2168 * \text{Status Conjugual}=\text{Solteiro} \\ - 0.1458 * \text{Rendim. Tributáveis} \\ + 15.401$$

Conjunto de Treino

Modelo: Árvore de Regressão



# Exemplo de Árvore Modelo

<i>Tid</i>	Restituição	Status Conjugal	Renda	Atraso	Atraso Predito	Diferença
1	S	Solteiro	125K	0	14,3874	14,3874
2	N	Casado	100K	1	0,1753	0,8247
3	N	Solteiro	70K	30	29,4118	0,5882
4	S	Casado	120K	2	-3,0467	5,0467
5	N	Solteiro	95K	24	25,7668	1,7668
6	N	Casado	60K	3	6,653	3,653
7	S	Solteiro	220K	1	-0,9171	1,9171
8	N	Solteiro	85K	36	27,2248	8,7752
9	N	Casado	75K	3	4,466	1,466
10	N	Solteiro	90K	30	26,4958	3,5042

Erro médio absoluto:  
4,193

Raiz do erro médio  
quadrático: 5,874

# Árvores de Decisão para Problemas de Regressão

- Principal mudança: medida de divisão de nós
  - Exemplo: standard deviation reduction (SDR)
    - Mesma fórmula genérica do “ganho”
    - Em vez de entropia ou Gini, apenas calcular o desvio padrão do atributo alvo para as instâncias de cada nó e ponderá-las pelas frequências

$$SDR = SD(v_{pai}) - \sum_{t=1}^k \frac{N(v_t)}{N} SD(v_t)$$

# Árvore elementar: Calculando o Índice GINI

*categorico*  
*Categorico*  
*contínuo*  
*alvo*

<i>Tid</i>	<i>Restitui ção</i>	<i>Status Conjugal</i>	<i>Renda</i>	<i>Atraso</i>
1	S	Solteiro	125K	0
2	N	Casado	100K	1
3	N	Solteiro	70K	30
4	S	Casado	120K	2
5	N	Solteiro	95K	24
6	N	Casado	60K	3
7	S	Solteiro	220K	1
8	N	Solteiro	85K	36
9	N	Casado	75K	3
10	N	Solteiro	90K	30

13

Média	13,00
Desvio Padrão	14,93

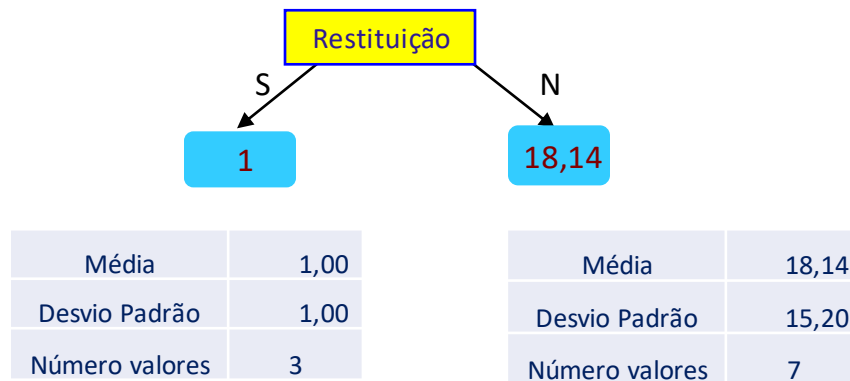
Conjunto de Treino

# Atributos Categóricos: Calculando o Índice GINI

*categórico*  
*Categórico*  
*contínuo*  
*alvo*

Tid	Restituição	Status Conjugal	Renda	Atraso
1	S	Solteiro	125K	0
2	N	Casado	100K	1
3	N	Solteiro	70K	30
4	S	Casado	120K	2
5	N	Solteiro	95K	24
6	N	Casado	60K	3
7	S	Solteiro	220K	1
8	N	Solteiro	85K	36
9	N	Casado	75K	3
10	N	Solteiro	90K	30

