

# Inteligencia Artificial

## Adquisición automática del conocimiento

Primavera 2007

profesor: Luigi Ceccaroni

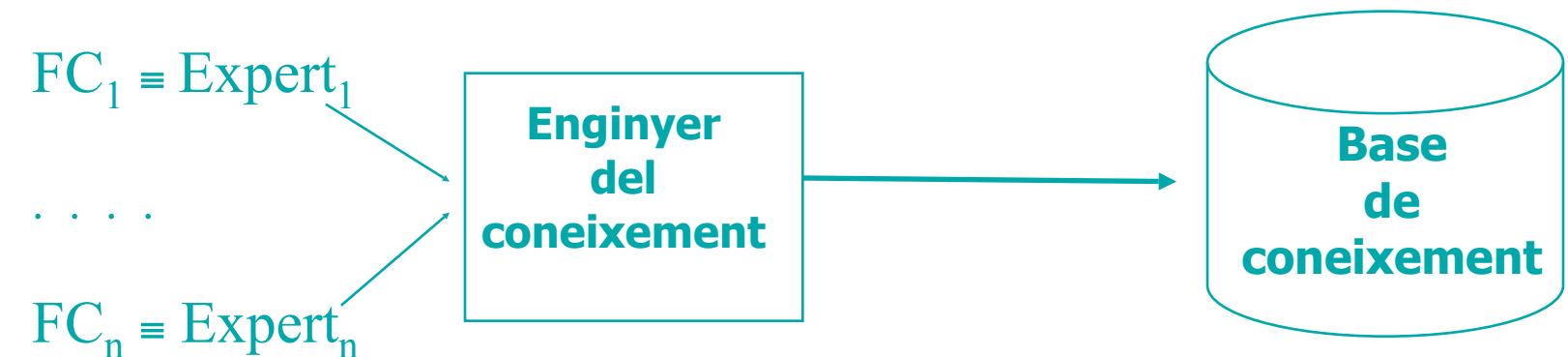


# Ingeniería del conocimiento

- Fa referència a tot el procés de construcció d'un sistema basat en el coneixement:
  - **Adquisició de coneixement**
  - Representació de coneixement
  - Mètode de resolució
  - Construcció de motors d'inferència

# Adquisició del coneixement

- Adquisició del coneixement  $\equiv$  Traspàs del coneixement d'un o més experts (o **fonts de coneixement**) en un domini determinat, cap a un formalisme de representació **computable** del coneixement

