Aprendizado de Máquina II

KNN e Arvore de Decisão para Regressão

Profª. Renata De Paris

Especialização em Ciência de Dados

UCS
UNIVERSIDADE
DE CAXIAS DO SUL

Roteiro da Aula

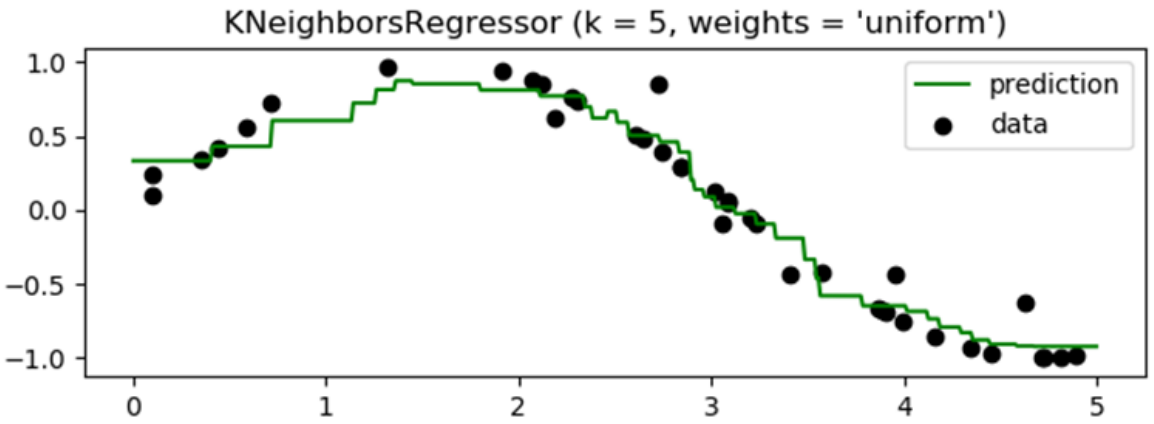
- KNN para regressão
- Árvore de Decisão para Regressão
- Avaliação do desempenho
- Atividade para entregar

KNN para Regressão

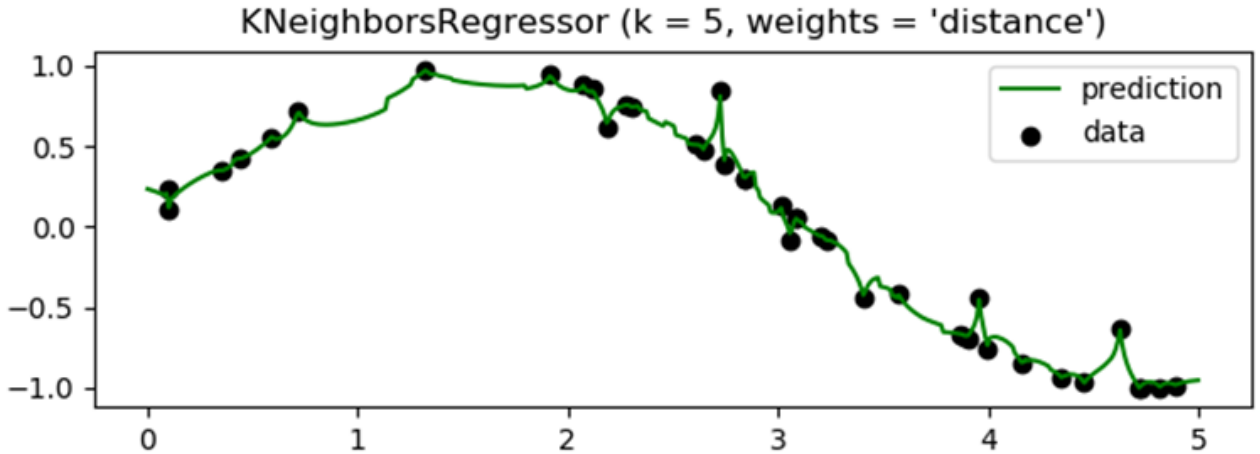
- Utiliza a distância euclidiana para calcular a distância entre as instâncias
- Diferença do KNN para classificação:
  - Ao invés de usar a classe de maior frequência dos vizinhos mais próximos, utiliza a média dos valores dessas instâncias.
- Principais estratégias:
  - Média dos k vizinhos mais próximos
    - Problema: valores com alto grau de dispersão.
    - Solução: utilizar pesos maiores para os vizinhos mais próximos – média ponderada
  - Média ponderada dos k vizinhos mais próximos

KNN para Regressão

- ***Uniform weights***: calcula o valor médio das instâncias mais próximas (k)



- ***Distance weights***: calcula a média ponderada onde vizinhos mais próximos de um ponto de consulta terão uma influência maior do que vizinhos mais distantes.



Árvore de Decisão para Regressão

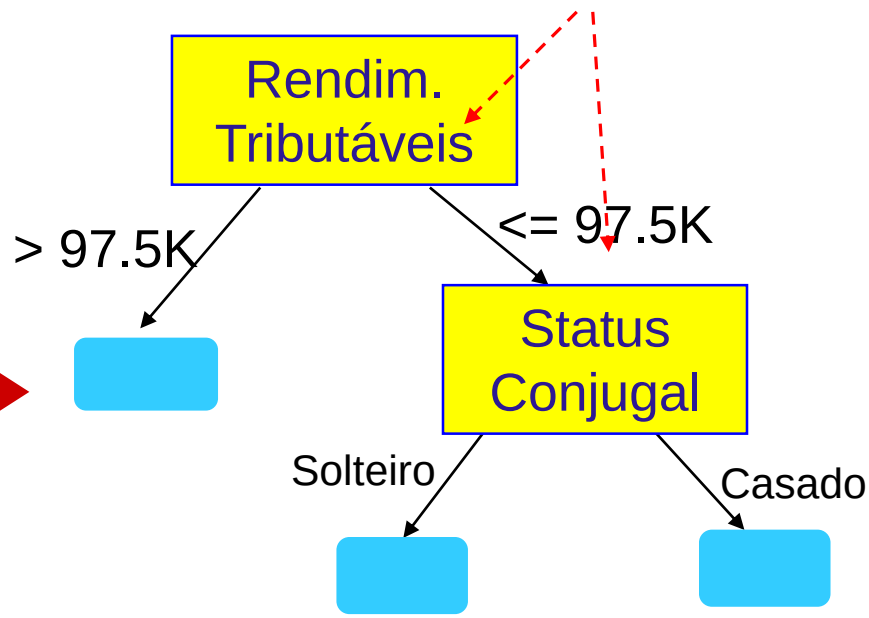
Exemplo de Árvore de Regressão

	Categórico	Categórico	Contínuo	Alvo
Tid	**Restituição**	**Status Conjugal**	**Rendim. Tributáveis**	**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

Modelo: Árvore de Regressão

Atributos Explanatórios



Nó folha contém uma constante, geralmente, uma média ou uma equação para o valor previsto de um determinado conjunto de dados.