Conjunto de Treino



$$\text{SDR} = 14,93 - 3/10 * 1,0 - 7/10 * 15,2$$

$$\text{SDR} = 14,93 - 0,3 - 10,64$$

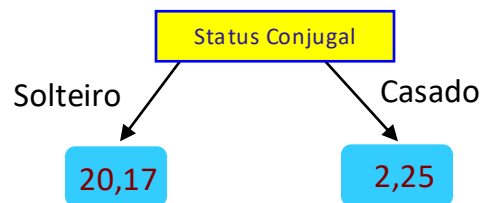
$$\text{SDR} = 3,99$$

## Atributos Categóricos: Calculando o Índice GINI

*categorico*  
*Categorico*  
*contínuo*  
*alvo*

Tid	Restituição	Status Conjugal	Renda	Atraso
1	S	Solteiro	125K	0
2	N	Casado	100K	1
3	N	Solteiro	70K	30
4	S	Casado	120K	2
5	N	Solteiro	95K	24
6	N	Casado	60K	3
7	S	Solteiro	220K	1
8	N	Solteiro	85K	36
9	N	Casado	75K	3
10	N	Solteiro	90K	30

Conjunto de Treino



Média	20,17
Desvio Padrão	15,70
Número valores	6

Média	2,25
Desvio Padrão	0,96
Número valores	4

$$\text{SDR} = 14,93 - 6/10 * 15,7 - 4/10 * 0,96$$

$$\text{SDR} = 14,93 - 9,42 - 0,38$$

$$\text{SDR} = 5,13$$

## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.

The screenshot shows the Weka Classifier Tree Visualizer window. The decision tree structure is as follows:

- Root Node: `Restitucao=`
  - Left Branch: `NAO`
    - Node: `Calote=`
      - Left Branch: `SIM` → Leaf Node: **95.0**
      - Right Branch: `NAO` → Node: `Estado_Civil=`
        - Left Branch: `SOLTEIRO` → Leaf Node: [ ]
        - Right Branch: `CASADO, DIVORCIADO` → Leaf Node: [ ]
  - Right Branch: `SIM` → Node: `Estado_Civil=`
    - Left Branch: `SOLTEIRO, CASADO` → Leaf Node: [ ]
    - Right Branch: `DIVORCIADO` → Leaf Node: [ ]

The data table below the tree is:

No.	Restitucao Nominal	Estado_Civil Nominal	Calote Nominal	Rendimento_Anual Numeric	Rend_Anual Predito Numeric
1	SIM	Solteiro	NAO	125.0	
2	NAO	Casado	NAO	100.0	
3	NAO	Solteiro	NAO	70.0	
4	SIM	Casado	NAO	120.0	
5	NAO	Divorciado	SIM	95.0	
6	NAO	Casado	NAO	60.0	
7	SIM	Divorciado	NAO	220.0	

A red arrow points from the value 95.0 in the table (row 5) to the leaf node labeled 95.0 in the decision tree.

ERRO

$\Sigma$  [ ]

## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.

The image shows two windows from the Weka software. The top window, 'Weka Classifier Tree Visualizer', displays a decision tree. The root node is 'Restituicao=' with branches 'NAO' and 'SIM'. The 'NAO' branch leads to a node 'Calote=' with branches 'SIM' (leading to leaf 95.0) and 'NAO' (leading to a node 'Estado\_Civil=' with branches 'SOLTEIRO' (leading to leaf 70.0) and 'CASADO, DIVORCIADO' (leading to an empty leaf). The 'SIM' branch leads to a node 'Estado\_Civil=' with branches 'SOLTEIRO, CASADO' (leading to an empty leaf) and 'DIVORCIADO' (leading to an empty leaf). The bottom window, 'Viewer', shows a table with 7 rows and 5 columns. A red arrow points from the '70.0' value in the 'Rendimento\_Anual' column of row 3 to the '70.0' leaf node in the tree.