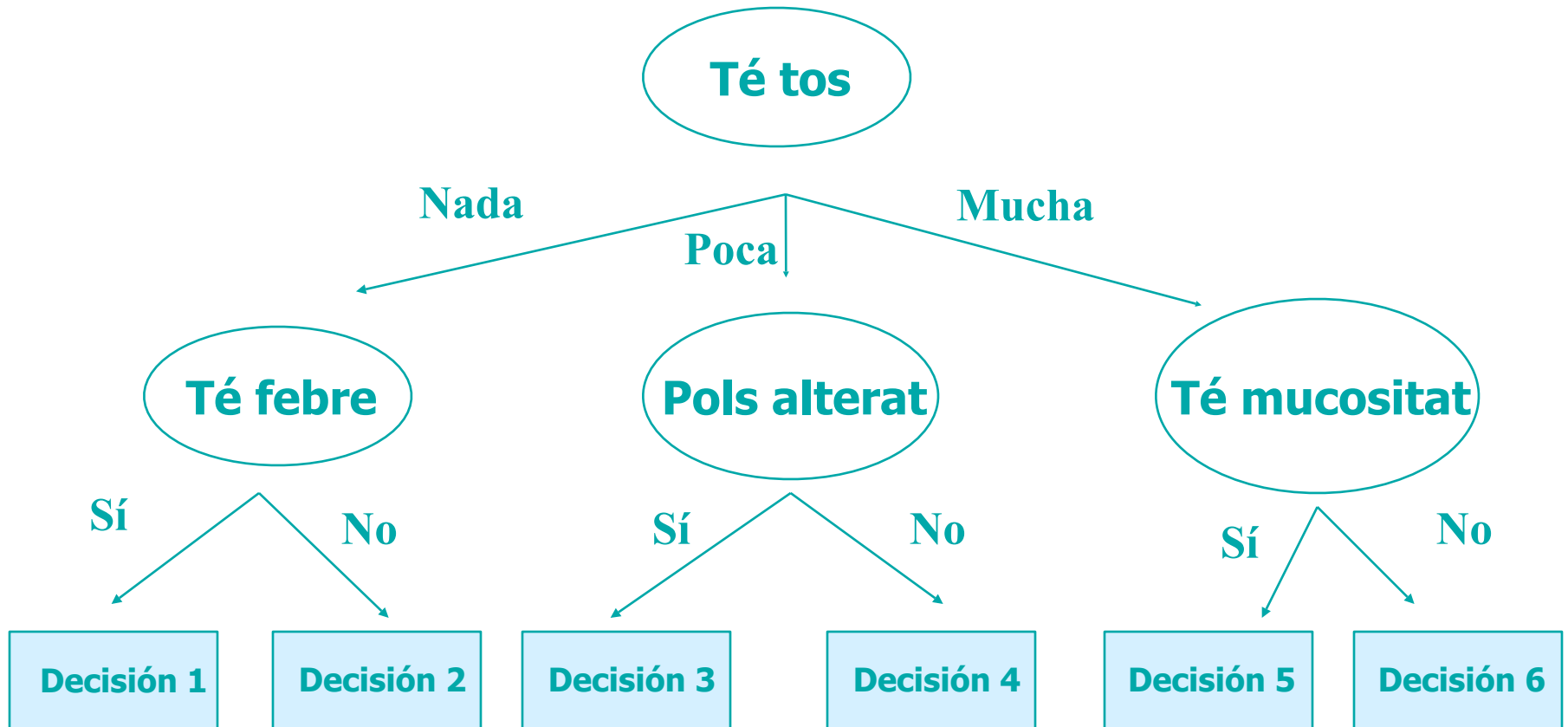


Coneixement  $\begin{cases} \text{Domini (fets, relacions, associacions)} \\ \text{Procés de resolució (heurístiques, mètodes)} \end{cases}$

# Metodologies para la adquisici3n de coneixement

- Com obtenir el coneixement?
  - Interacci3n amb entrevistes
    - Construcci3n d'arbres de decisi3n:
      - Els nodes representen atributs.
      - Les branques representen els possibles valors de l'atribut.
  - Eines automàtiques d'explicitaci3n de coneixements
  - **Tècniques basades en l'aprenentatge automàtic inductiu**

# Arboles de decisión



# Aprendizaje automático inductivo

- Tècniques orientades a problemes d'anàlisi (classificació/interpretació)
- L'expert expressa el seu coneixement en una forma habitual per a ell: *observacions/exemples*.
- Es transforma aquesta representació en la del sistema.
- Es requereix una *validació* de l'expert.

# Aprendizaje automático inductivo

- Exemple: dades sobre un gimnàs

Client	Sexe	Horari	Edat	Anys al	Act1	Act2	Piscina	Desc.
1	Dona	Tarda	40	2	Aerobics	Stretch	Si	0
4	Home	Tarda	35	6	Aerobics	Stretch	Si	0
6	Dona	Mati	30	3	Aerobics	Ioga	Si	0
7	Home	Tarda	28	4	TBC	Steps	No	10
10	Home	Tarda	32	4	TBC	Stretch	Si	10
11	Home	Mati	60	10	Ioga	TBC	Si	0
12	Home	Mati	34	8	Stretch	TBC	Si	0
13	Home	Tarda	87	2	TBC	TBC	Si	0

# Aprendizaje automático inductivo

- Objectiu: **Agrupar** objectes semblants
  - Hi ha poca informació del domini i es vol començar a tenir-ne una idea més clara.
- Tècniques:
  - Mètodes d'agrupació (clustering)
- Exemple: Es descobreix que hi ha dues agrupacions:
  - Classe 1
  - Classe 2

# Aprendizaje automático inductivo

- Objectiu: **Classificar nous objectes**
- *Es parteix d'una situació més informada, sabent que existeixen grups ja definits.*
- **Determinar les característiques peculiars** de cada grup, per poder ubicar un nou objecte en la classe que li correspon.

