

Data Mining

Regras de Classificação

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2021

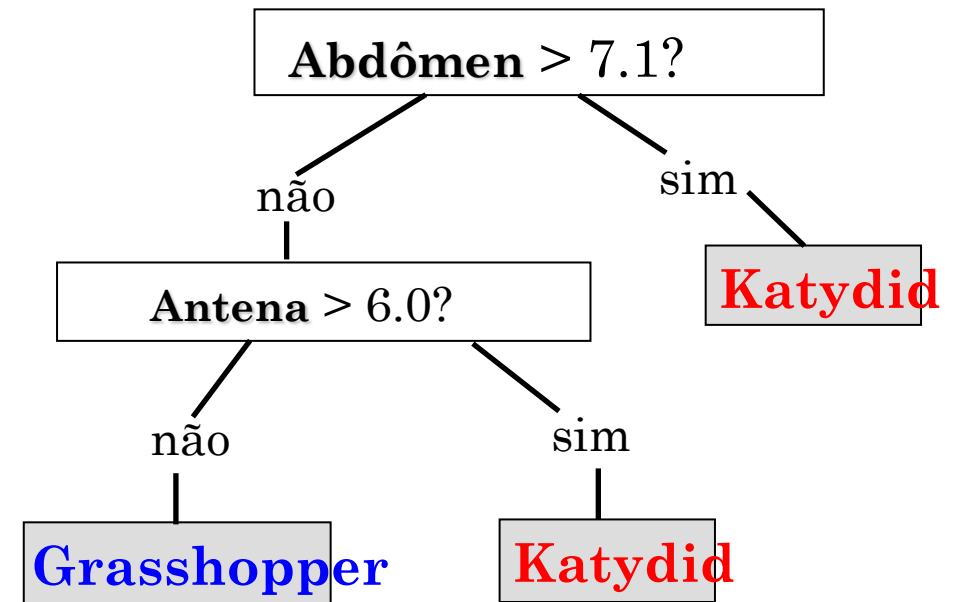
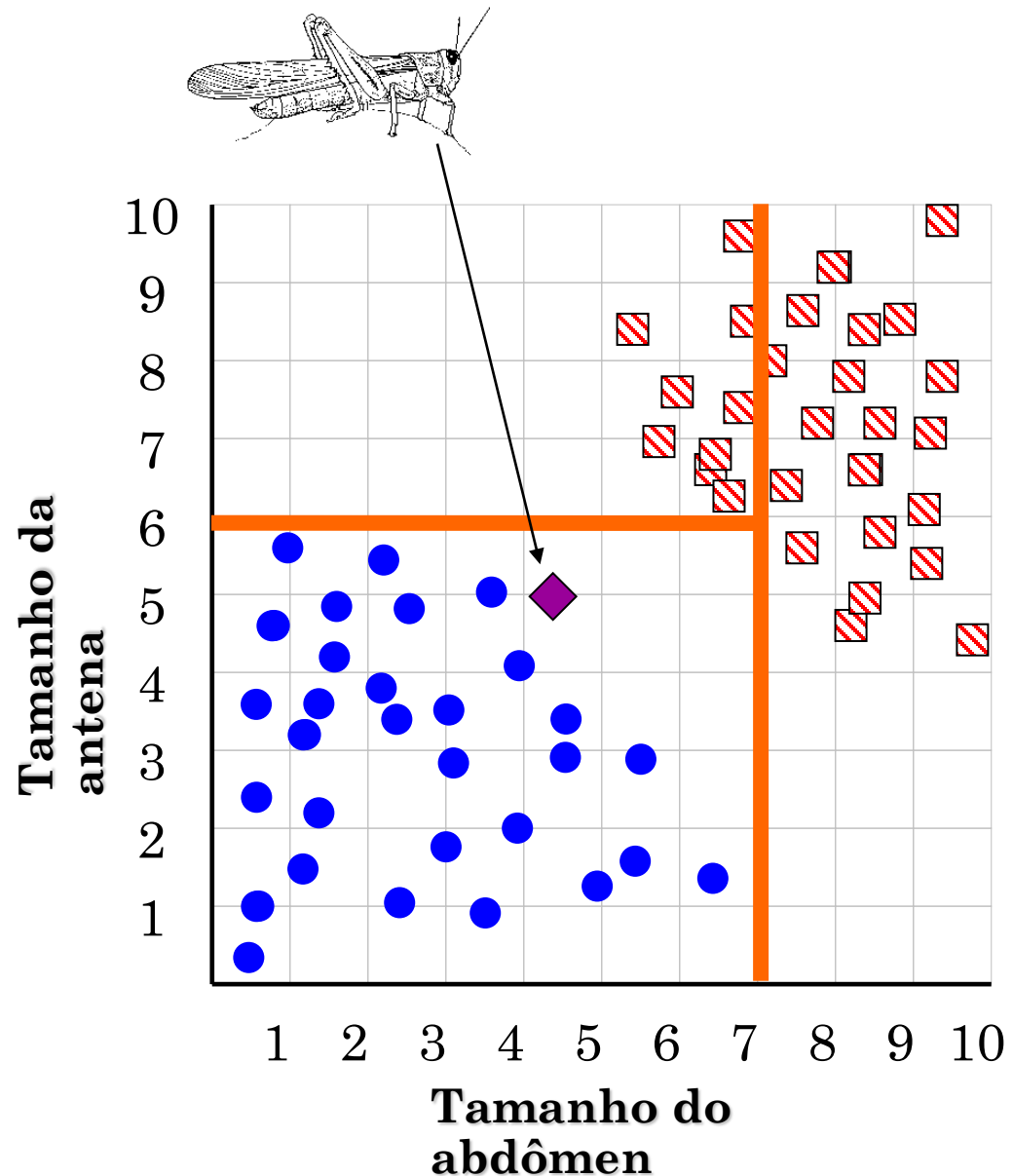
Notas legais

- Este material foi cedido pelo Dr. Eamonn Keogh (University of California - Riverside, US) para as aulas de mineração de dados na UFSM.
- Tradução e adaptação: Dr. Joaquim Assunção.