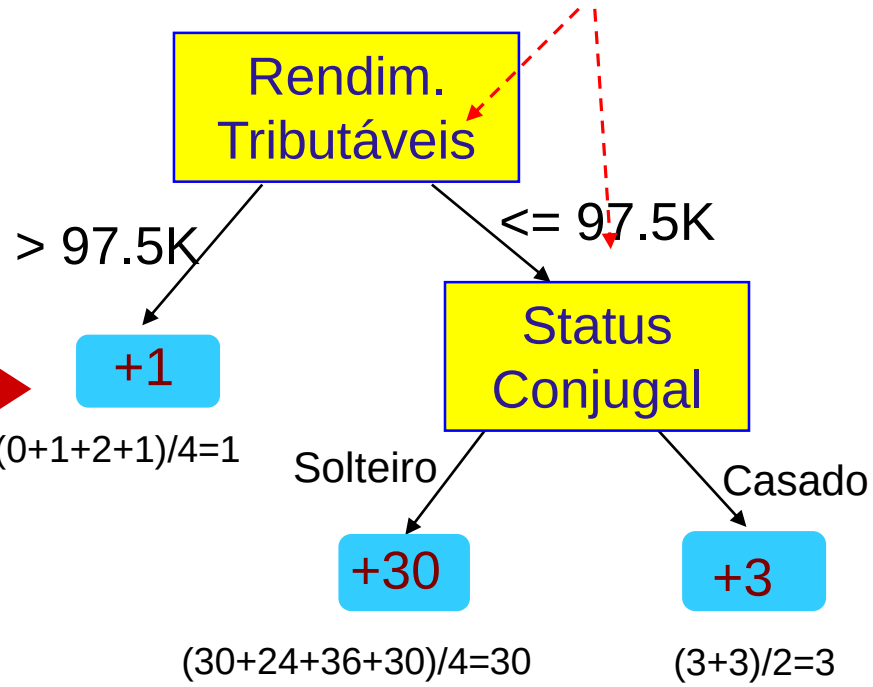
Exemplo de Árvore de Regressão

Tid	Restituição (Categórico)	Status Conjugal (Categórico)	Rendim. Tributáveis (Contínuo)	Atraso (Alvo)
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

Atributos Explanatórios



Exemplo de Árvore de Regressão

Erro Médio Absoluto (EMA):

$$EMA = \sum_{i=1}^{N} (|previsto - real|)/N$$

Tid	Restituição	Status Conjugal	Rendim. Tributáveis	Atraso	Atraso Predito	Diferença
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

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

Raiz do erro médio quadrático:

$$\sqrt{\frac{1+1+36+36}{10}} =$$

$$\sqrt{\frac{74}{10}} = \sqrt{7{,}4} =$$

$$2{,}72$$

Exemplo de Árvore Modelo

Tid	Restituição	Status Conjugal	Rendim. Tributáveis	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

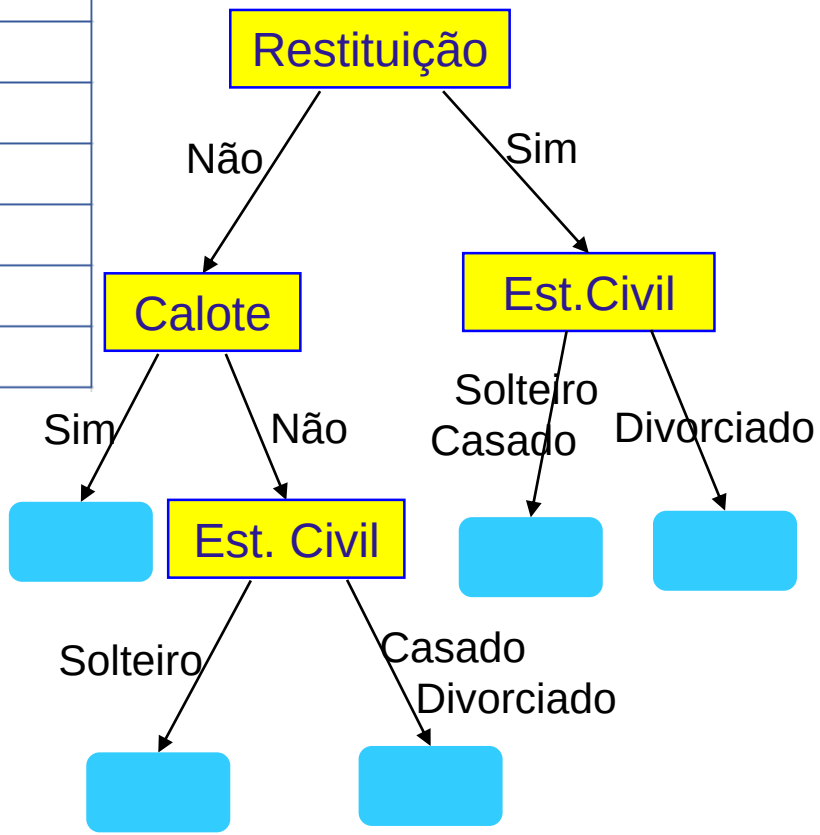
Exercícios

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

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

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0		
NÃO	Casado	NÃO	100,0		
NÃO	Solteiro	NÃO	70,0		
SIM	Casado	NÃO	120,0		
NÃO	Divorciado	SIM	95,0		
NÃO	Casado	NÃO	60,0		
SIM	Divorciado	NÃO	220,0		