Client	Sexe	Horari	Edat	Anys al	Act1	Act2	Piscina	Classe
1	Dona	Tarda	40	2	Aerobics	Stretch	Si	Classe1
4	Home	Tarda	35	6	Aerobics	Stretch	Si	Classe2
6	Dona	Mati	30	3	Aerobics	Ioga	Si	Classe1
7	Home	Tarda	28	4	TBC	Steps	No	Classe1
10	Home	Tarda	32	4	TBC	Stretch	Si	Classe1
11	Home	Mati	60	10	Ioga	TBC	Si	Classe2
12	Home	Mati	34	8	Stretch	TBC	Si	Classe2
13	Home	Tarda	87	2	TBC	TBC	Si	Classe2

# Aprendizaje automático inductivo

- Métodes:
  - Arbres de decisió:  
CART, ID3, ASSISTANT, C4.5, C5.1
  - Regles de classificació:  
If Act1 is steps  
Then  
Act2 is ioga  
Rule's probability: 0.9  
The rule exists in 52 records

# ID3

- **ID3**  $\equiv$  Induction Decision Tree [Quinlan, 1979, 1986]
- Tècnica d'aprenentatge automàtic
- Inducció d'arbres de decisió
- Estratègia top-down
- A partir d'un conjunt d'**exemples/instàncies** i la classe a la qual pertanyen, crea l'arbre de decisió *millor* que expliqui les instàncies.

# ID3

És un algorisme voraç (*greedy*)  
que selecciona a cada pas el  
*millor* atribut



El *millor* és el més discriminant  
(potencialment més útil)

# ID3

- El procés de construcció és **iteratiu**:
  1. Es selecciona un subconjunt (finestra) de exemples del conjunt d'entrenament (*training set*).
  2. Es construeix l'arbre de decisió que permeti discriminar el conjunt d'exemples de la finestra
  3. **Si** l'arbre de decisió induït explica la resta d'exemples del conjunt d'entrenament
    - Llavors**  
l'arbre de decisió es el definitiu
    - Sinó**  
els exemples mal classificats (excepcions)  
s'afegeixen a la finestra i es torna a (2)
    - fSi**