Decision Tree Classifier

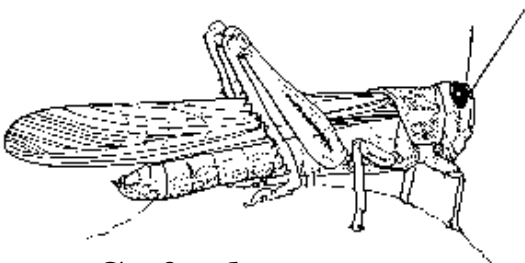


Ross Quinlan



Antena é mais curta que o corpo?

Yes



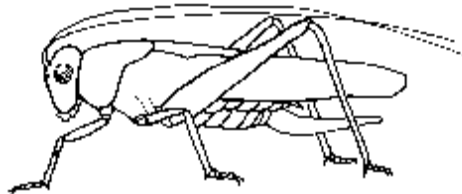
Gafanhoto

No

3 Tarsi?



Yes



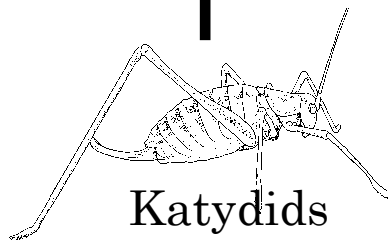
Grilo



No

Foretíba tem orelhas?

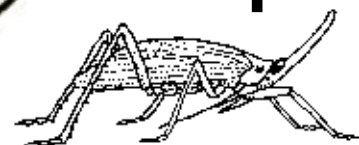
Yes



Katydid



No



Grilo Camelo

Uma árvore de decisão ...

- É uma estrutura de árvore do tipo fluxograma.
- Nó interno indica um teste em um atributo.
- Uma ramificação representa um resultado do teste.
- Nós de folha representam rótulos de classe ou distribuição de classe

A geração da árvore de decisão consiste em duas fases

- Construção de árvores
 - No início, todos os exemplos de treinamento estão na raiz.
 - Exemplos de partição são recursivamente criados com base nos atributos selecionados.
- Poda de árvores
 - Identifique e remova ramificações que refletem ruído ou *outliers*.

Como construir uma árvore



Condições de parada

- Todas as amostras de um determinado nó pertencem à mesma classe.
- Não há atributos remanescentes para particionamento adicional - votação majoritária é empregada para classificar a folha.
- Não há amostras restantes.

Como construir uma árvore



Ganho de informação como critério de divisão

- Selecione o atributo com o maior ganho de informação (ganho de informação é a redução esperada na entropia).
- Suponha que existem duas classes, P e N
 - Deixe o conjunto de exemplos S conter p elementos da classe P e n elementos da classe N
 - A quantidade de informação necessária para decidir se um exemplo arbitrário em S pertence a P ou N é definido como:

$$E(S) = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

Como construir uma árvore



Ganho de informação como critério de divisão

- Selecione o atributo com o maior ganho de informação (ganho de informação é a redução esperada na entropia).
- Suponha que existem duas classes, P e N
 - Deixe o conjunto de exemplos S conter p elementos da classe P e n elementos da classe N
 - A quantidade de informação necessária para decidir se um exemplo arbitrário em S pertence a P ou N é definido como:

$$E(S) = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

...










Suponha que, usando o atributo A , um conjunto atual seja particionado em algum número de conjuntos filho

A informação de codificação que seria obtida pela ramificação em $A \rightarrow$

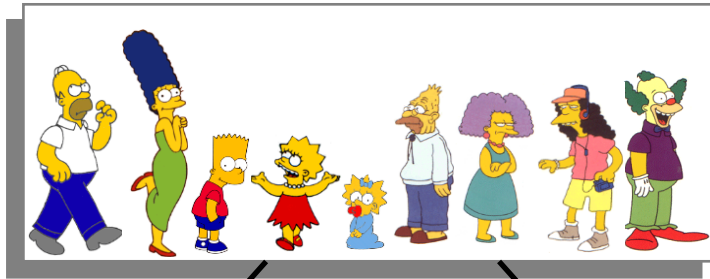
$$Gain(A) = E(Current\ set) - \sum E(all\ child\ sets)$$

Nota: entropia está no mínimo se a coleção de objetos for uniforme.

Exemplo

Person	Hair Length	Weight	Age	Class
 Homer	0"	250	36	M
 Marge	10"	150	34	F
 Bart	2"	90	10	M
 Lisa	6"	78	8	F
 Maggie	4"	20	1	F
 Abe	1"	170	70	M
 Selma	8"	160	41	F
 Otto	10"	180	38	M
 Krusty	6"	200	45	M

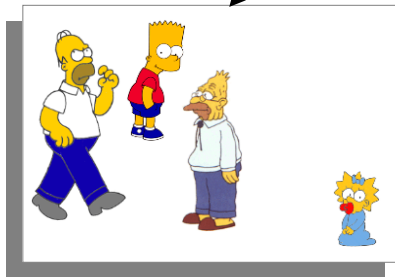
	Comic	8"	290	38	?
---	-------	----	-----	----	----------



$$Entropy(S) = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

$$Entropy(4\mathbf{F}, 5\mathbf{M}) = -(4/9) \log_2(4/9) - (5/9) \log_2(5/9) = \mathbf{0.9911}$$

yes
no
Hair Length <= 5?