EMA = _____



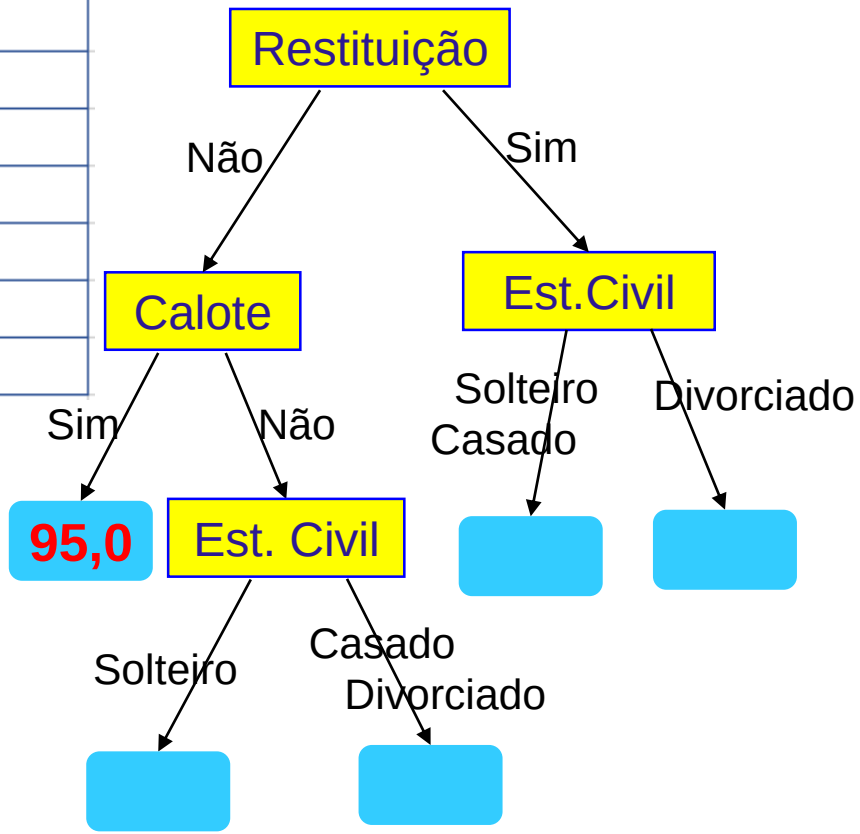
Exercícios

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

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

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0		
NÃO	Casado	NÃO	100,0		
NÃO	Solteiro	NÃO	70,0		
SIM	Casado	NÃO	120,0		
NÃO	Divorciado	SIM	**95,0**		
NÃO	Casado	NÃO	60,0		
SIM	Divorciado	NÃO	220,0		

EMA = _____



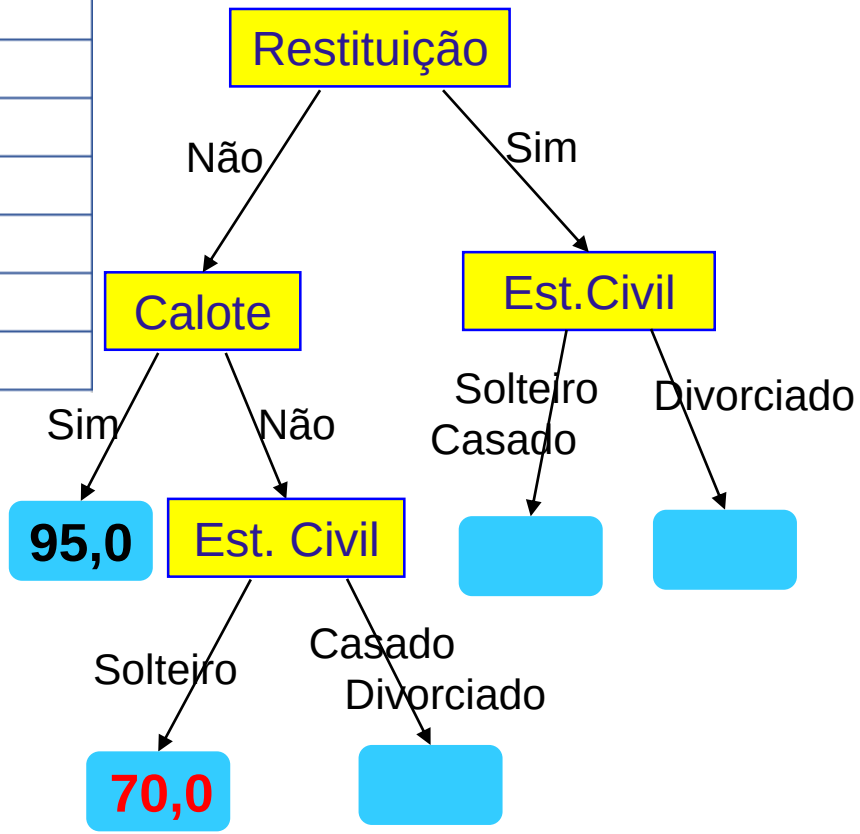
Exercícios

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

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

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0		
NÃO	Casado	NÃO	100,0		
NÃO	Solteiro	NÃO	**70,0**		
SIM	Casado	NÃO	120,0		
NÃO	Divorciado	SIM	95,0		
NÃO	Casado	NÃO	60,0		
SIM	Divorciado	NÃO	220,0		

EMA = _____



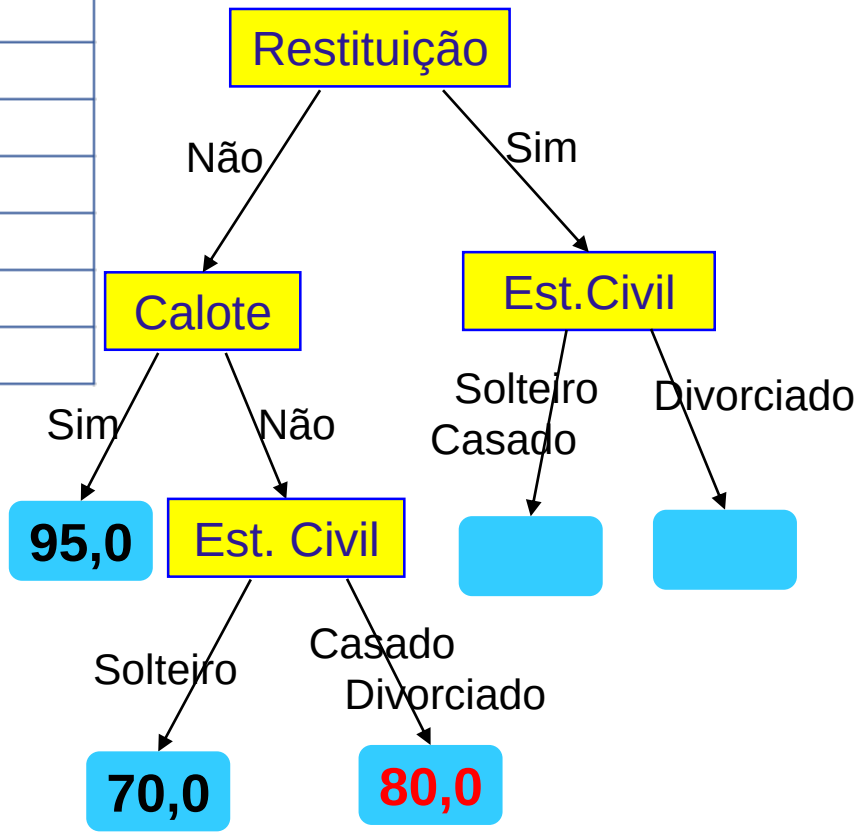
Exercícios

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

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

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0		
NÃO	Casado	NÃO	100,0		
NÃO	Solteiro	NÃO	70,0		
SIM	Casado	NÃO	120,0		
NÃO	Divorciado	SIM	95,0		
NÃO	Casado	NÃO	60,0		
SIM	Divorciado	NÃO	220,0		

