COMP0271: Inteligência Artificial
Árvores de decisão

Professor: Hendrik Macedo

Universidade Federal de Sergipe, Brasil

Subconjunto do censo 1990 US: 48.000 registros, 16 atributos [Kohavi 1995]

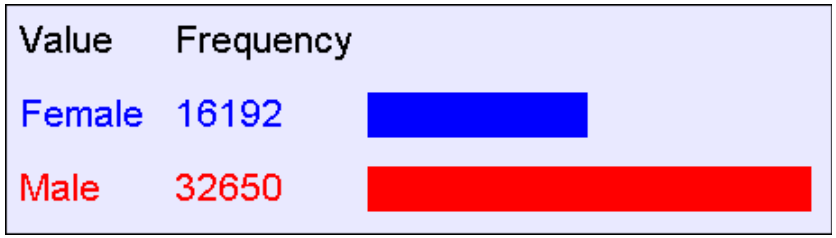
age	employment	education	edunum	marital	…	job	relation	race	gender	hours	country	wealth
					…							
39	State_gov	Bachelors	13	Never_mar	…	Adm_cleric	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_not_i	Bachelors	13	Married	…	Exec_mana	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	…	Handlers_c	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	…	Handlers_c	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	…	Prof_specia	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	…	Exec_mana	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	…	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp_not_i	HS_grad	9	Married	…	Exec_mana	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	…	Prof_specia	Not_in_fam	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	…	Exec_mana	Husband	White	Male	40	United_Sta	rich
37	Private	Some_colle	10	Married	…	Exec_mana	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	…	Prof_specia	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	…	Adm_cleric	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_acdı	12	Never_mar	…	Sales	Not_in_fam	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	…	Craft_repai	Husband	Asian	Male	40	*MissingVa	rich
34	Private	7th_8th	4	Married	…	Transport_ı	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp_not_i	HS_grad	9	Never_mar	…	Farming_fis	Own_child	White	Male	35	United_Sta	poor
33	Private	HS_grad	9	Never_mar	…	Machine_o	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	…	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_not_i	Masters	14	Divorced	…	Exec_mana	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	…	Prof_specia	Husband	White	Male	60	United_Sta	rich
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:

Atributos

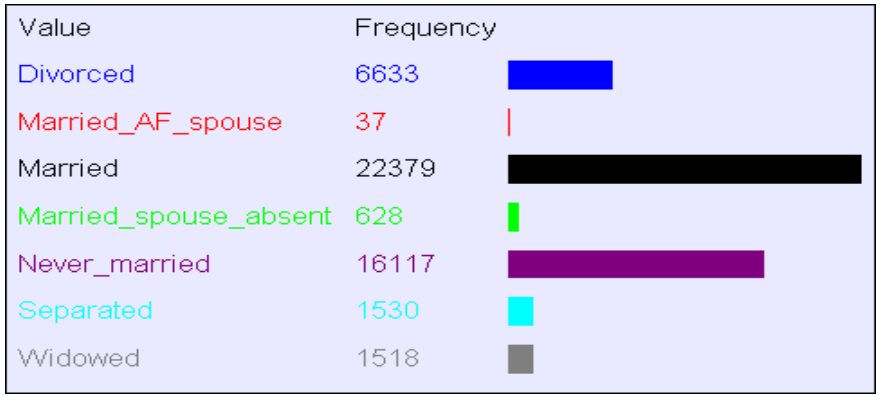
age	edunum	race	hours_worked
employment	marital	gender	country
taxweighting	job	capitalgain	wealth
education	relation	capitalloss	agegroup

O que podemos fazer com essas informações?

- Montar Histogramas



Gender



Marital Status

O que podemos fazer com essas informações?

- Montar Tabelas de Contingência (1-d, 2-d, 3-d)
  1. Escolha *d* atributos: $a_1, a_2, \ldots a_d$.
  2. Para cada combinação possível de valores: $a_1 = x_1$, $a_2 = x_2$, ... $a_d = x_d$, grave o quão frequentemente essa combinação ocorre

Tabela de Contingência 2-d

wealth values:		poor	rich
agegroup	10s	2507	3
	20s	11262	743
	30s	9468	3461
	40s	6738	3986
	50s	4110	2509
	60s	2245	809
	70s	668	147
	80s	115	16
	90s	42	13

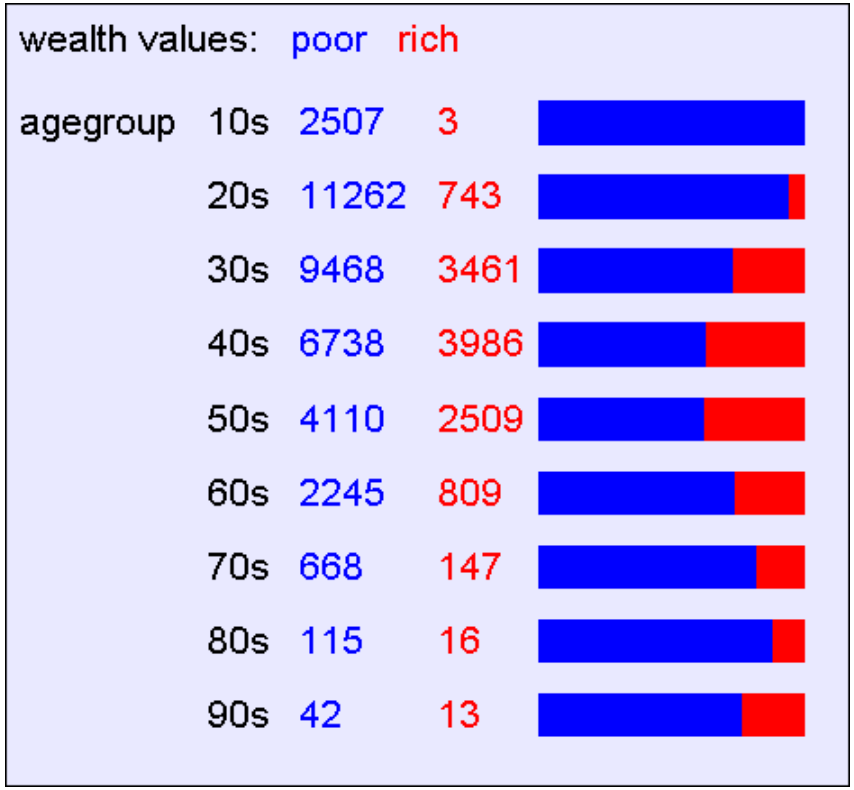


Tabela de Contingência 2-d

job values:		Adm_clerical	Craft_repair		Farming_fishing			Machine_op_inspct			Priv_house_serv			Protective_serv		Tech_support
MissingValue		Armed_Forces	Exec_managerial		Handlers_cleaners			Other_service			Prof_specialty			Sales		Transport_moving
marital	Divorced	270	1192	0	679	890	90	197	434	762	46	795	121	664	239	254	
	Married_AF_spouse	5	6	0	4	3	1	1	1	5	0	4	1	5	0	1	
	Married	928	1495	7	3818	3600	869	724	1469	1088	27	3182	583	2491	609	1489	
	Married_spouse_absent	45	84	0	77	52	35	32	37	92	9	64	7	55	9	30	
	Never_married	1242	2360	8	1301	1260	434	1029	872	2442	99	1849	237	1992	506	486	
	Separated	97	224	0	160	126	23	63	123	275	21	145	23	146	48	56	
	Widowed	222	250	0	73	155	38	26	86	259	40	133	11	151	35	39	