Tree View

Restituicao=

NAO

SIM

Calote=

SIM

95.0

NAO

Estado\_Civil=

SOLTEIRO

70.0

CASADO, DIVORCIADO

Estado\_Civil=

SOLTEIRO, CASADO

DIVORCIADO

Viewer

Relation: TAN\_DecisionTree-weka.filters.unsupervised.attribute.Remove-R1

No.	Restituicao Nominal	Estado_Civil Nominal	Calote Nominal	Rendimento_Anual Numeric	Rend_Anual Predito Numeric
1	SIM	Solteiro	NAO	125.0	
2	NAO	Casado	NAO	100.0	
3	NAO	Solteiro	NAO	70.0	
4	SIM	Casado	NAO	120.0	
5	NAO	Divorciado	SIM	95.0	
6	NAO	Casado	NAO	60.0	
7	SIM	Divorciado	NAO	220.0	

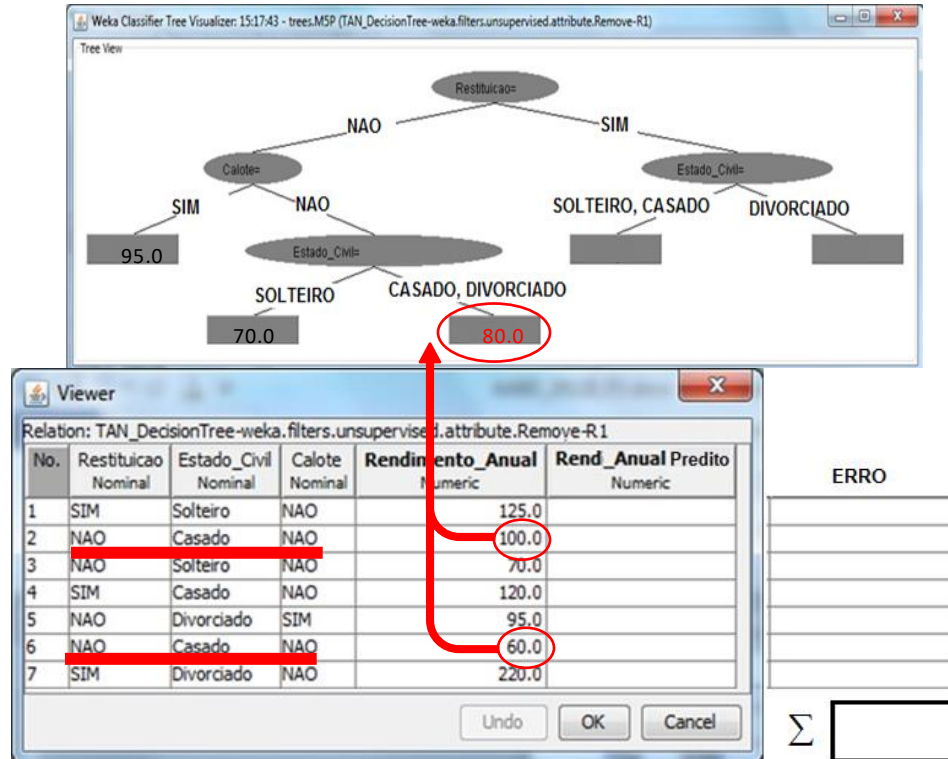
ERRO

$\Sigma$

## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.





## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.