EMA = _____



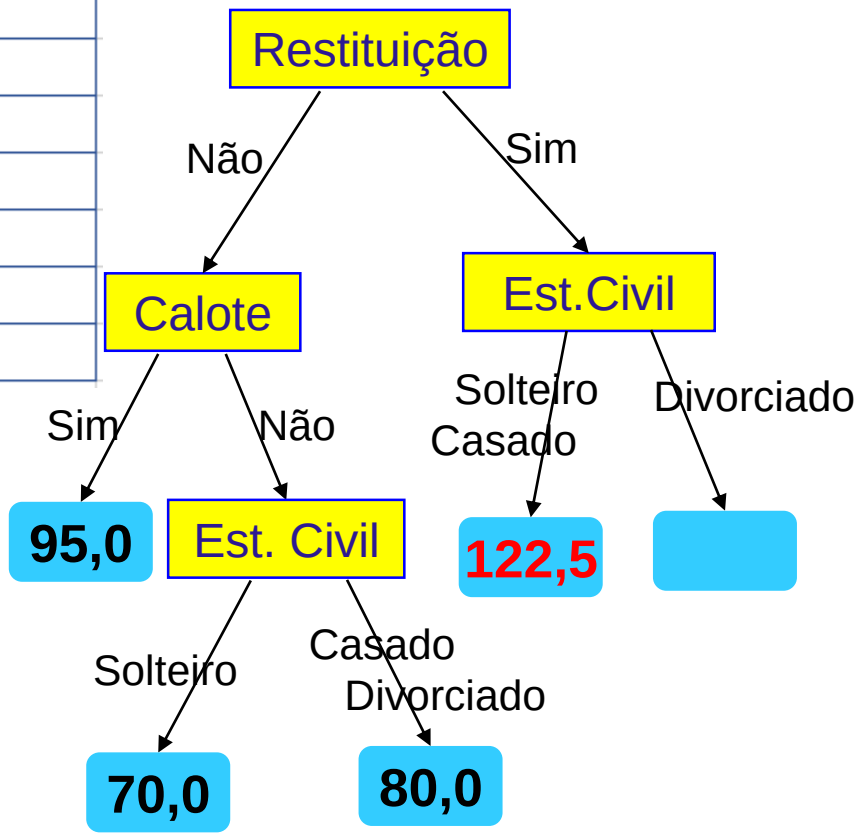
Exercícios

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

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

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0		
NÃO	Casado	NÃO	100,0		
NÃO	Solteiro	NÃO	70,0		
SIM	Casado	NÃO	120,0		
NÃO	Divorciado	SIM	95,0		
NÃO	Casado	NÃO	60,0		
SIM	Divorciado	NÃO	220,0		

EMA = _____



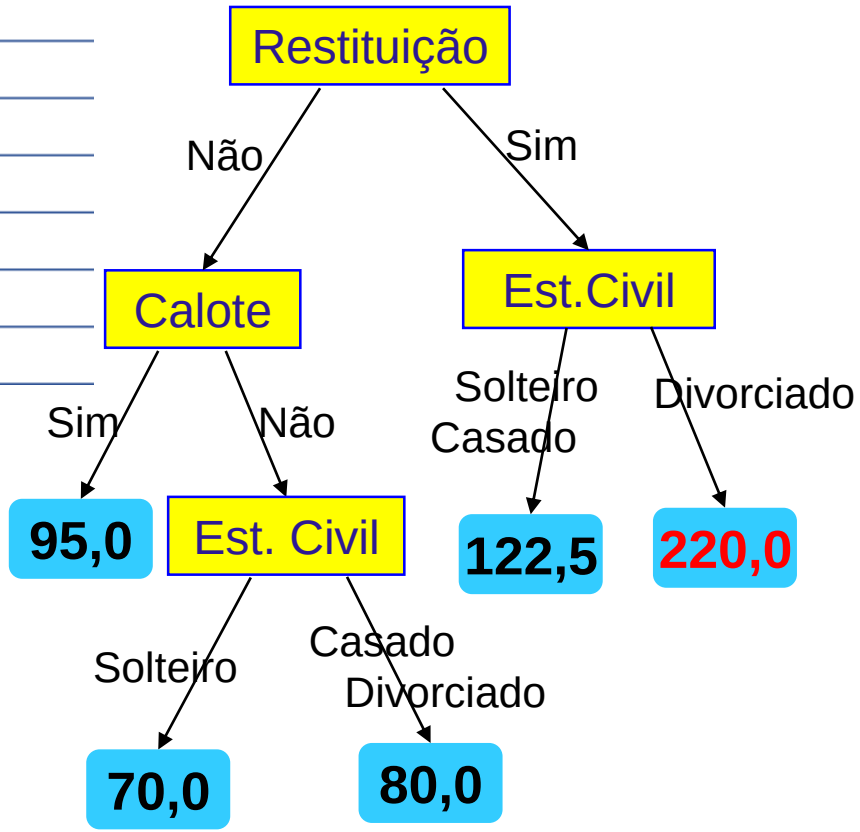
Exercícios

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

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

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0		
NÃO	Casado	NÃO	100,0		
NÃO	Solteiro	NÃO	70,0		
SIM	Casado	NÃO	120,0		
NÃO	Divorciado	SIM	95,0		
NÃO	Casado	NÃO	60,0		
SIM	Divorciado	NÃO	220,0		

EMA = _____



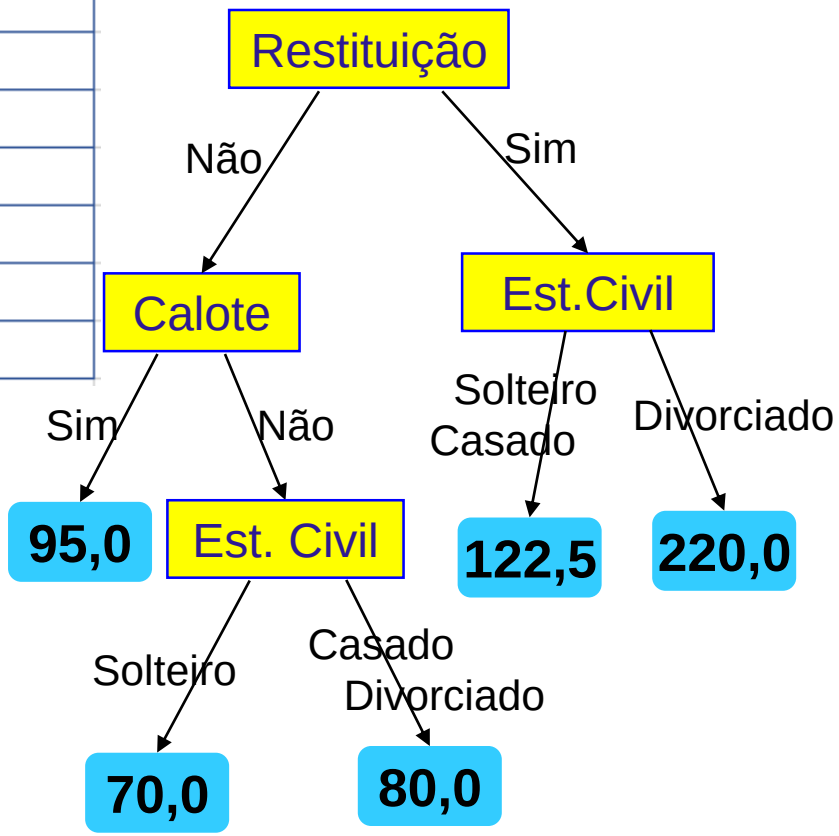
Exercícios

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

b) Preencha a coluna Rend. Anual Predito com os valores que a árvore sugere.