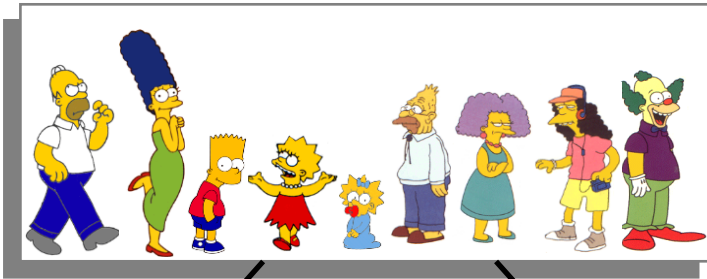
Vamos tentar dividir
em *Hair length*

$$Entropy(1\mathbf{F}, 3\mathbf{M}) = -(1/4) \log_2(1/4) - (3/4) \log_2(3/4) = \mathbf{0.8113}$$

$$Entropy(3\mathbf{F}, 2\mathbf{M}) = -(3/5) \log_2(3/5) - (2/5) \log_2(2/5) = \mathbf{0.9710}$$

$$Gain(A) = E(Current\ set) - \sum E(all\ child\ sets)$$

$$Gain(Hair\ Length\ \leq 5) = \mathbf{0.9911} - (4/9 * \mathbf{0.8113} + 5/9 * \mathbf{0.9710}) = \mathbf{0.0911}$$



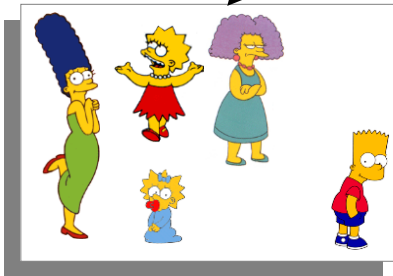
$$Entropy(S) = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

$$Entropy(4\mathbf{F}, 5\mathbf{M}) = -(4/9) \log_2(4/9) - (5/9) \log_2(5/9) = \mathbf{0.9911}$$

yes

no

Weight <= 160?



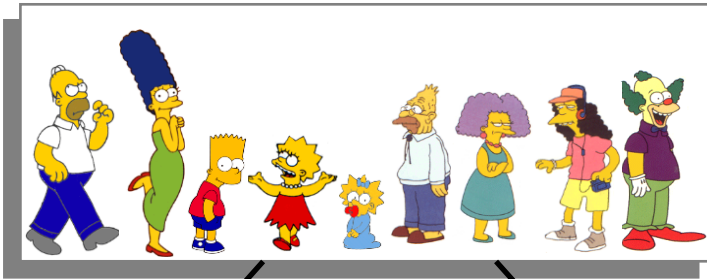
Vamos tentar dividir em *Weight*

$$Entropy(4\mathbf{F}, 1\mathbf{M}) = -(4/5) \log_2(4/5) - (1/5) \log_2(1/5) = \mathbf{0.7219}$$

$$Entropy(0\mathbf{F}, 4\mathbf{M}) = -(0/4) \log_2(0/4) - (4/4) \log_2(4/4) = \mathbf{0}$$

$$Gain(A) = E(Current\ set) - \sum E(all\ child\ sets)$$

$$Gain(Weight \leq 160) = \mathbf{0.9911} - (5/9 * \mathbf{0.7219} + 4/9 * \mathbf{0}) = \mathbf{0.5900}$$



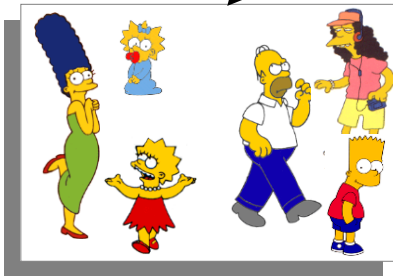
$$Entropy(S) = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

$$Entropy(4\text{F}, 5\text{M}) = -(4/9) \log_2(4/9) - (5/9) \log_2(5/9) = 0.9911$$

yes

no

age <= 40?



Vamos tentar dividir em Age

$$Entropy(3\text{F}, 3\text{M}) = -(3/6) \log_2(3/6) - (3/6) \log_2(3/6) = 1$$

$$Entropy(1\text{F}, 2\text{M}) = -(1/3) \log_2(1/3) - (2/3) \log_2(2/3) = 0.9183$$