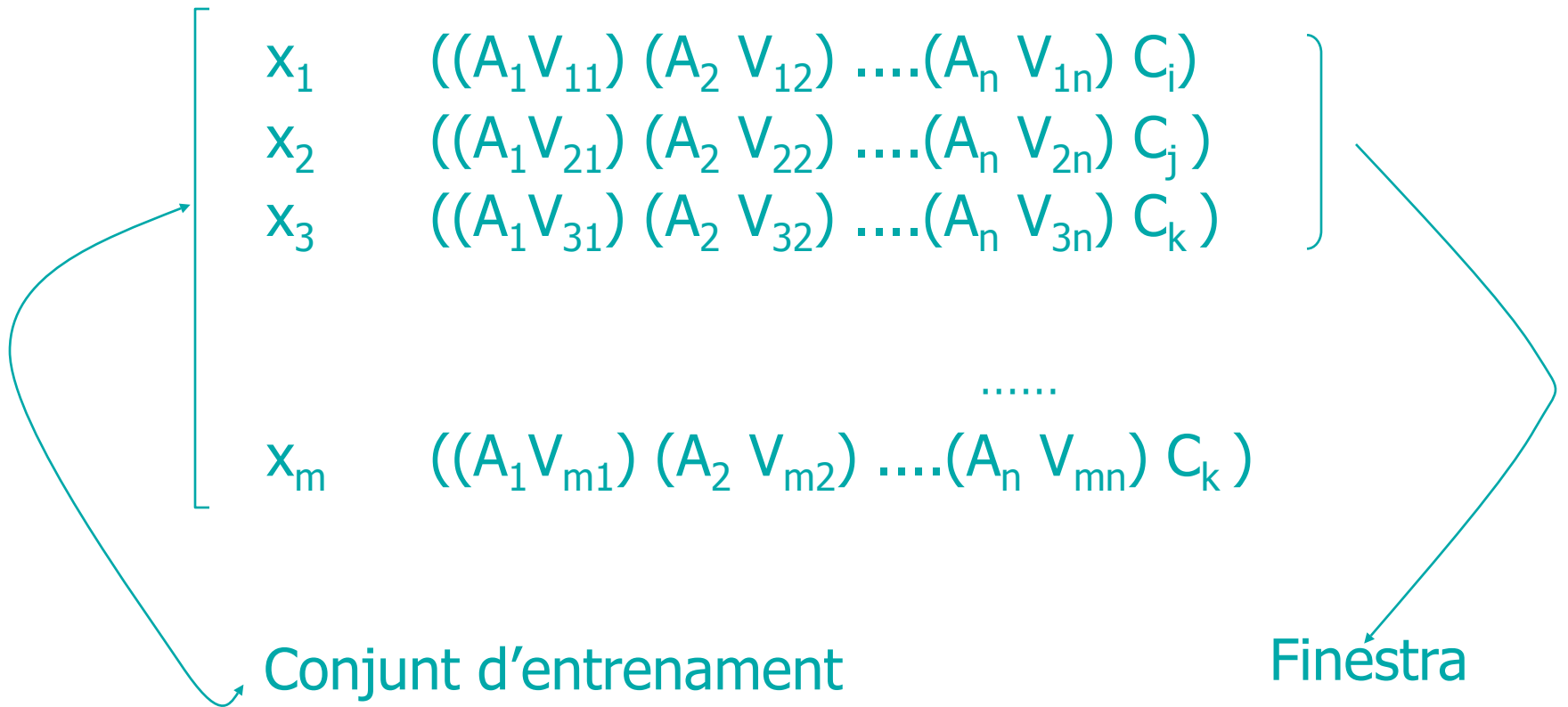
# ID3: formalización

- $X \equiv$  conjunt d'exemples  $\equiv \{x_j\}_{j=1,m}$
- $A \equiv$  conjunt d'atributs  $\equiv \{A_k\}_{k=1,n}$
- $C \equiv$  conjunt de classes  $\equiv \{C_i\}_{i=1,p}$
- $\# \equiv$  Cardinalitat

$A_i \in A$ , un atribut  
 $x \in X$ , un exemple  
 $v$ , un valor

$V(A_i) \equiv$  conjunt de valors d' $A_i$   
 $\equiv \{v_l\}_{l=1,q}$   
 $A_i(x) \equiv$  valor de  $x$  per  $A_i$   
 $A_i^{-1}(v) = \{x \in X : A_i(x) = v\}$

# ID3



# ID3: idea básica

- Seleccionar a cada pas l'atribut que pot discriminar més.
- La selecció es fa maximitzant una certa funció  $G(X,A)$ .

# ID3: criterio de selecció

- Seleccionar  $A_k$  tal que *maximitza* el guany d'informació

$$G(X, A_k) = I(X, C) - E(X, A_k) \Leftrightarrow \text{minimitzem } E(X, A_k) \approx 0$$

On

informació

$$I(X, C) = - \sum_{C_i \in C} p(X, c_i) * \log_2 p(X, c_i)$$

$C_i \in C$

Probabilitat de que un exemple pertanyi a la classe  $C_i$

$$P(X, c_i) = \# C_i / \#X$$

entropia

$$E(X, A_k) = \sum_{v_l \in V(A_k)} [ p(X, v_l) * I(A_k^{-1}(v_l), C) ]$$

$v_l \in V(A_k)$

$$P(X, v_l) = \# A_k^{-1}(v_l) / \#X$$

Probabilitat de que un exemple tingui el valor  $v_l$  en l'atribut  $A_k$

# ID3: ejemplo

	Ulls	Cabell	Estatura	Classe
E1	Blaus	Ros	Alt	C+
E2	Blaus	Moreno	Mitjà	C+
E3	Marrons	Moreno	Mitjà	C-
E4	Verds	Moreno	Mitjà	C-
E5	Verds	Moreno	Alt	C+
E6	Marrons	Moreno	Baix	C-
E7	Verds	Ros	Baix	C-
E8	Blaus	Moreno	Mitjà	C+