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0	122,5	
NÃO	Casado	NÃO	100,0	80,0	
NÃO	Solteiro	NÃO	70,0	70,0	
SIM	Casado	NÃO	120,0	122,5	
NÃO	Divorciado	SIM	95,0	95,0	
NÃO	Casado	NÃO	60,0	80,0	
SIM	Divorciado	NÃO	220,0	220,0	

EMA = _____



Exercícios

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

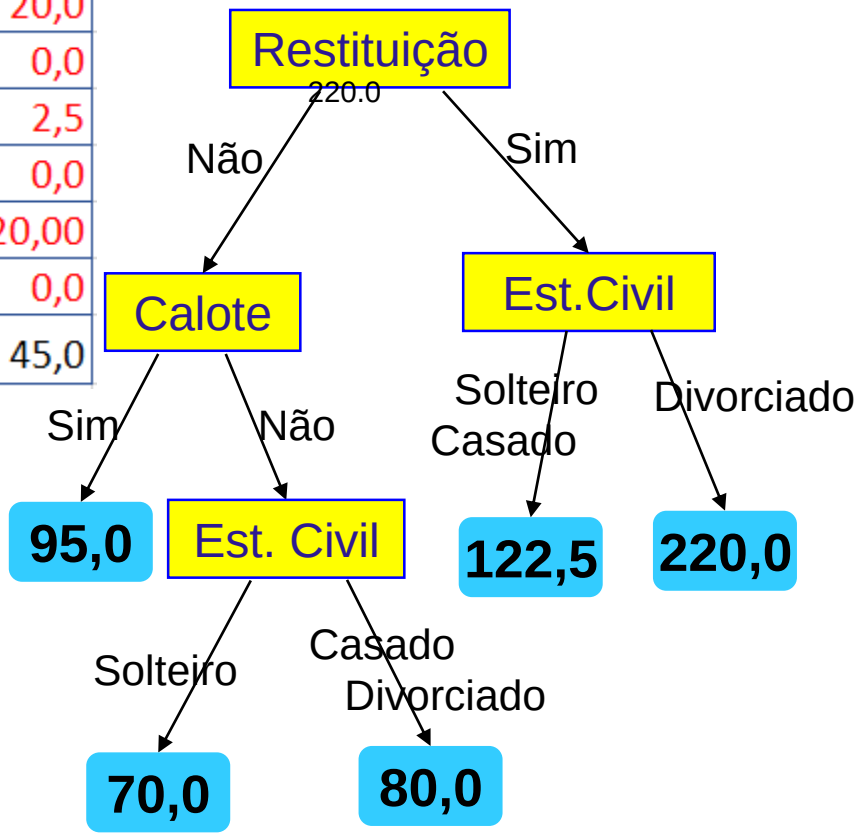
c) Calcule o Erro Médio Absoluto (EMA).

Restituição	Etado Civil	Calote	Rend. Anual	Rend. Anual Predito	Erro
SIM	Solteiro	NÃO	125,0	122,5	2,5
NÃO	Casado	NÃO	100,0	80,0	20,0
NÃO	Solteiro	NÃO	70,0	70,0	0,0
SIM	Casado	NÃO	120,0	122,5	2,5
NÃO	Divorciado	SIM	95,0	95,0	0,0
NÃO	Casado	NÃO	60,0	80,0	20,00
SIM	Divorciado	NÃO	220,0	220,0	0,0
				Σ =	45,0

Atenção! Erro é sempre Positivo!!