$$Gain(A) = E(\text{Current set}) - \sum E(\text{all child sets})$$

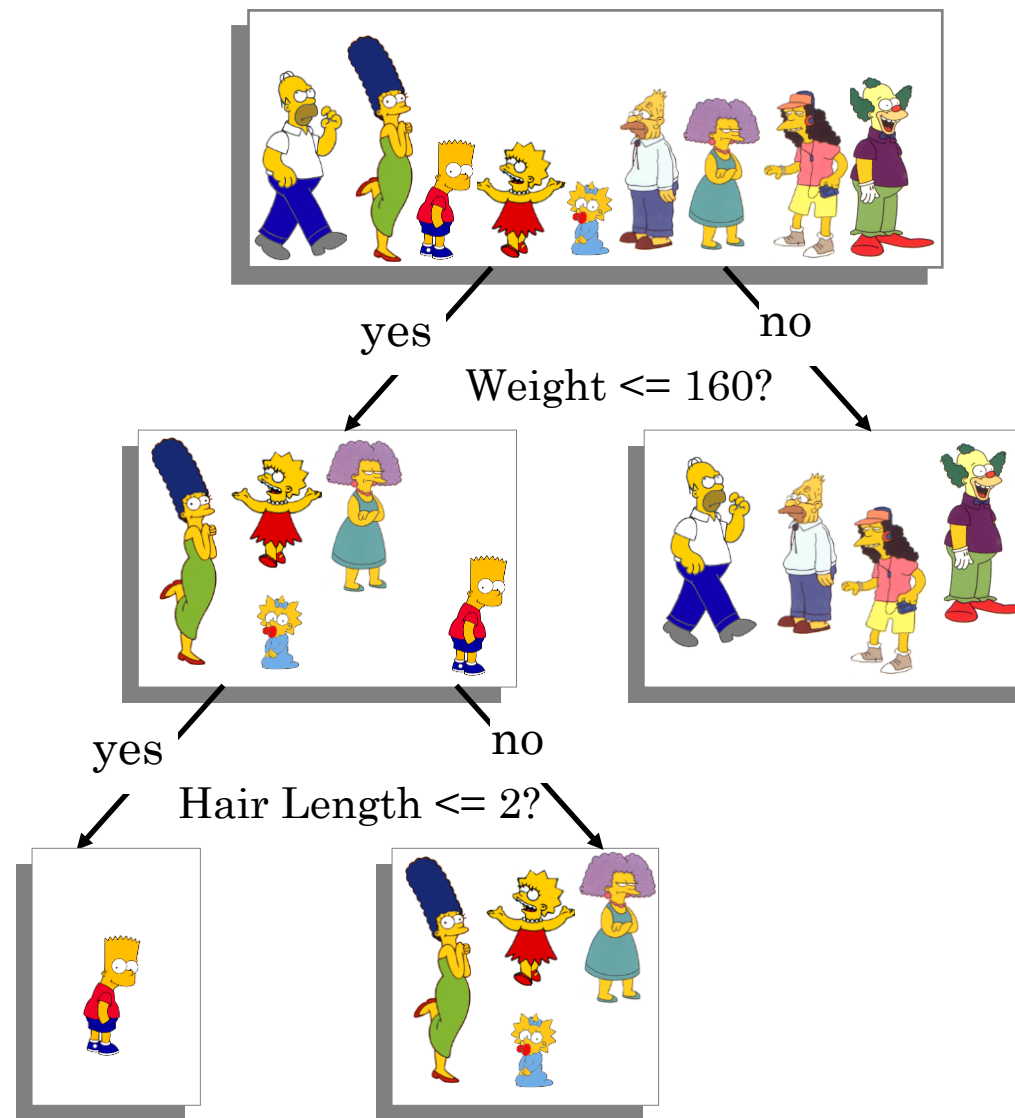
$$Gain(\text{Age} \leq 40) = 0.9911 - (6/9 * 1 + 3/9 * 0.9183) = 0.0183$$

!?!
...

Das três características que tivemos, o peso foi o melhor.