# ID3: ejemplo

$$I(X, C) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$(1,2,5,8)$                        $(3,4,5,7)$   
**C+**                                      **C-**

$$E(X, \text{Ulls}) = \frac{3}{8} (-1 \log_2 1 - 0 \log_2 0) + \frac{2}{8} (-0 \log_2 0 - 1 \log_2 1) + \frac{3}{8} (-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}) = 0.344$$

$$E(X, \text{Cabell}) = \frac{2}{8} (-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}) + \frac{6}{8} (-\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6}) = 1$$

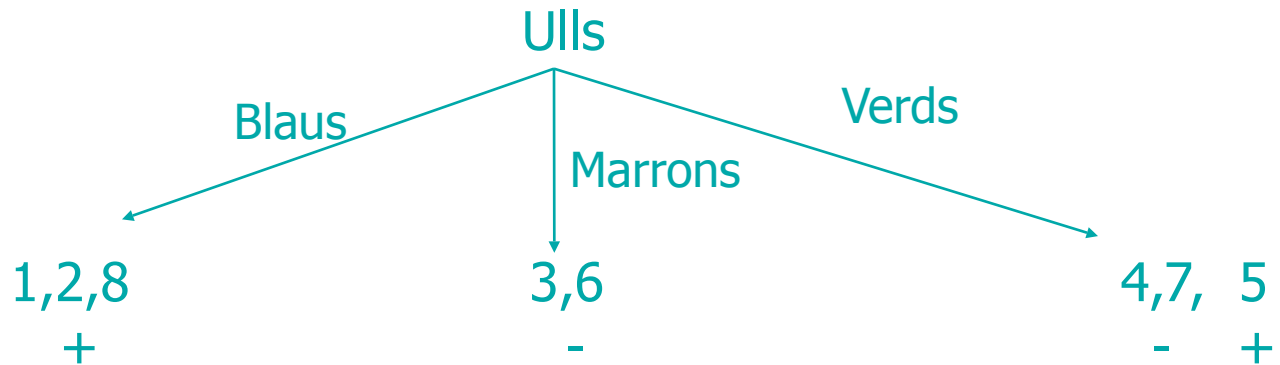
$$E(X, \text{Estatura}) = \frac{2}{8} (-1 \log_2 1 - 0 \log_2 0) + \frac{4}{8} (-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}) + \frac{2}{8} (-0 \log_2 0 - 1 \log_2 1) = 0.5$$