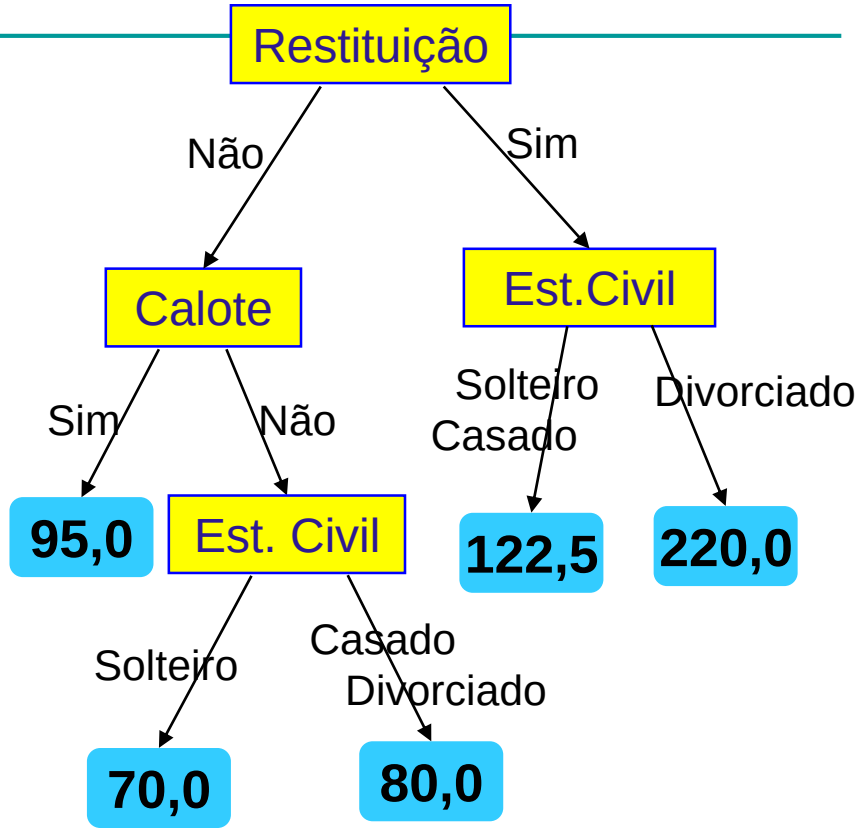
EMA = 45/7 = **6,43**

$$EMA = \sum_{i=1}^{N} (|previsto - real|)/N$$



Regressão

Criando uma árvore de regressão



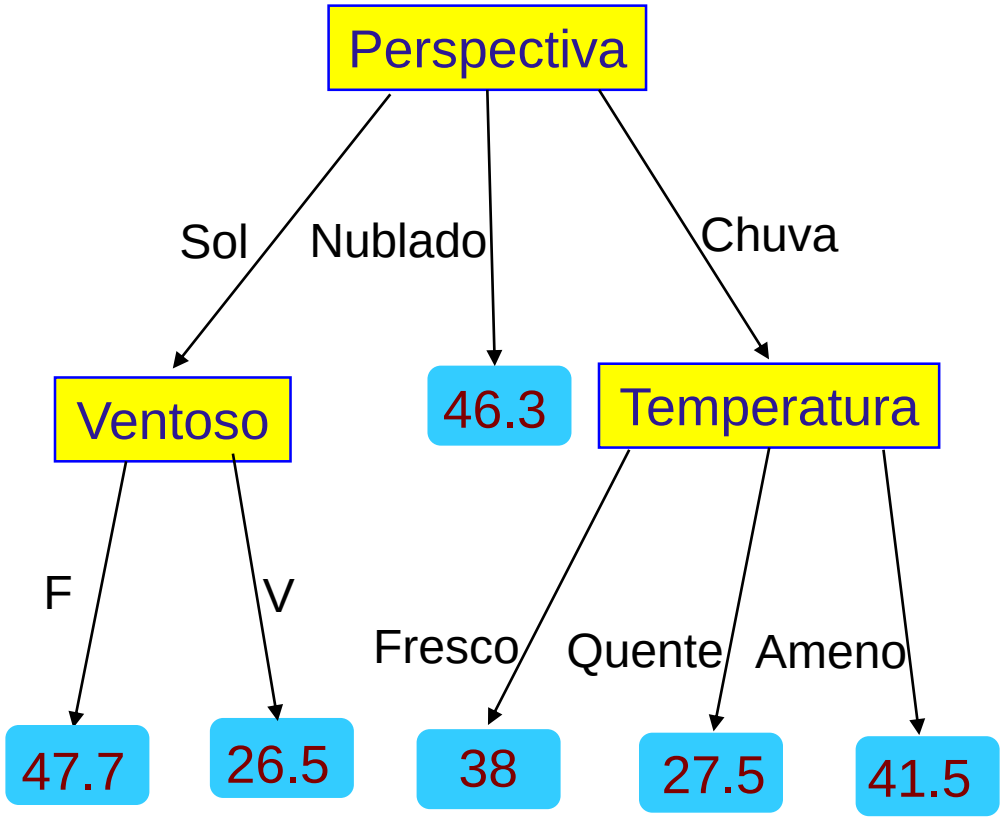

```
if sizeof (node.instances) < 5 or sd(node.instances)  < 0.05*SD
     node.type = LEAF
else
     node.type = INTERIOR
     for each atribute
          for all possible split positions of the atribute
               calculate the atribute's SDR
          node.atribute = atribute with maximum SDR
          split (node.left)
          split (node.right)
```

Criando uma Árvore de Regressão

Perspectiva (Categórico)	Temperatura (Categórico)	Humidade (Categórico)	Vento (Categórico)	Horas Jogadas (Alvo)
Chuva	Quente	Alta	F	25
Chuva	Quente	Alta	V	30
Nublado	Quente	Alta	F	46
Sol	Amena	Alta	F	45
Sol	Fresca	Normal	F	52
Sol	Fresca	Normal	V	23
Nublado	Fresca	Normal	V	43
Chuva	Amena	Alta	F	35
Chuva	Fresca	Normal	F	38
Sol	Amena	Normal	F	46
Chuva	Amena	Normal	V	48
Nublado	Amena	Alta	V	52
Nublado	Quente	Normal	F	44
Sol	Amena	Alta	V	30

Conjunto de Treino: Tempo

Modelo de Árvore de Regressão



Criando uma Árvore de Regressão

Perspectiva	Temperatura	Humidade	Ventoso	Horas Jogadas
Chuva	Quente	Alta	F	25
Chuva	Quente	Alta	V	30
Nublado	Quente	Alta	F	46
Sol	Amena	Alta	F	45
Sol	Fresca	Normal	F	52
Sol	Fresca	Normal	V	23
Nublado	Fresca	Normal	V	43
Chuva	Amena	Alta	F	35
Chuva	Fresca	Normal	F	38
Sol	Amena	Normal	F	46
Chuva	Amena	Normal	V	48
Nublado	Amena	Alta	V	52
Nublado	Quente	Normal	F	44
Sol	Amena	Alta	V	30

Conjunto de Treino

- *Passo 1*: Calcular o desvio padrão do atributo alvo.

  σ **Horas Jogadas = 9,32**

- *Passo 2:* O conjunto é dividido nos diferentes atributos e são calculados:
  - O desvio padrão de cada ramo.
  - A redução do desvio padrão (***SDR – Standard Deviation Reduction***) do nó.

Calculando o SDR para Perspectiva

Perspectiva	Temperatura	Humidade	Ventoso	Horas Jogadas
Chuva	Quente	Alta	F	25
Chuva	Quente	Alta	V	30
Nublado	Quente	Alta	F	46
Sol	Amena	Alta	F	45
Sol	Fresca	Normal	F	52
Sol	Fresca	Normal	V	23
Nublado	Fresca	Normal	V	43
Chuva	Amena	Alta	F	35
Chuva	Fresca	Normal	F	38
Sol	Amena	Normal	F	46
Chuva	Amena	Normal	V	48
Nublado	Amena	Alta	V	52
Nublado	Quente	Normal	F	44
Sol	Amena	Alta	V	30

Conjunto de Treino

Perspectiva	Quant.	Horas Jogadas (σ)
Chuva	5	7,78
Sol	5	10,87
Nublado	4	3,49
SDR = 1,66		

$$SDR(T) = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

Calculando o SDR para Perspectiva

Perspectiva	Quant.	Horas Jogadas (σ)
Chuva	5	7,78
Sol	5	10,87
Nublado	4	3,49
SDR = 1,66		

$$SDR(T) = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

- $S(T,X) = \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$

S(Horas, Perspectiva) =
*P(Chuva)*sd(Chuva)+P(Sol)*sd(Sol)+P(Nublado)*sd(Nublado)*
*S(Horas,Perspectiva)=(5/14)*7,78+(5/14)*10,87+(4/14)*3,49 = 7,66*

- $SDR(T,X) = sd(T) - S(T,X)$

SDR(Horas,Perspectiva)=sd(Horas)-S(Horas,Perspectiva)
SDR(Horas,Perspectiva)=9,32-7,66 = 1,66

SDR para todos os Ramos

Perspectiva	Quant.	Horas Jogadas (σ)
Chuva	5	7,78
Sol	5	10,87
Nublado	4	3,49
SDR = 1,66		

Temperatura	Quant.	Horas Jogadas (σ)
Fresca	4	10,51
Amena	6	7,65
Quente	4	8,95
SDR = 0,17		

Humidade	Quant.	Horas Jogadas (σ)
Alta	7	9,36
Normal	7	8,73
SDR = 0,28		

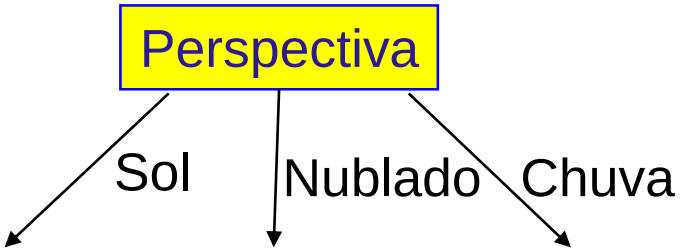
Ventoso	Quant.	Horas Jogadas (σ)
F	8	7,87
V	6	10,59
SDR = 0,29		

- *Passo 3*: O atributo que possui a MAIOR redução do desvio padrão é escolhido para o nodo de decisão.