The image shows two windows from the Weka software. The top window, 'Weka Classifier Tree Visualizer', displays a decision tree. The root node is 'Restituicao='. It branches into 'NAO' and 'SIM'. The 'NAO' branch leads to a node 'Calote=' which branches into 'SIM' (leaf node 95.0) and 'NAO' (leaf node 70.0). The 'SIM' branch leads to a node 'Estado\_Civil=' which branches into 'SOLTEIRO, CASADO' (leaf node 122.5) and 'DIVORCIADO' (leaf node 80.0). The bottom window, 'Viewer', shows the training data table. The table has columns: No., Restituicao Nominal, Estado\_Civil Nominal, Calote Nominal, Rendimento\_Anual Numeric, and Rend\_Anual Predito Numeric. The data rows are numbered 1 to 7. Red lines and circles highlight the data points used to calculate the leaf node values: row 1 (SIM, Solteiro, NAO, 125.0) for the 95.0 leaf, row 2 (NAO, Casado, NAO, 100.0) for the 70.0 leaf, row 4 (SIM, Casado, NAO, 120.0) for the 122.5 leaf, and row 5 (NAO, Divorciado, SIM, 95.0) for the 80.0 leaf. A red arrow points from the 122.5 value in the tree to the 120.0 value in the table. To the right of the table is a section labeled 'ERRO' with five empty rows, and a summation symbol  $\Sigma$  followed by an empty box.

No.	Restituicao Nominal	Estado_Civil Nominal	Calote Nominal	Rendimento_Anual Numeric	Rend_Anual Predito Numeric
1	SIM	Solteiro	NAO	125.0	
2	NAO	Casado	NAO	100.0	
3	NAO	Solteiro	NAO	70.0	
4	SIM	Casado	NAO	120.0	
5	NAO	Divorciado	SIM	95.0	
6	NAO	Casado	NAO	60.0	
7	SIM	Divorciado	NAO	220.0	