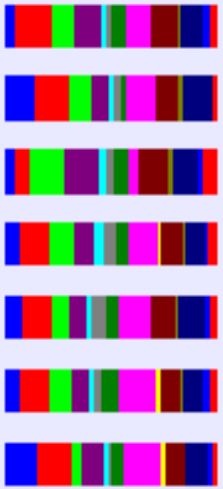
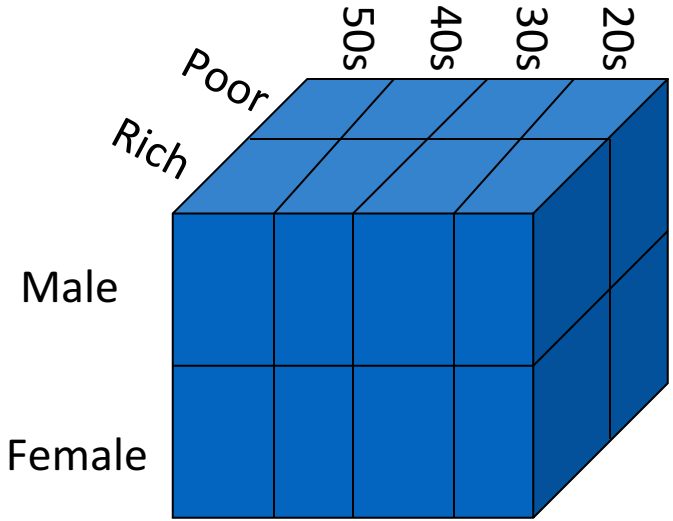


Tabela de Contingência 3-d

- Mais difícil de visualizar



Refletindo…

- Com 16 atributos,
  - quantas tabelas 1-d existem? 16
  - quantas 2-d? 16 * 15 / 2 = 120
  - 3-d? 16*15*14 / 6 = 560
- Com 100 atributos, quantas tabelas 3-d existem?
  - 100*99*98 / 6 = 161.700 !!!!!

Imagina ter que olhar/analisar isso manualmente ???

Mineração dos dados

- Automatização do processo de busca por padrões nos dados.

Que padrões são interessantes?

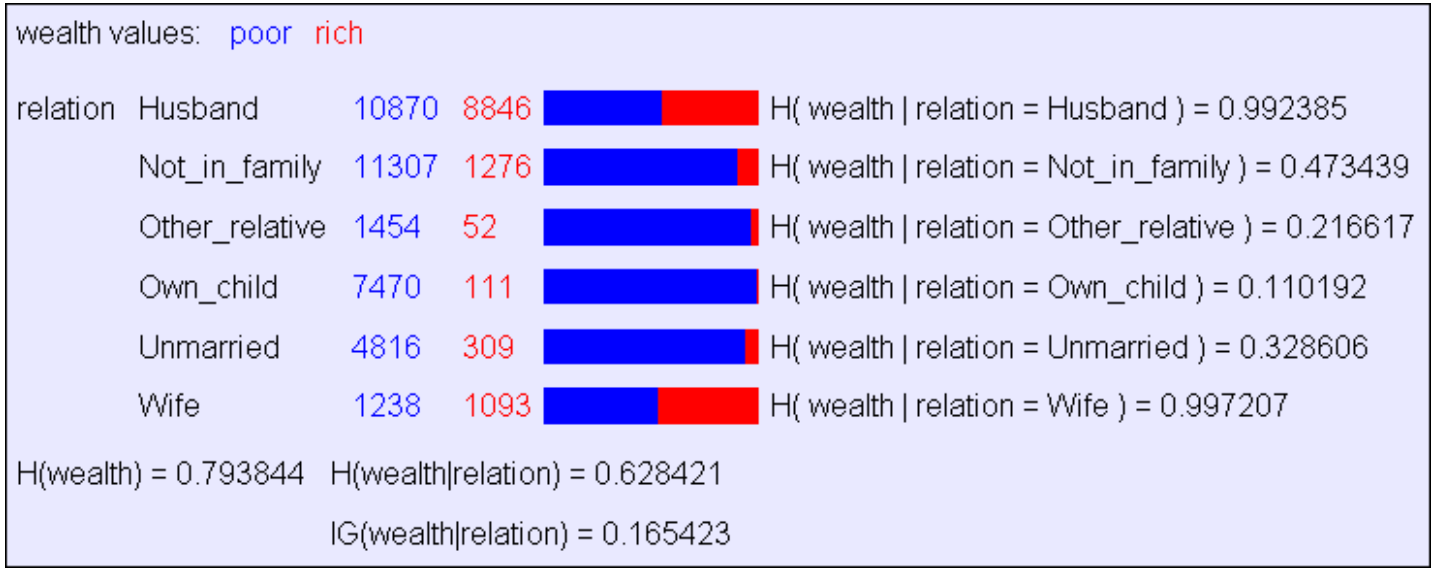
Quais seriam meras ilusões?

Como podem ser explorados?

usar Teoria da Informação (Shannon, 1948)

Pedindo ajuda à máquina…

- Considerando algo que você quer prever (ex: wealth), é mais fácil pedir ao computador para encontrar o atributo de maior **Ganho de Informação**



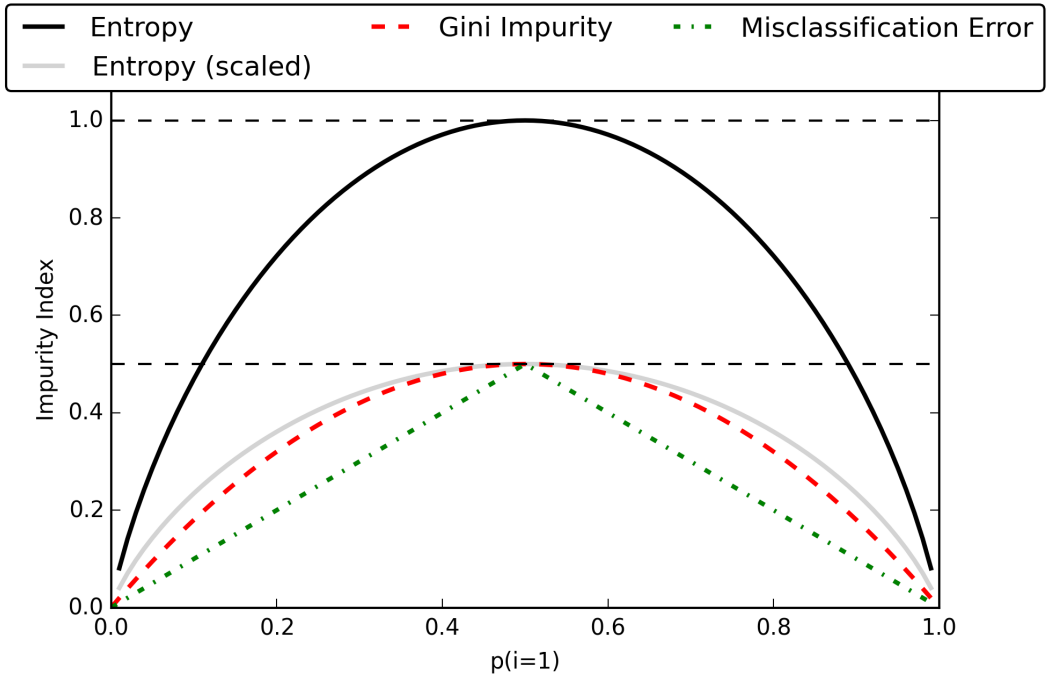
Ganho de informação e Impureza

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

(Impureza)

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) \log_2 p(i|t)$$

(Entropia)