Nodo Raíz

Perspectiva	Temperatura	Humidade	Ventoso	Horas Jogadas
Chuva	Quente	Alta	F	25
Chuva	Quente	Alta	V	30
Chuva	Amena	Alta	F	35
Chuva	Fresca	Normal	F	38
Chuva	Amena	Normal	V	48
Nublado	Quente	Alta	F	46
Nublado	Fresca	Normal	V	43
Nublado	Amena	Alta	V	52
Nublado	Quente	Normal	F	44
Sol	Amena	Alta	F	45
Sol	Fresca	Normal	F	52
Sol	Fresca	Normal	V	23
Sol	Amena	Normal	F	46
Sol	Amena	Alta	V	30

- *Passo 4*: O dataset é dividido baseado nos valores do atributo selecionado.



- *Passo 5*: O conjunto de ramo que possuir desvio padrão maior do que zero precisa continuar dividindo.

Recursividade para os demais Atributos

Temperatura	Humidade	Vento so	Horas Jogadas
Amena	Alta	F	45
Fresca	Normal	F	52
Fresca	Normal	V	23
Amena	Normal	F	46
Amena	Alta	V	30

- Volta para o Passo 2.
- $\sigma_{Sol} = 10{,}87$
- Calcula o SDR para cada atributo.

Humidade	Quant.	Horas Jogadas (σ)
Alta	2	7,50
Normal	3	12,50
SDR = 0,37		

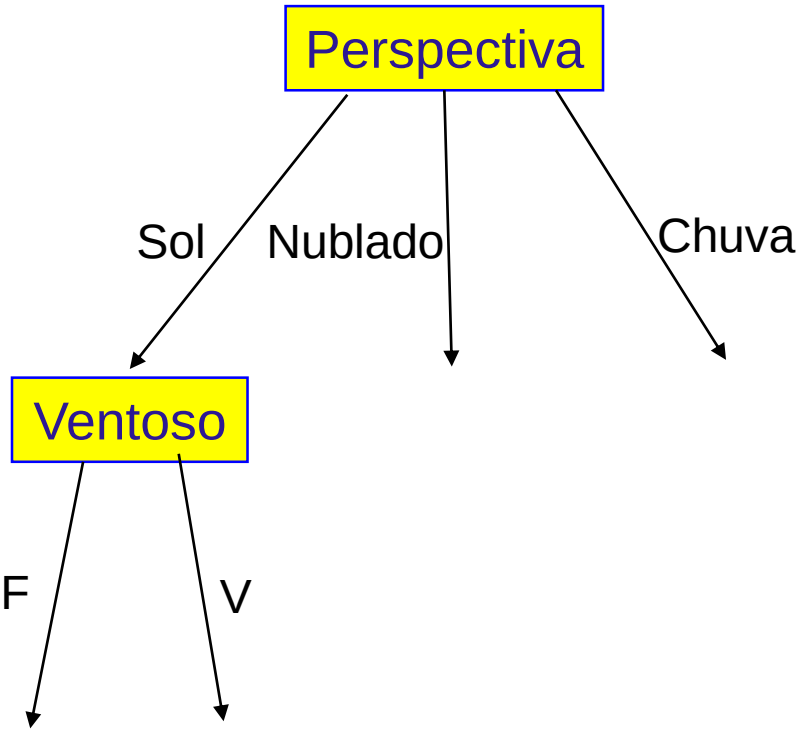
Ventoso	Quant.	Horas Jogadas (σ)
F	3	3,09
V	2	3,50
SDR = 7,62		

Temperatura	Quant.	Horas Jogadas (σ)
Fresca	2	14,5
Amena	3	7,32
Quente	0	0.00
SDR = 0,68		

Adicionando Ramos à Àrvore

- O atributo que possui a MAIOR redução do desvio padrão é escolhido compor o ramo da opção “Sol” do ramo Perspectiva.

Ventoso	Quant.	Horas Jogadas (σ)
F	3	3,09
V	2	3,50
SDR = 7,62		



Recursividade para os demais Atributos

Temperatura	Humidade	Ventoso	Horas Jogadas
Quente	Alta	F	25
Quente	Alta	V	30
Amena	Alta	F	35
Fresca	Normal	F	38
Amena	Normal	V	48

- Volta para o Passo 2.
- $\sigma_{Chuva} = 7{,}78$
- Calcula o SDR para cada atributo.

Humidade	Quant.	Horas Jogadas (σ)
Alta	3	4,08
Normal	2	5,00
SDR = 3,33		

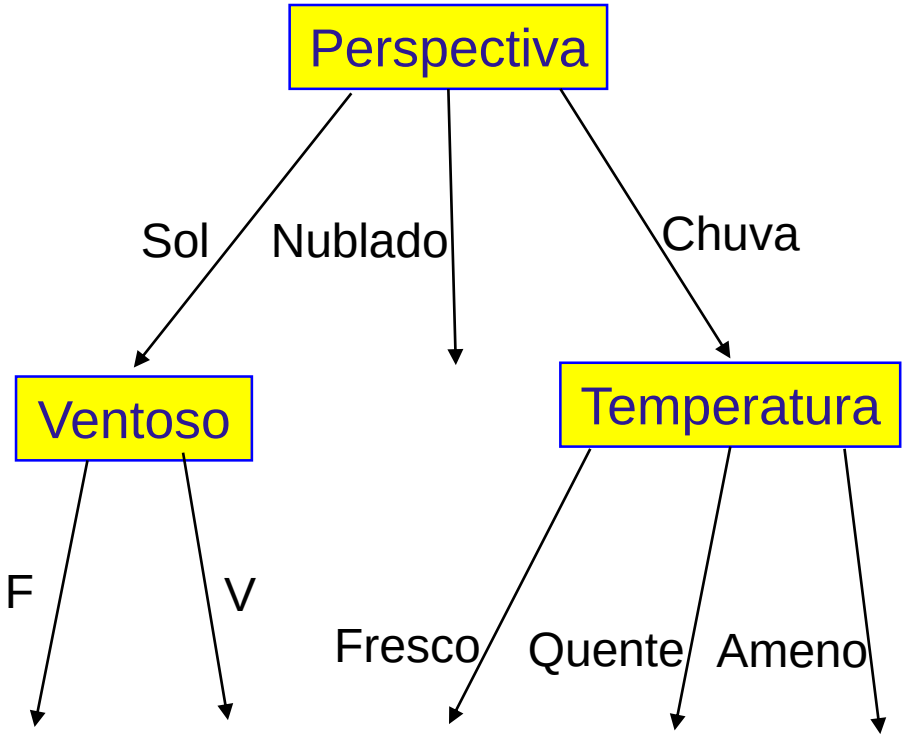
Ventoso	Quant.	Horas Jogadas (σ)
F	3	5,56
V	2	9,00
SDR = 0,85		

Temperatura	Quant.	Horas Jogadas (σ)
Fresca	1	0.00
Amena	2	6,50
Quente	2	2,50
SDR = 4,18		