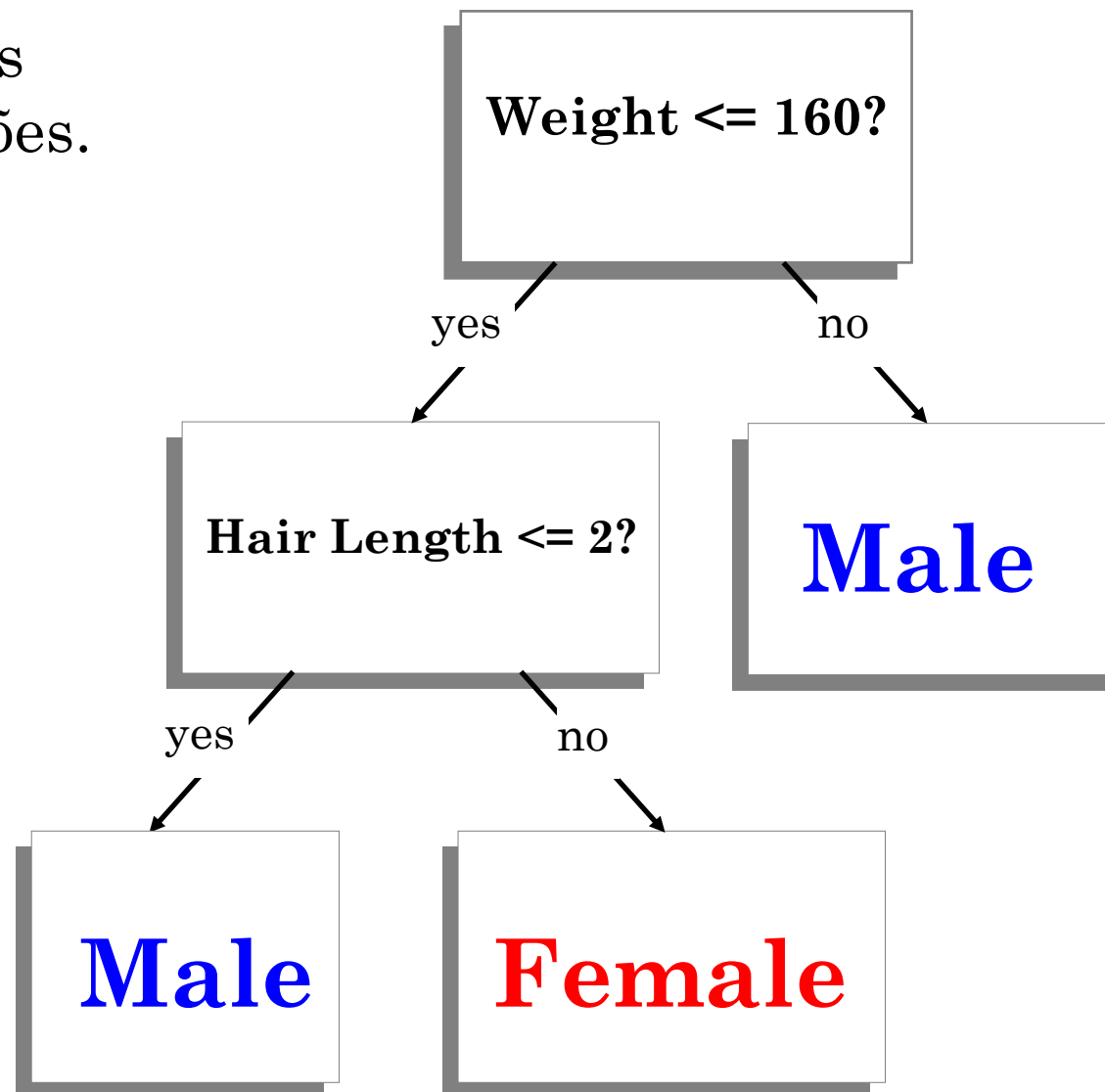
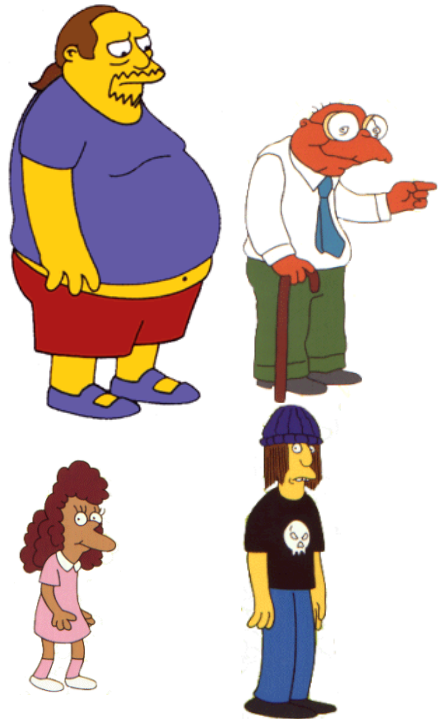
Mas enquanto as pessoas que pesam mais de 160 libras são perfeitamente classificadas (como homens), as pessoas com menos de 160 libras não são perfeitamente classificadas ... Então, usamos recursão!

- Veja que o comprimento do cabelo resolveu o problema.



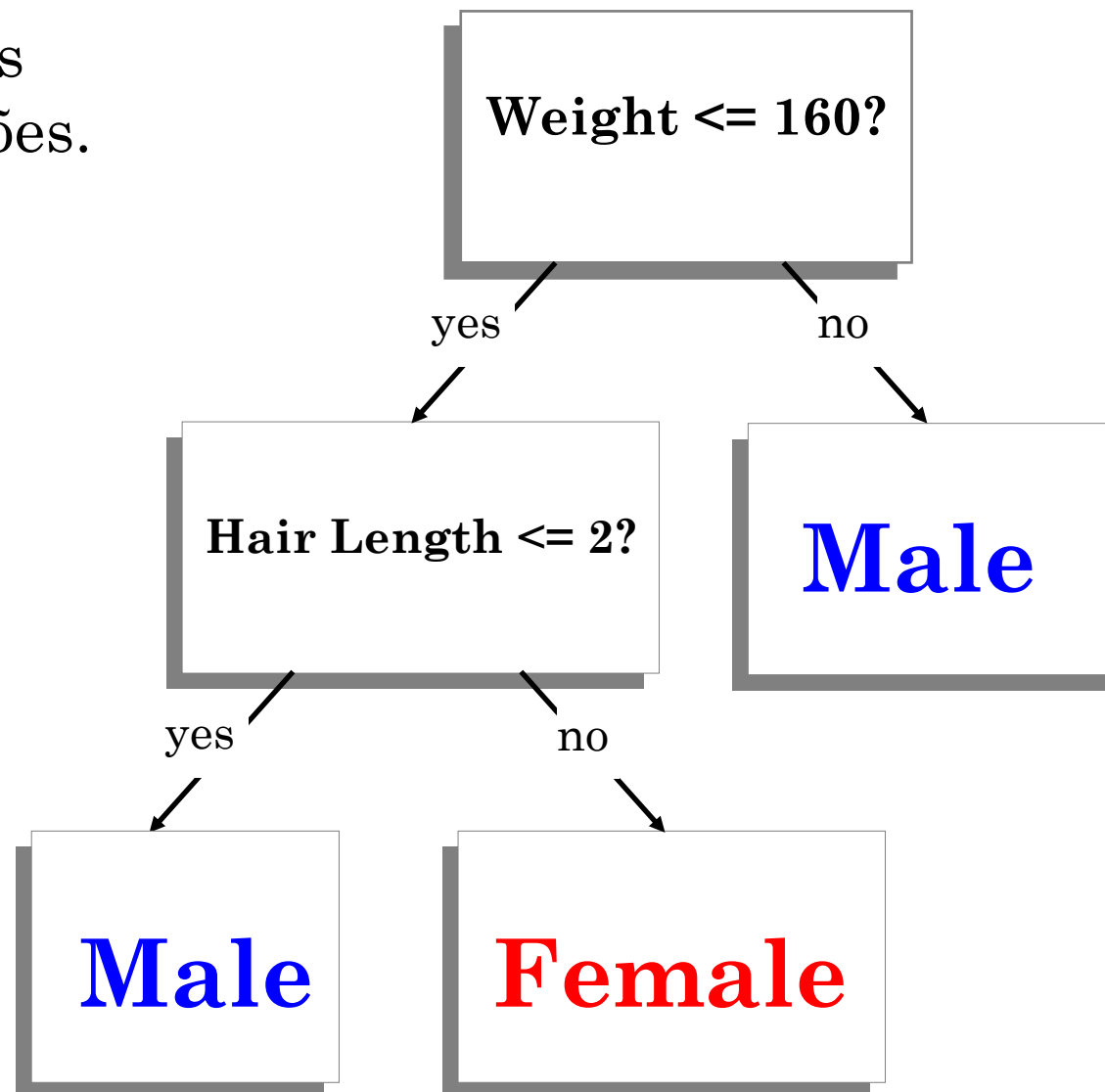
Não precisamos manter os dados, somente as condições.