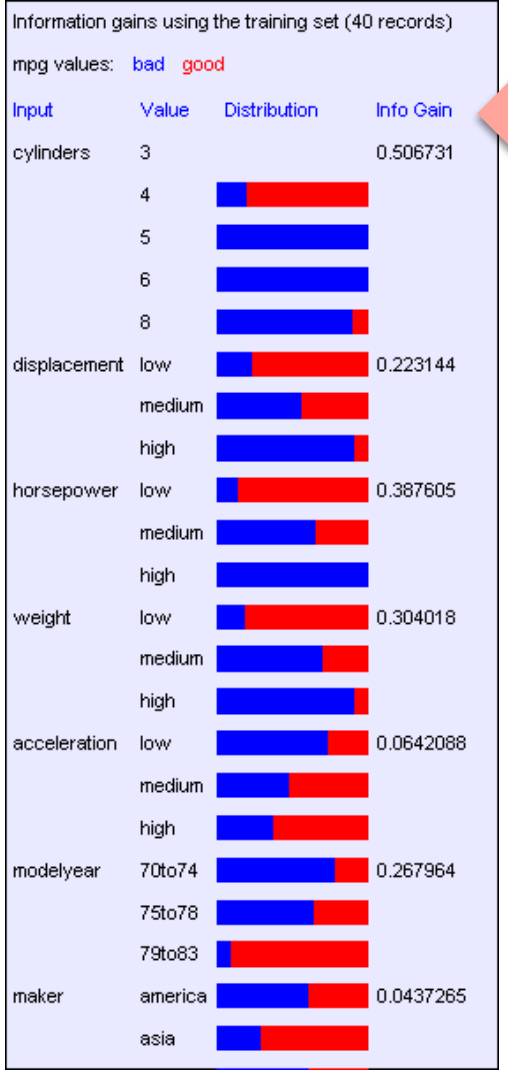
Árvore de Decisão

- Plano estruturado em árvore de um conjunto de atributos para testar a fim de se prever a saída.
- Qual atributo deve ser testado primeiro? Aquele que produzir **maior** Ganho de Informação
  - E aí, recorrentemente...

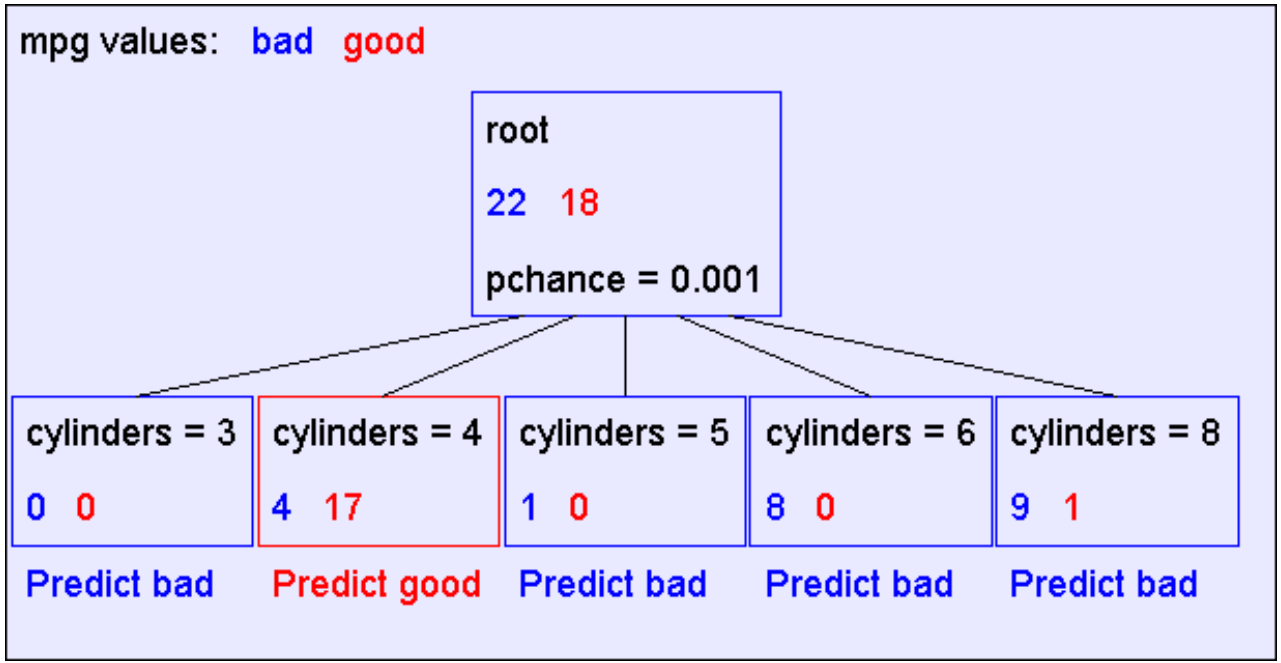
Outro dataset 40 exemplos

mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe

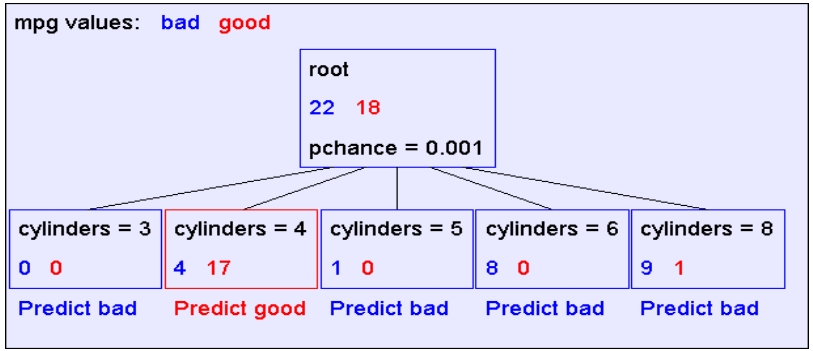
Considere prever MPG (milhas por galão)…



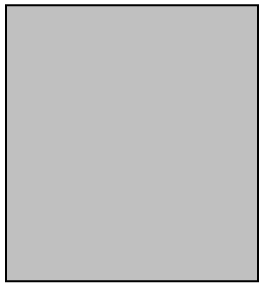
Um pedaço da árvore



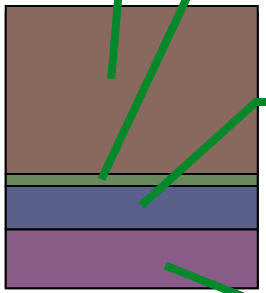
Passo recursivo



Pega o dataset original..



... e particiona de acordo com o valor do atributo.