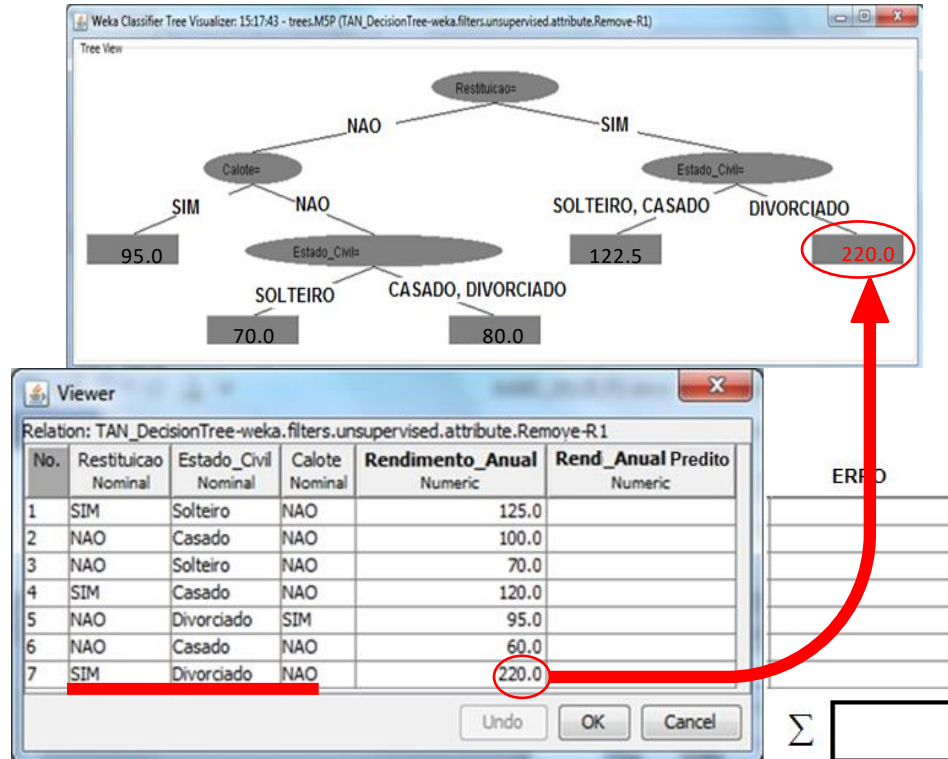
ERRO

$\Sigma$

## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

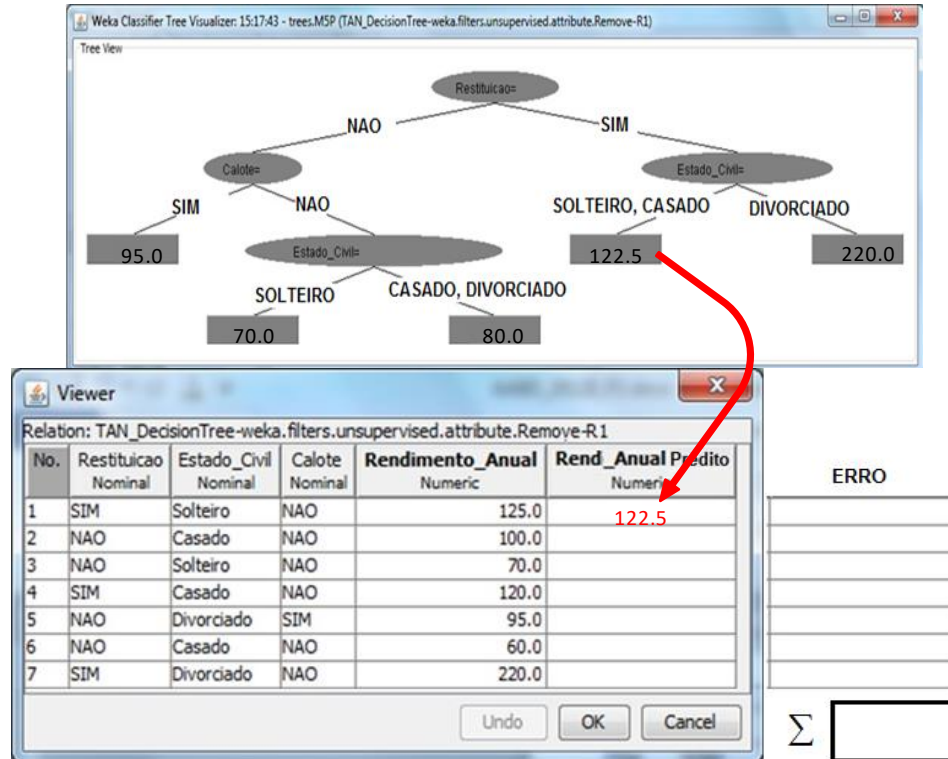
1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.



## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

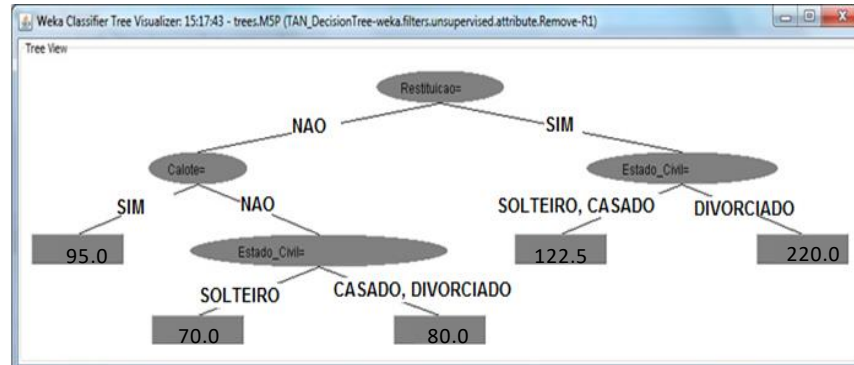
1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.



## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.