# ID3: ejemplo

$$G(X, \text{Ulls}) = 1 - 0.366 = \mathbf{0.634}$$

$$G(X, \text{Cabell}) = 1 - 1 = 0$$

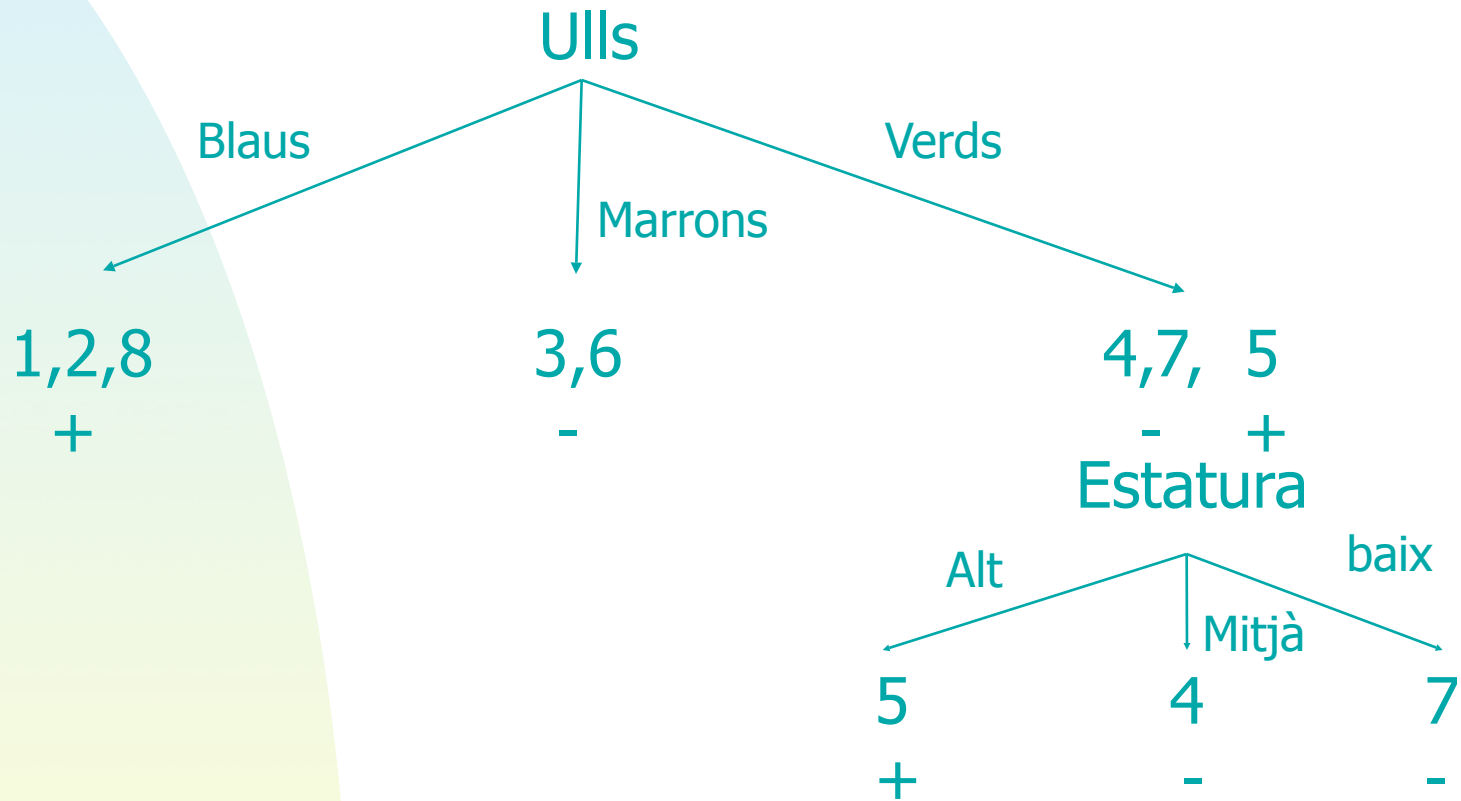
$$G(X, \text{Estatura}) = 1 - 0.5 = 0.5$$



	Cabell	Estatura	Classe
E4	Moreno	Mitjà	C-
E5	Moreno	Alt	C+
E7	Ros	Baix	C-



# ID3: ejemplo



# ID3: ejemplo

Ulls = Blaus  $\rightarrow C+$

Ulls = Marrons  $\rightarrow C-$

Ulls = Verds  $\wedge$  Estatura = Alt  $\rightarrow C+$

Ulls = Verds  $\wedge$  Estatura = Mitjà  $\rightarrow C-$

Ulls = Verds  $\wedge$  Estatura = Baix  $\rightarrow C-$

# ID3: algoritmo

**funció** ID3 (X,A són conjunts) **retorna** arbre\_de\_decisió **és**

**var** arbre1,arbre2 **són** arbre\_de\_decisió **fvar**

**opció**

**cas**  $(\exists C_i: \forall x_j \in X \rightarrow x_j \in C_i)$  **fer**

arbre1 := crear\_arbre (C<sub>i</sub>)

**cas no**  $(\exists C_i: \forall x_j \in X \rightarrow x_j \in C_i)$  **fer**

**opció**

**cas**  $A \neq \emptyset$  **fer**

$A_{\max} := \max_{A_k \in A} \{G(X, A_k)\};$

arbre1 := crear\_arbre(A<sub>max</sub>);

**pertot**  $v \in V(A_{\max})$  **fer**

arbre2 := ID3(A<sup>-1</sup><sub>max</sub> (v), A-{A<sub>max</sub>});

arbre1 := afegir\_branca(arbre1, arbre2, v)

**fpertot**

**cas**  $A = \emptyset$  **fer**

arbre1 := crear\_arbre(classe\_majoritària(X))

**fopció**

**fopció**

**retorna** arbre1

**ffunció**

# Lecturas recomendadas

## Artículos

- Brian R. Gaines & Mildred L. G. Shaw, **ELICITING KNOWLEDGE AND TRANSFERRING IT EFFECTIVELY TO A KNOWLEDGE-BASED SYSTEM**, *IEEE Transactions on Knowledge and Data Engineering* (September 1992)