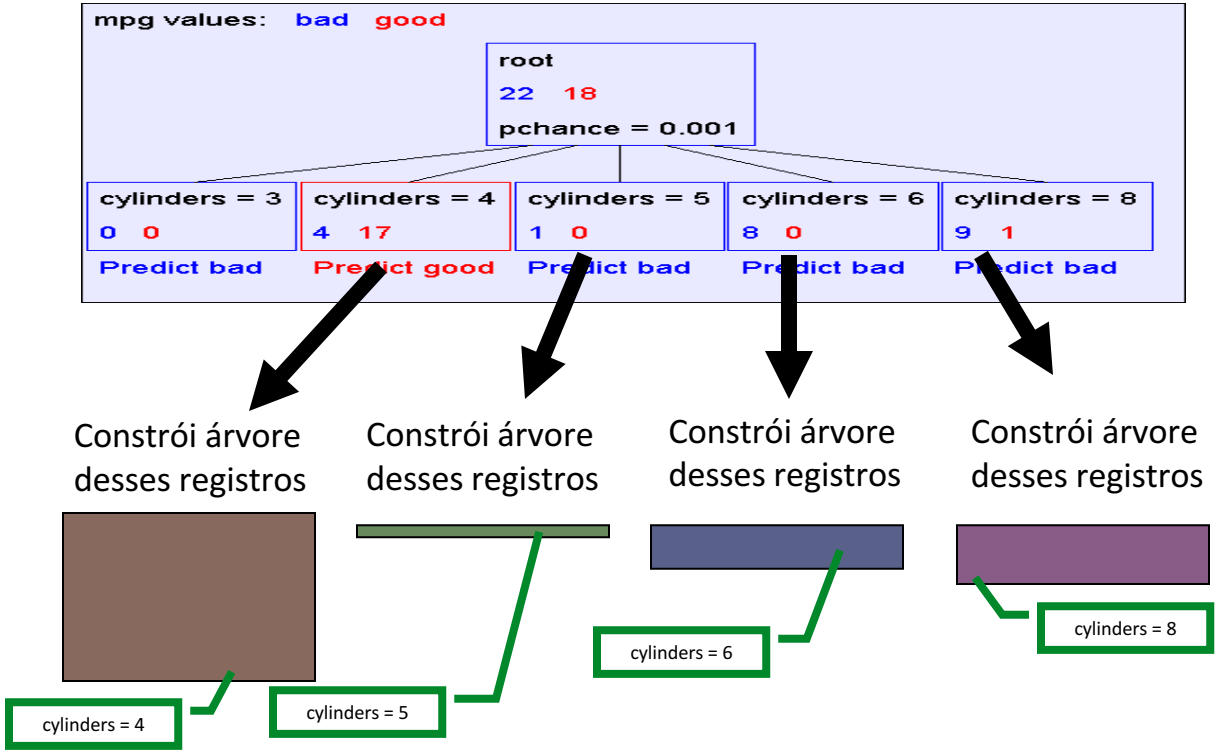
Registros onde cylinders = 4

cylinders = 5

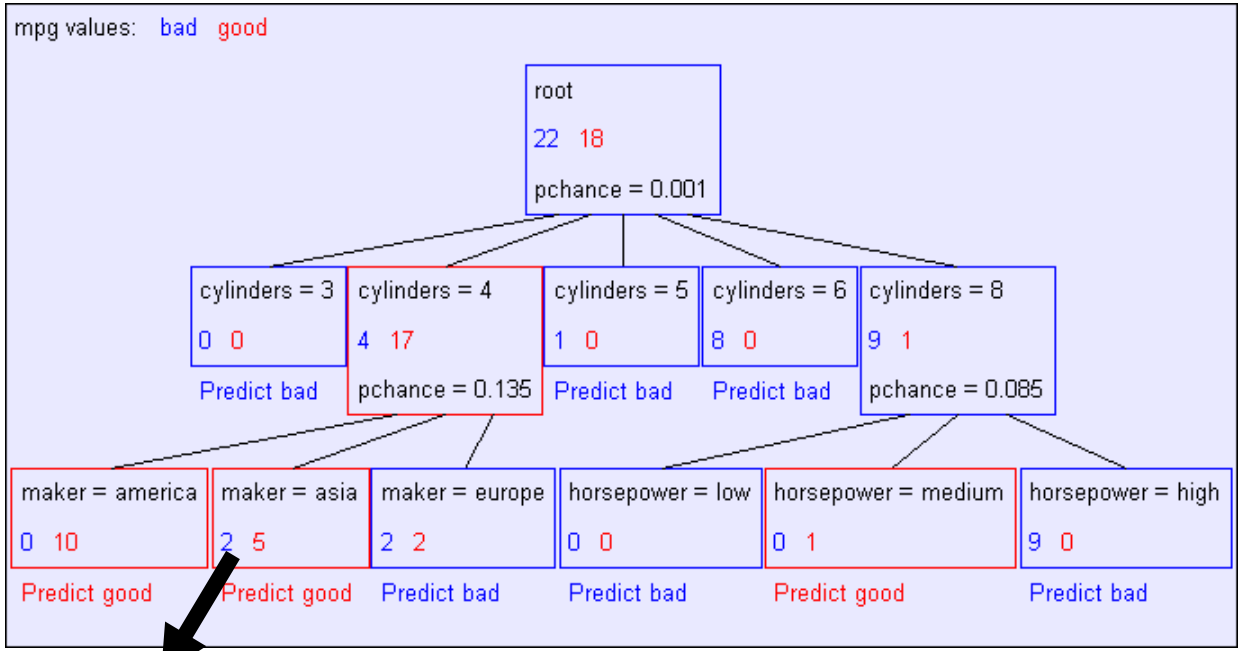
cylinders = 6

cylinders = 8

Passo recursivo

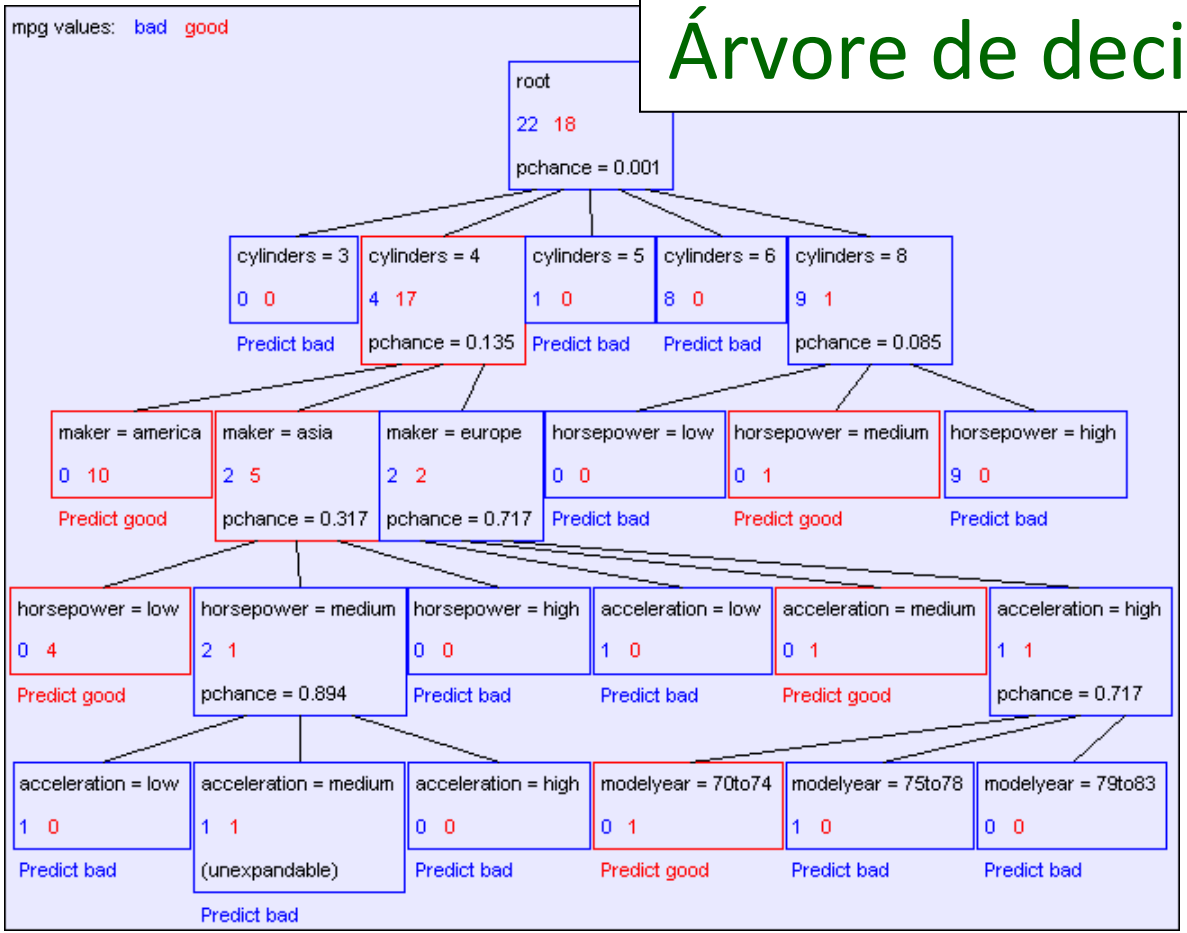


Segundo nível da árvore...