Como estas pessoas seriam classificadas?



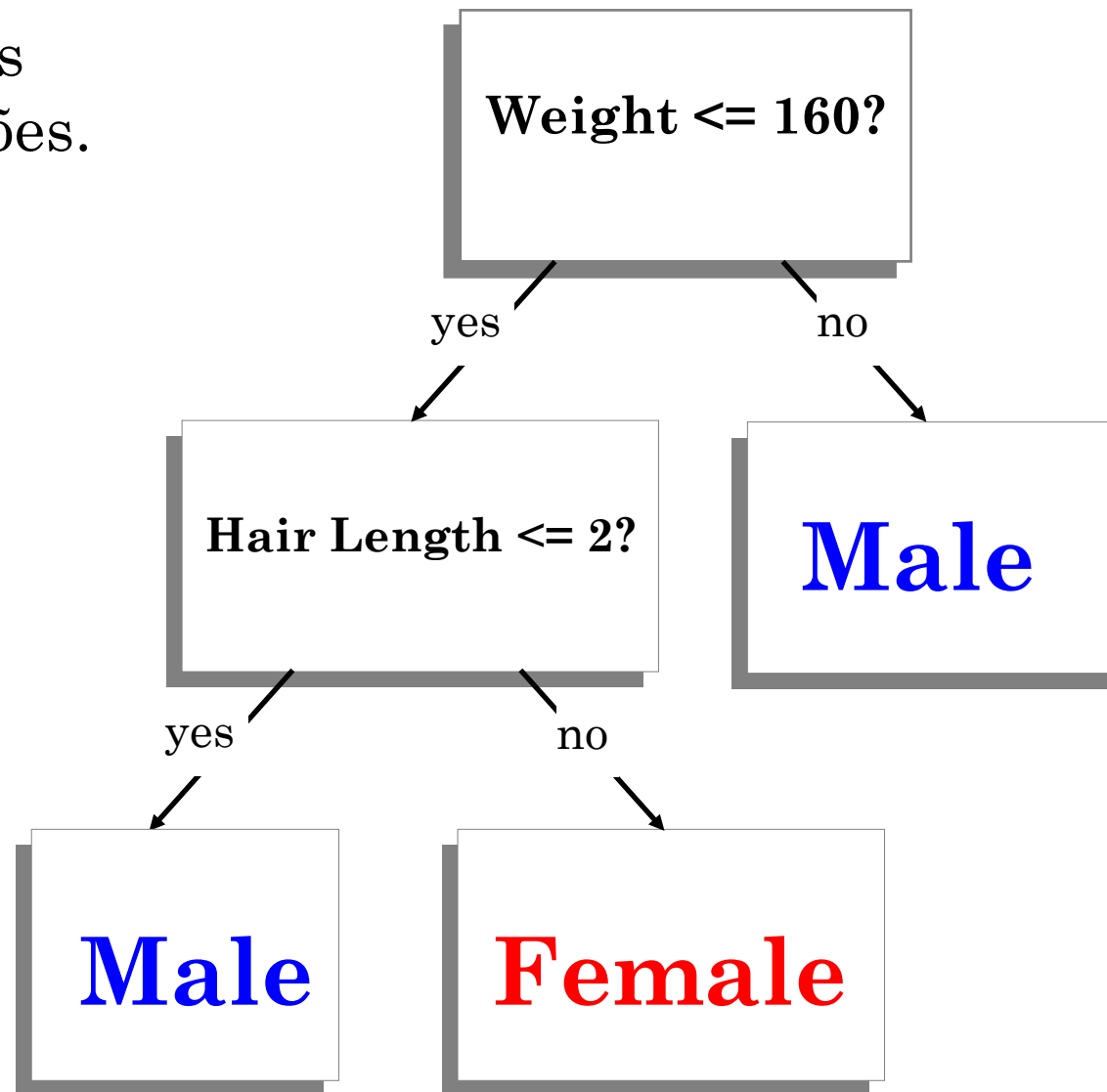
Não precisamos manter os dados, somente as condições.