Adicionando Ramos à Àrvore

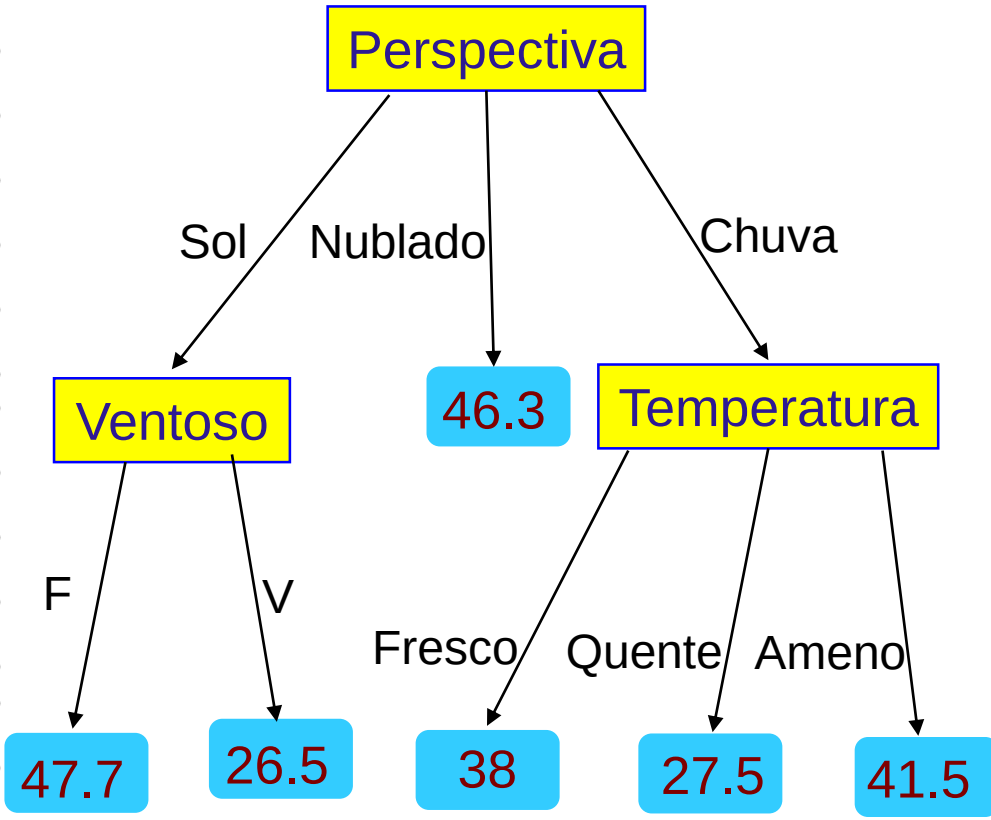
- O atributo que possui a MAIOR redução do desvio padrão é escolhido compor o ramo da opção “Chuva” do ramo Perspectiva.

Temperatura	Quant.	Horas Jogadas (σ)
Fresca	1	0.00
Amena	2	6,50
Quente	2	2,50
SDR = 4,18		



Calcula a média para nodos finais

Perspectiva	Temperatura	Humidade	Ventoso	Horas Jogadas
Chuva	Quente	Alta	F	25
Chuva	Quente	Alta	V	30
Chuva	Amena	Alta	F	35
Chuva	Fresca	Normal	F	38
Chuva	Amena	Normal	V	48
Nublado	Quente	Alta	F	46
Nublado	Fresca	Normal	V	43
Nublado	Amena	Alta	V	52
Nublado	Quente	Normal	F	44
Sol	Amena	Alta	F	45
Sol	Fresca	Normal	F	52
Sol	Fresca	Normal	V	23
Sol	Amena	Normal	F	46
Sol	Amena	Alta	V	30



Conjunto de Treino: Tempo

Critérios de Parada

- Árvores de regressão trabalham com dois critérios de parada, os quais em alguns algoritmos podem ser previamente definidos.
  1) Quando o desvio padrão de um ramo torna-se **menor do que uma certa fração** (padrão = 5%) do desvio padrão de todo o dataset.
  2) Se restarem **poucas instâncias** no ramo (por exemplo, 4). **Número mínimo de instâncias.**

- Quando a árvore de regressão estiver pronta, calcula-se a **média** como valores finais para o atributo alvo.

Exercícios

Calcular o EMA da árvore gerada a partir do conjunto de treino Tempo.

Perspectiva	Temperatura	Humidade	Vento so	Horas Jogadas	Predito	Erro
Chuva	Quente	Alta	F	25		
Chuva	Quente	Alta	V	30		
Nublado	Quente	Alta	F	46		
Sol	Amena	Alta	F	45		
Sol	Fresca	Normal	F	52		
Sol	Fresca	Normal	V	23		
Nublado	Fresca	Normal	V	43		
Chuva	Amena	Alta	F	35		
Chuva	Fresca	Normal	F	38		
Sol	Amena	Normal	F	46		
Chuva	Amena	Normal	V	48		
Nublado	Amena	Alta	V	52		
Nublado	Quente	Normal	F	44		
Sol	Amena	Alta	V	30		

EMA = _____

Prática:
Algoritmo DecisionTreeRegressor e RandomForestRegressor

Atividade para entregar

- A partir do dataset escolhido para trabalhar com os métodos supervisionados e utilizando a biblioteca scikitlearn da linguagem de programação Python, realize as seguintes tarefas:

1. Escolha apenas atributos contínuos e escolha um deles como sendo o classificador
2. Execute um dos algoritmos para Regressão do sklearn aplicando *utilizando a função GridSearchCV* do sklearn.

   1.1 Altere os parâmetros do algoritmo, tais como quantidade de arvores geradas na floresta (n_estimator), número máximo de features por árvore (max_features), entre outros

   1.2 Encontre os melhores parâmetros (hiperparâmetros) e melhor pontuação.
3. Execute novamente o algoritmo escolhido com o método holdout utilizando entre 25 a 30% dos dados para teste e os melhores parâmetros encontrados pelo *GridSearchCV.*

   2.1 Analise e compare os resultados obtidos, utilizando o erro médio absoluto (MAE).

   2.2 O objetivo é encontrar o melhor modelo para o dataset, baseando-se nas medidas de avaliação dos métodos supervisionados.

Referências

- Breiman, L., Freidman, J., Olshen, R. e Stone, C. (1984). *Classification and Regression Trees.* Wadsworth International Group., USA.
- Faceli, K.; Lorena, A.C.; Gama, J.; de Carvalho, A.C.P.L.F. *Inteligência Artificial: Uma abordagem de aprendizado de máquina.* LTC, Rio de Janeiro, 2011.
- Quilan, R. (1979). *Discovering rules by induction from large collections of examples.* In: Michie, D. (Ed.) Expert Systems in the Microelectronic Age, p. 168-201. Edinburgh University Press.
- Quilan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA.
- TAN, P-N; STEINBACH, M.; KUMAR,V. *Introduction to Data Mining.* Pearson, 2006.