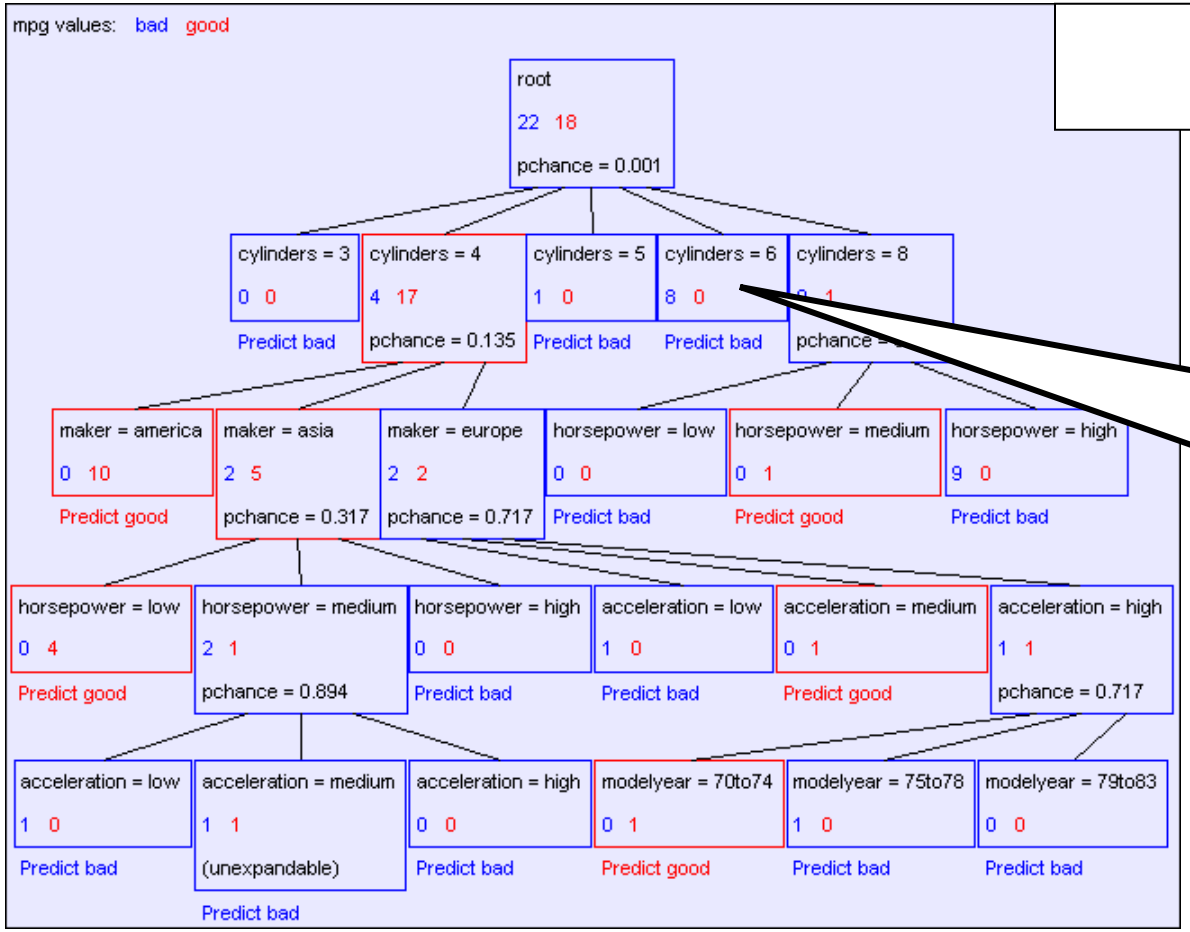


Constrói uma árvore a partir dos 7 (sete) registros em que cylinders = 4 e maker = Asia

Árvore de decisão final

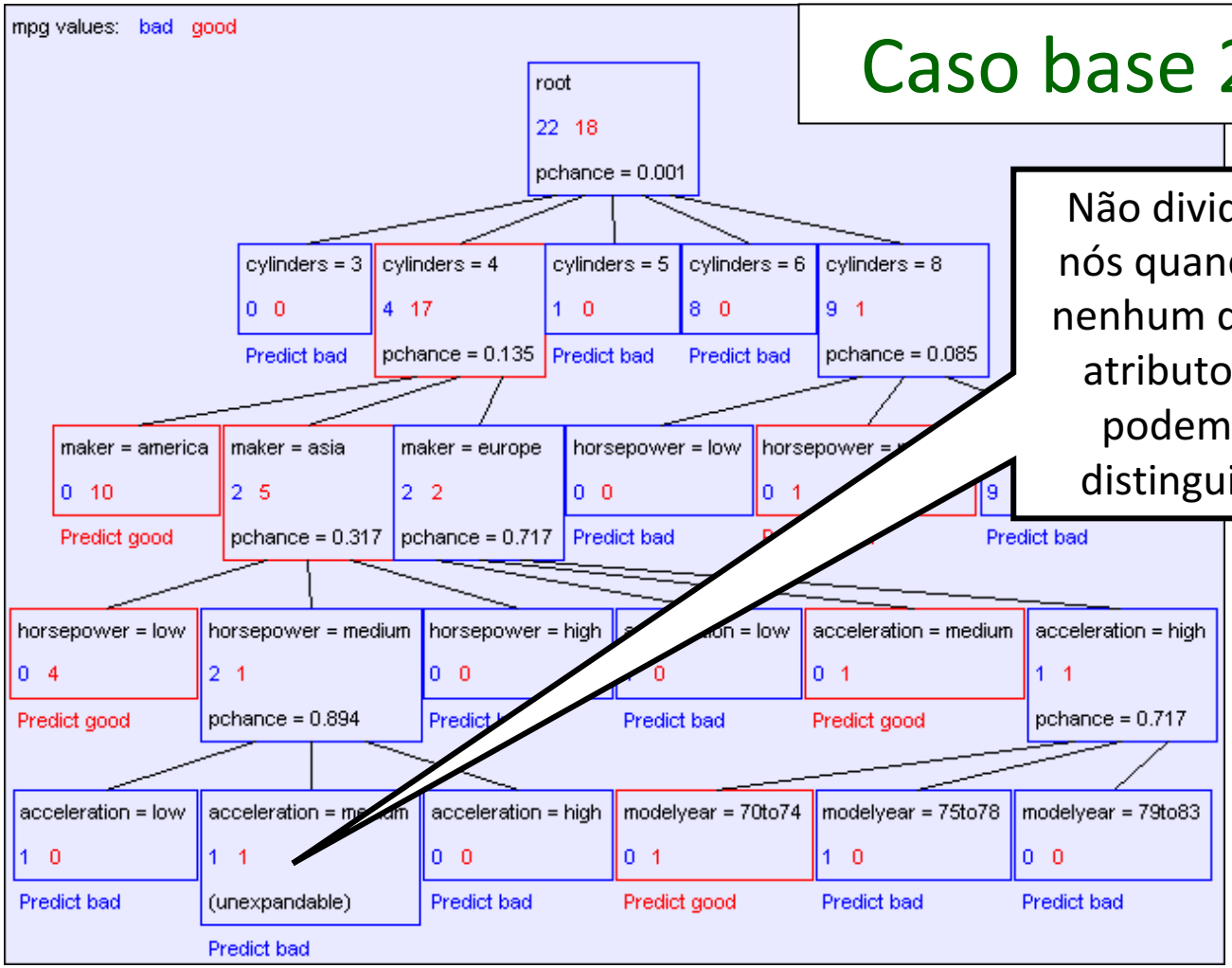


Caso Base 1



Não dividir nós onde todos os registros possuem mesmo valor de saída

Caso base 2



Não dividir nós quando nenhum dos atributos podem distinguir

Information gains using the training set (2 records)

mpg values: bad good

Input	Value	Distribution	Info Gain
cylinders	3		0
	4		
	5		
	6		
	8		
displacement	low		0
	medium		
	high		
horsepower	low		0
	medium		
	high		
weight	low		0
	medium		
	high		
acceleration	low		0
	medium		
	high		
modelyear	70to74		0
	75to78		
	79to83		
maker	america		0
	asia		
	europe		