Como estas pessoas seriam classificadas?



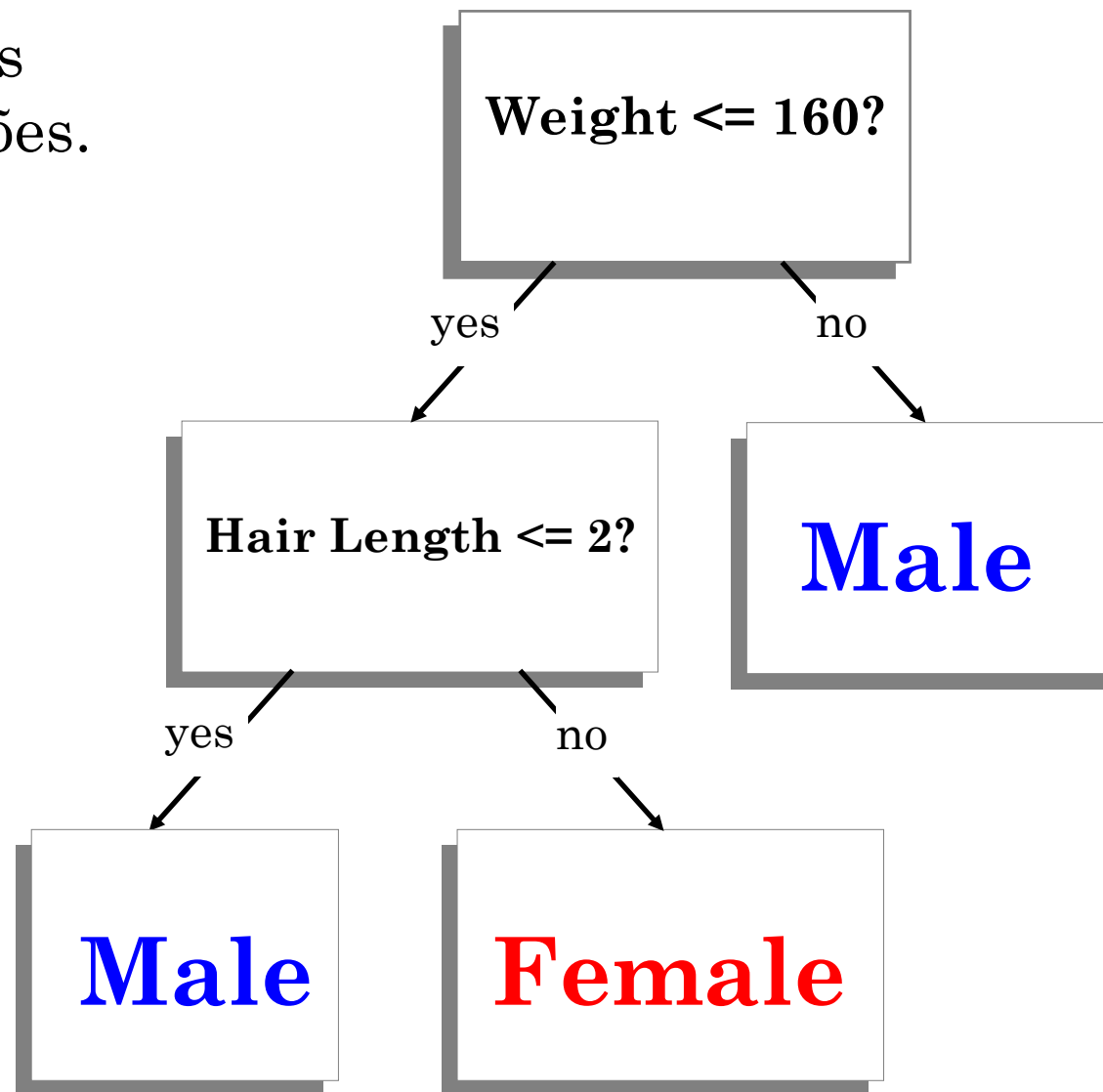
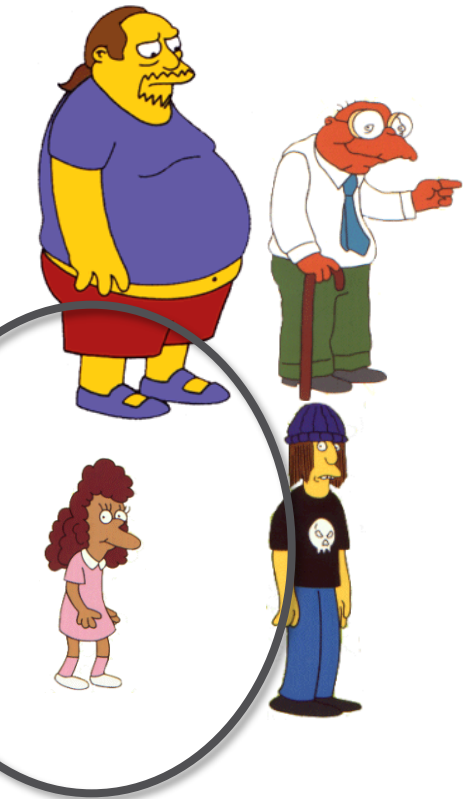
Não precisamos manter os dados, somente as condições.

Como estas pessoas seriam classificadas?