Viewer

Relation: TAN\_DecisionTree-weka.filters.unsupervised.attribute.Remove-R1

No.	Restituicao Nominal	Estado_Civil Nominal	Calote Nominal	Rendimento_Anual Numeric	Rend_Anual Predito Numeric
1	SIM	Solteiro	NAO	125.0	122.5
2	NAO	Casado	NAO	100.0	80.0
3	NAO	Solteiro	NAO	70.0	70.0
4	SIM	Casado	NAO	120.0	122.5
5	NAO	Divorciado	SIM	95.0	95.0
6	NAO	Casado	NAO	60.0	80.0
7	SIM	Divorciado	NAO	220.0	220.0

Undo OK Cancel

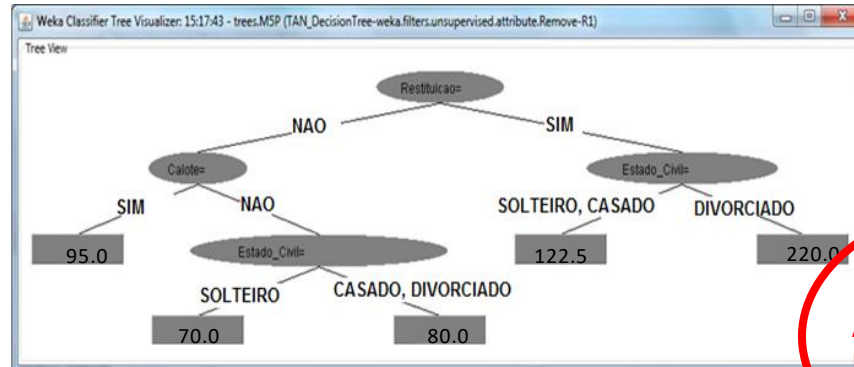
ERRO


$\Sigma$

## Exemplo Ilustrativo

Considere a seguinte árvore de regressão e a tabela logo a seguir, cujo atributo alvo é `Rendimento_Anual`.

1: Preencha os valores para os nodos folha da árvore, a partir da tabela abaixo.



Viewer

Relation: TAN\_DecisionTree-weka.filters.unsupervised.attribute.Remove-R1

No.	Restituicao Nominal	Estado_Civil Nominal	Calote Nominal	Rendimento_Anual Numeric	Rend_Anual Predito Numeric
1	SIM	Solteiro	NAO	125.0	122.5
2	NAO	Casado	NAO	100.0	80.0
3	NAO	Solteiro	NAO	70.0	70.0
4	SIM	Casado	NAO	120.0	122.5
5	NAO	Divorciado	SIM	95.0	95.0
6	NAO	Casado	NAO	60.0	80.0
7	SIM	Divorciado	NAO	220.0	220.0

Undo OK Cancel

Atenção! Erro é sempre Positivo!!!

ERRO

2.5
20.0
0.0
2.5
0.0
20.0
0.0

$\Sigma$

6.43

45 / 7

# Exemplos de Algoritmos

- ID3 (Quinlan 1986)

- Iterative Dichotomiser 3
- Lida apenas com atributos nominais
- Medida de impureza: ganho de informação
- Tipo de poda: pré-poda (limite de instâncias)

- C4.5 (Quinlan 1993)

- J48 (Weka), C5.0 (comercial)
- Atributos discretos e contínuos
- Medida de impureza: gain ratio
- Tipo de poda: pós-poda (error-based pruning)

# Exemplos de Algoritmos

- CART (Breiman et al. 1984)
  - Classification and Regression Trees
  - Árvores de Classificação e Regressão
  - Atributos discretos e contínuos
  - Divisões sempre binárias (agrega categorias)
  - Medida de impureza: índice Gini / twoing / sum of squares
  - Tipo de poda: pós-poda (cost-complexity pruning)

# Exemplos de Algoritmos

- M5 (Quinlan 1992)
  - M5P (Weka)
  - Árvores de Regressão e Árvores de Modelos
  - Atributos discretos e contínuos
  - Medida de impureza: SDR
  - Tipo de poda: erro corrigido (leva em conta o número de parâmetros dos modelos lineares)