Algoritmo sumarizado

BuildTree(*DataSet,Output*)

If all output values are the same in *DataSet*, return a leaf node that says “predict this unique output”

If all input values are the same, return a leaf node that says “predict the majority output”

Else find attribute *X* with highest Info Gain

Suppose *X* has n_X distinct values (i.e. X has arity n_X).

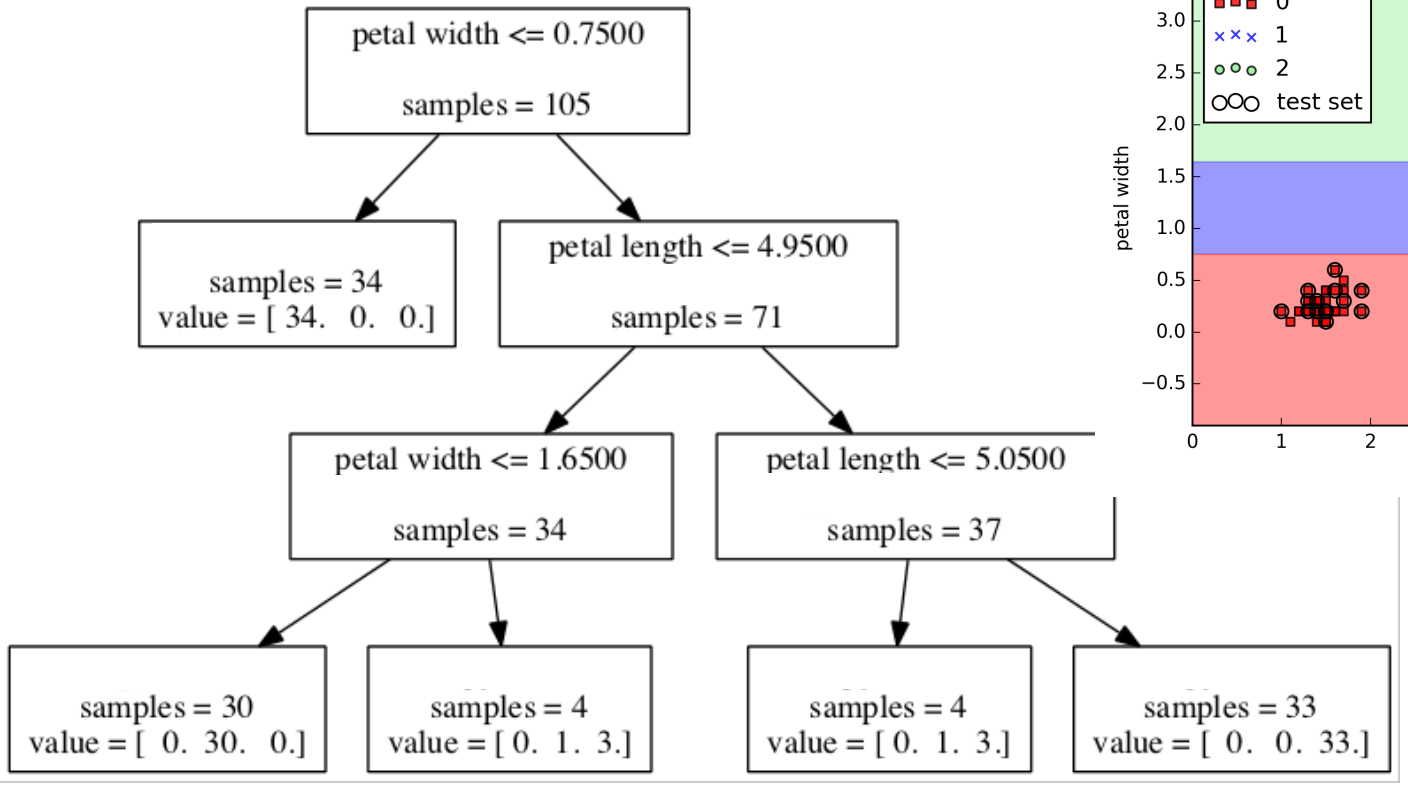
Create and return a non-leaf node with n_X children.

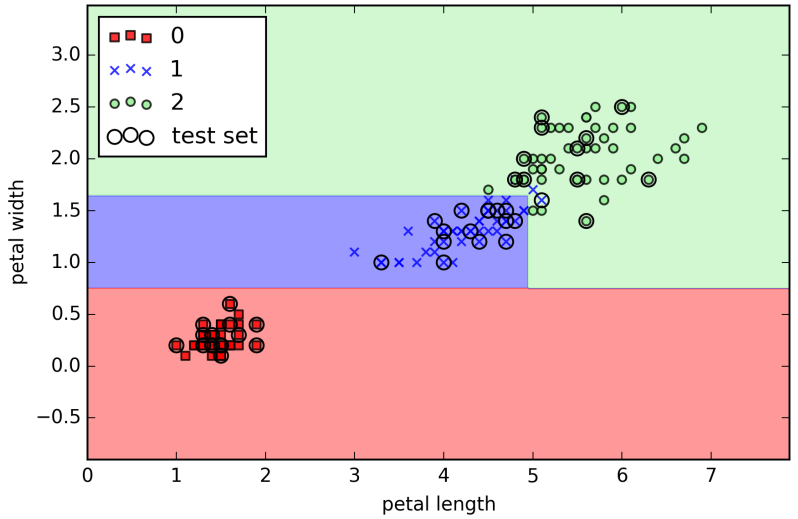
The *i*’th child should be built by calling

BuildTree(DS_i,*Output*)