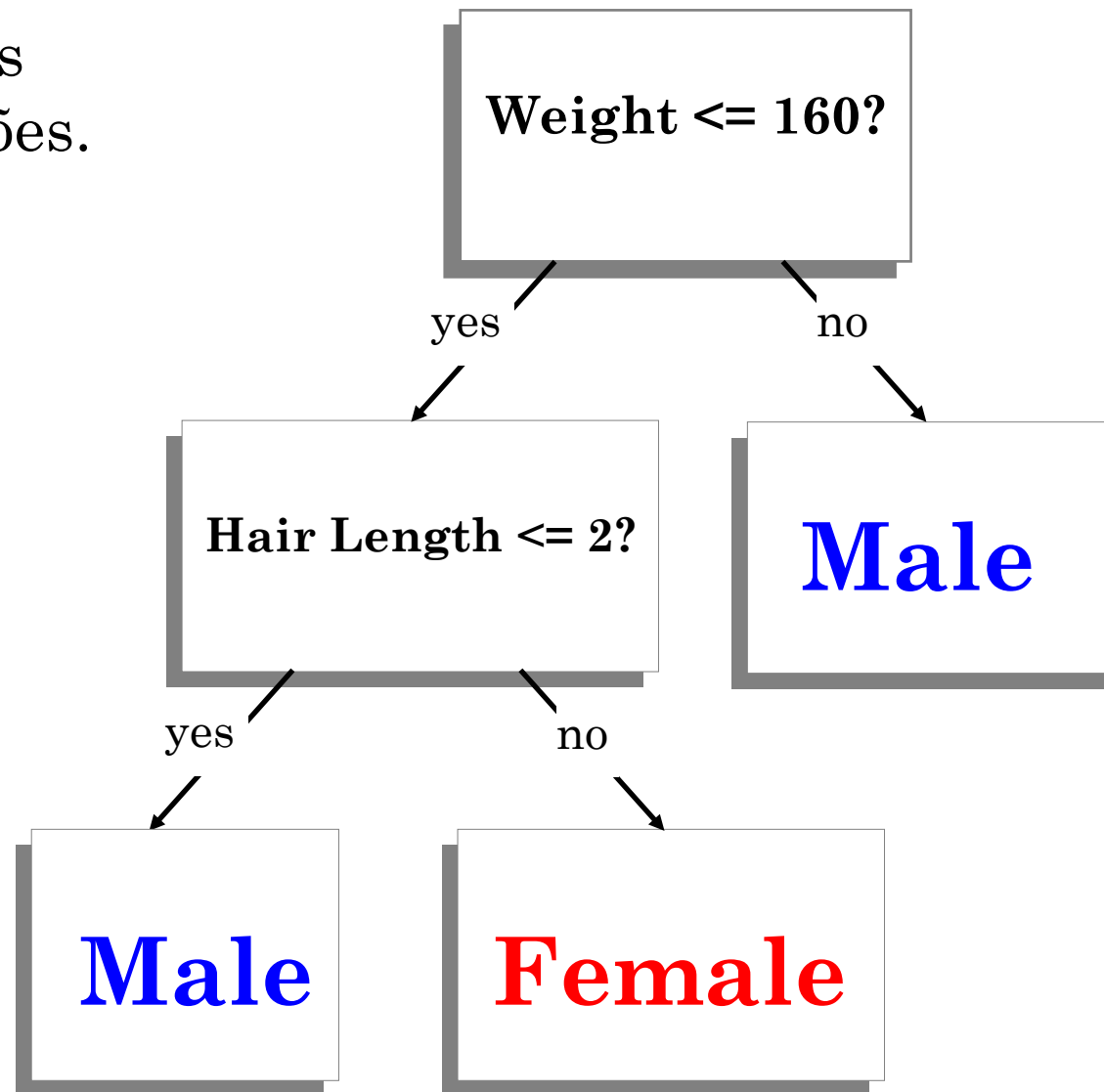
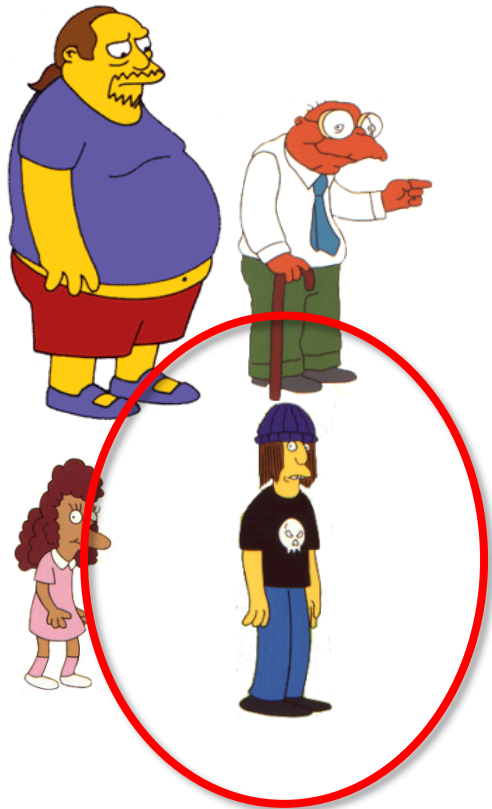
Não precisamos manter os dados, somente as condições.

Como estas pessoas seriam classificadas?



Não precisamos manter os dados, somente as condições.

Como estas pessoas seriam classificadas?



Hands On!

1. Converta a seguinte árvore em um algoritmo para testar se o indivíduo X é homem ou mulher.
2. Calcule a entropia gerada por 35 instancias, onde 21 foram classificadas de um lado (19A e 2B) e 14 de outro (3A 11B).