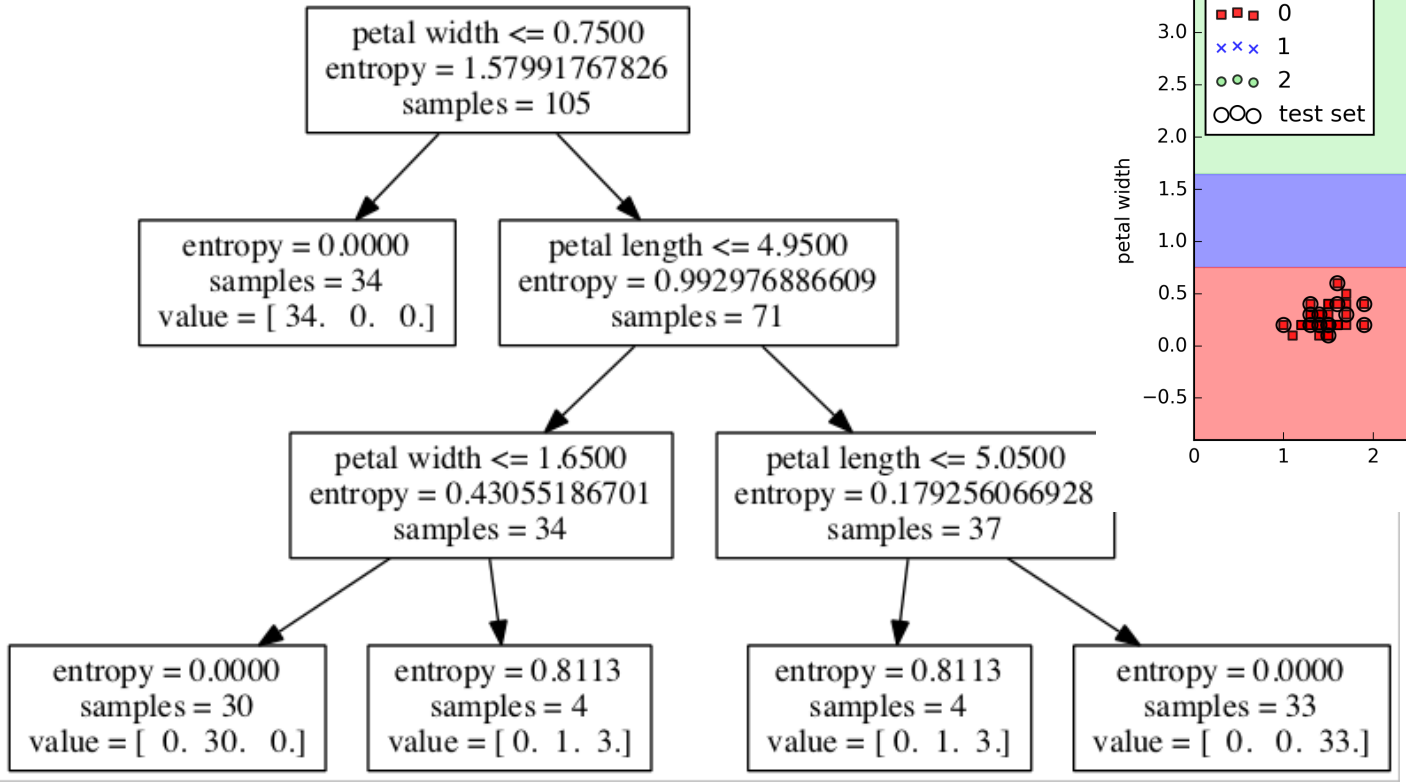
(Where DS_i built consists of all those records in DataSet for which X = *i*th distinct value of X)

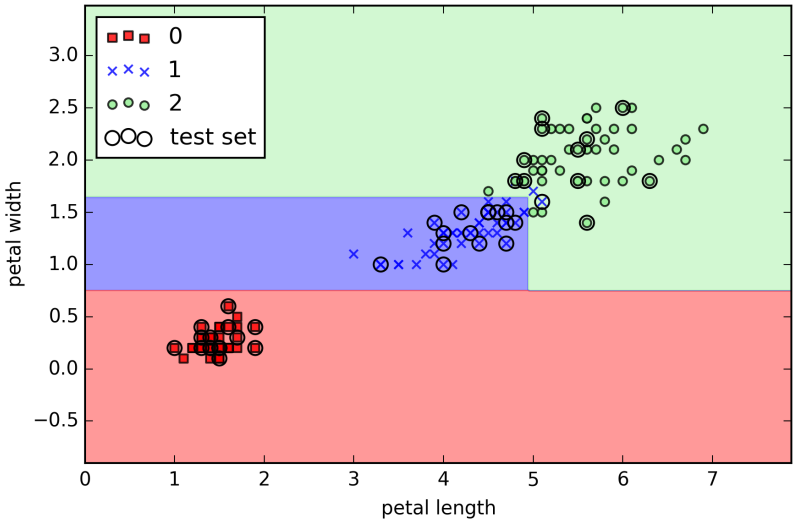
Dataset Iris





Dataset Iris





Exemplo brinquedo (livro): Devo jogar tenis?

Dia	Aspecto	Temp.	Umidade	Vento	Jogar Tênis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

O conteúdo de informação da árvore é calculado a partir das probabilidades das diferentes classificações.

p(sim) = 9/14; p(nao) = 5/14

$I[JogarTenis] = -p(sim)*\log_2(p(sim)) -p(nao)*\log_2(p(nao))$

$= -(9/14)*\log_2(9/14) -(5/14)*\log_2(5/14)$

$= 0,9404858$

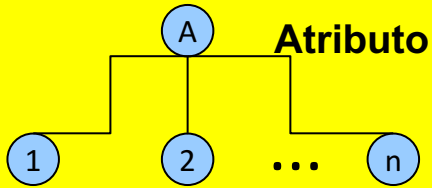
Ganho de Informação

$$Ganho\ \ A = I\ \ C - \sum_{v \in valores\ \ A} \frac{|C_v|}{|C|} \cdot I\ \ C_v$$

Mede a redução na Entropia da (sub)árvore obtida pela divisão escolhida.

Objetivo: escolher a divisão que obtém maior redução (maior Ganho)

Considere um conjunto C de exemplos de treinamento



Atributo A como raiz irá dividir C....

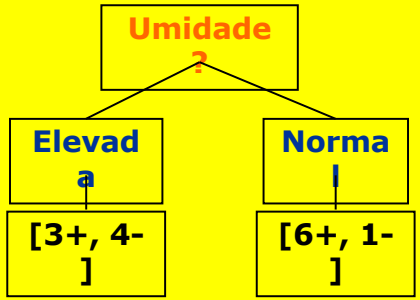
...em subconjuntos {C_1, C_2, ...C_n}

Primeiro passo: são analisados todos os atributos, começando pela Umidade.

$$C = [9\ ,\ 5-]$$

$$I(C) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0,940$$

).	Umidade	Vento	Jogar Tênis
te	Elevada	Fraco	Não
te	Elevada	Forte	Não
te	Elevada	Fraco	Sim
no	Elevada	Fraco	Sim
co	Normal	Fraco	Sim
co	Normal	Forte	Não
co	Normal	Fraco	Sim
no	Elevada	Fraco	Não
co	Normal	Fraco	Sim
no	Normal	Forte	Sim
no	Normal	Forte	Sim
no	Elevada	Forte	Sim
te	Normal	Fraco	Sim
no	Elevada	Forte	Não



$$Ganho\ Umidade = 0,940 - \frac{7}{14} \times 0,985 - \frac{7}{14} \times 0,592$$

$$Ganho\ Umidade = 0,151$$

E(Umidade)

I(C)

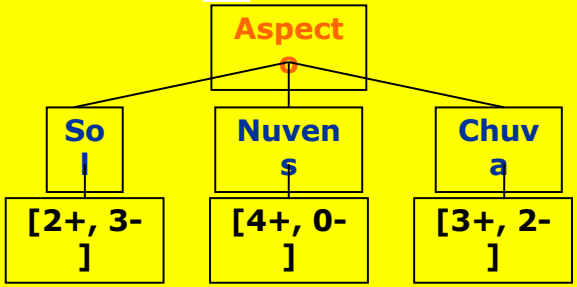
I(Elevada)

I(Normal)

Para o resto dos atributos....

C = [9+, 5−]

I = 0,940



$$MAX\begin{pmatrix} Ganho(\ Umidade) = 0{,}151 \\ Ganho(\ Vento) = 0{,}048 \\ Ganho(\ Aspecto) = 0{,}247 \\ Ganho(\ Temp.) = 0{,}029 \end{pmatrix} = Ganho(\ Aspecto)$$

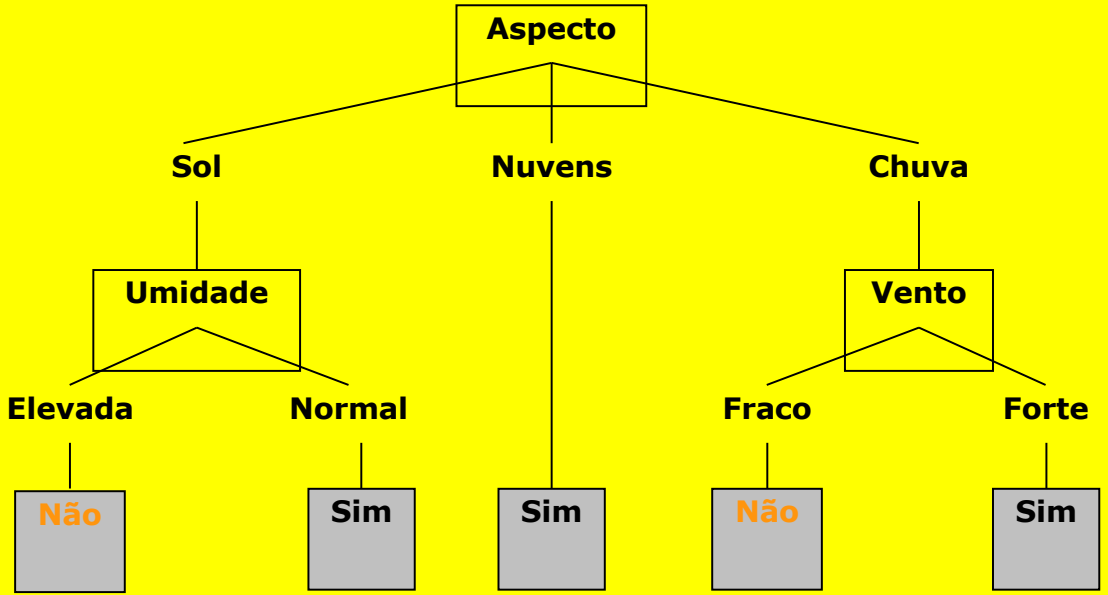
$$Ganho\ (\ Aspecto) = 0{,}940 - \left(5/14\right)\times 0{,}971 - \left(4/14\right)\times 0{,}0 - \left(5/14\right)\times 0{,}971$$

$$Ganho\ (\ Aspecto) = 0{,}247$$

Dia	Aspecto	Temp.	Umidade	Vento	Jogar Tênis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

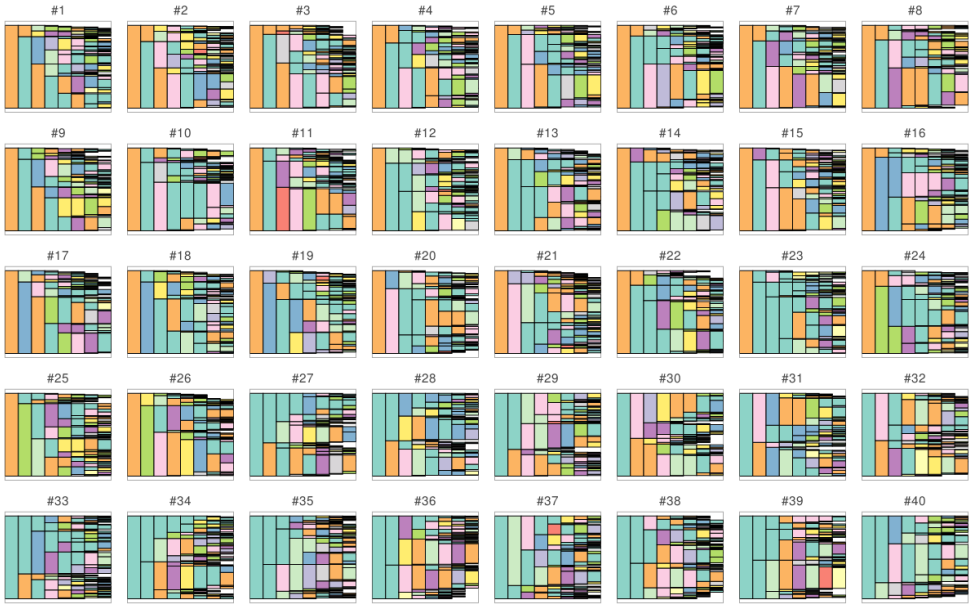
$$MAX\begin{pmatrix} Ganho(C_{Sol}, Umidade) = 0{,}971 - \left(\tfrac{3}{5}\right)\times 0{,}0 - \left(\tfrac{2}{5}\right)\times 0{,}0 = 0{,}971 \\ Ganho(C_{Sol}, Temp.) = 0{,}971 - \left(\tfrac{2}{5}\right)\times 0{,}0 - \left(\tfrac{2}{5}\right)\times 1{,}0 - \left(\tfrac{1}{5}\right)\times 0{,}0 = 0{,}570 \\ Ganho(C_{Sol}, Vento) = 0{,}971 - \left(\tfrac{2}{5}\right)\times 1{,}0 - \left(\tfrac{3}{5}\right)\times 0{,}918 = 0{,}019 \end{pmatrix}$$

$$= Ganho(C_{Sol}, Umidade)$$



$\forall$ x Aspecto(x,Sol) $\wedge$ Umidade(x, Normal) $\Rightarrow$ JogarTenis(x)
$\forall$ x Aspecto(x,Nuvens) $\Rightarrow$ JogarTenis(x)
$\forall$ x Aspecto(x,Chuva) $\wedge$ Vento(x, Forte) $\Rightarrow$ JogarTenis(x)

Árvores de decisão --> Random forests



(ensemble learning)

Inúmeras árvores de decisão

Random forests

1. Draw a random **bootstrap** sample of size n (randomly choose n samples from the training set with replacement).
2. Grow a decision tree from the bootstrap sample. At each node:
    1. Randomly select d features without replacement.
    2. Split the node using the feature that provides the best split according to the objective function, for instance, by maximizing the information gain.
3. Repeat the steps 1 to 2 k times.
4. Aggregate the prediction by each tree to assign the class label by **majority vote.**

Random forests

Majority vote

versicolor

```python
from sklearn.ensemble import RandomForestClassifier

feat_labels = df_wine.columns[1:]

forest = RandomForestClassifier(n_estimators=10000,
                                criterion='entropy',
                                max_features='sqrt',
                                random_state=0,
                                n_jobs=-1)

forest.fit(X_train, y_train)
importances = forest.feature_importances_

indices = np.argsort(importances)[::-1]

for f in range(X_train.shape[1]):
    print("%2d) %-*s %f" % (f + 1, 30,
                            feat_labels[indices[f]],
                            importances[indices[f]]))

plt.title('Feature Importances')
plt.bar(range(X_train.shape[1]),
        importances[indices],
        color='lightblue',
        align='center')

plt.xticks(range(X_train.shape[1]),
           feat_labels[indices], rotation=90)
plt.xlim([-1, X_train.shape[1]])
plt.tight_layout()
#plt.savefig('./random_forest.png', dpi=300)
plt.show()
```

k árvores

d features

```
 1) Color intensity                 0.174273
 2) Flavanoids                      0.167982
 3) Proline                         0.149730
 4) OD280/OD315 of diluted wines    0.145087
 5) Alcohol                         0.096237
 6) Hue                             0.076172
 7) Total phenols                   0.065276
 8) Alcalinity of ash               0.032060
 9) Magnesium                       0.022632
10) Malic acid                      0.022628
11) Proanthocyanins                 0.021253
12) Nonflavanoid phenols            0.013354
13) Ash                             0.013316
```