Algoritmi de clasificare

ARBORI DE DECIZIE

Clasificarea

- Se dă o mulțime de antrenare: o mulțime de instanțe (vectori de antrenare, obiecte)
  - Datele de antrenare
- Instanțele au atribute
- Fiecare instanță are atribute cu anumite valori
- De obicei, ultimul atribut este clasa

Starea vremii	Temperatură	Umiditate	Vânt	Joc
Soare	Mare	Mare	Absent	Nu
Soare	Mare	Mare	Prezent	Nu
Înnorat	Mare	Mare	Absent	Da
Ploaie	Medie	Mare	Absent	Da
Ploaie	Mică	Normală	Absent	Da
Ploaie	Mică	Normală	Prezent	Nu
Înnorat	Mică	Normală	Prezent	Da
Soare	Medie	Mare	Absent	Nu
Soare	Mică	Normală	Absent	Da
Ploaie	Medie	Normală	Absent	Da
Soare	Medie	Normală	Prezent	Da
Înnorat	Medie	Mare	Prezent	Da
Înnorat	Mare	Normală	Absent	Da
Ploaie	Medie	Mare	Prezent	Nu

Tipuri de atribute

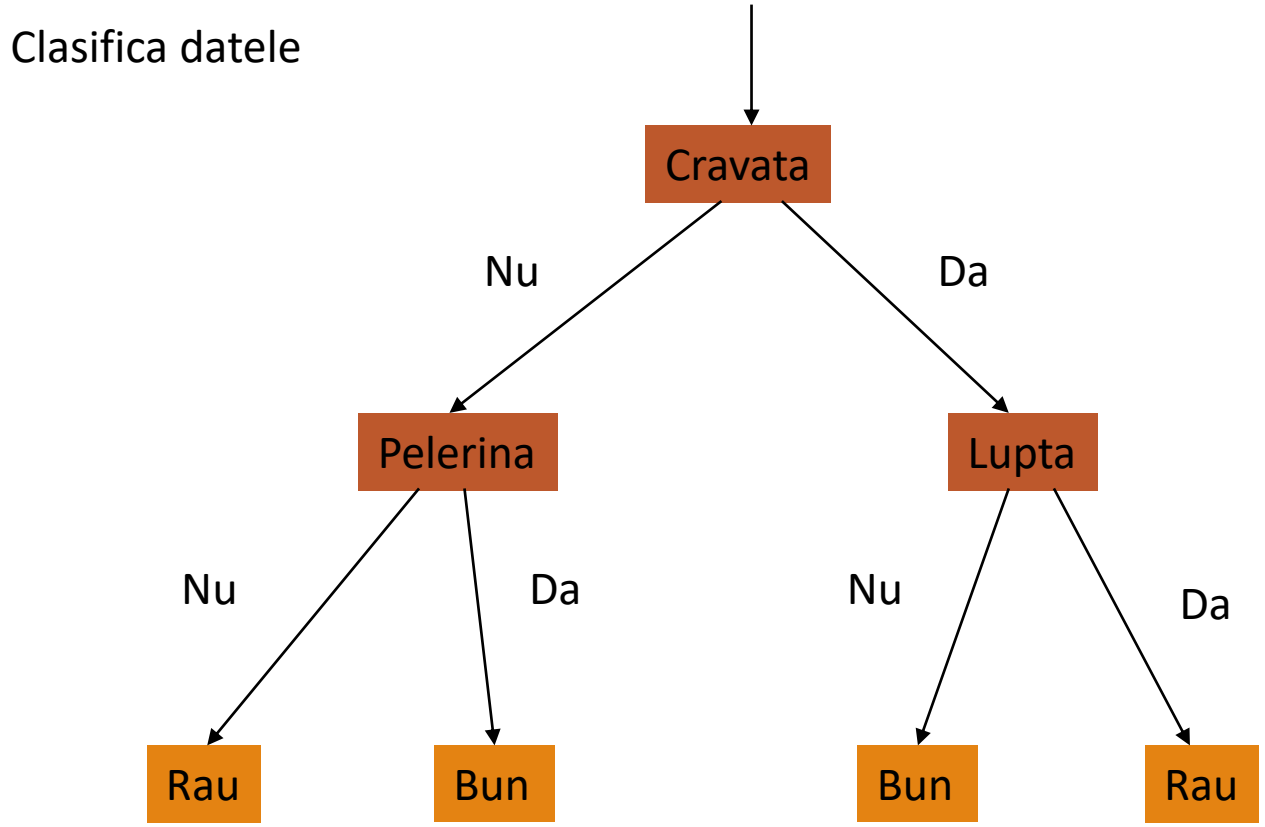
Există patru tipuri de atribute, organizate pe două coordonate:

- Atribute *simbolice (calitative)*: de tip *nominal* (*ex.* culoarea ochilor, nume, sex, CNP ca obiect, nu număr) şi *ordinal* (înălțime (mică, medie, mare), ranguri, calificative)
- Atribute *numerice (cantitative)*: de tip *interval* (*Temperatura în °C, date calendaristice)* şi *raţional* (lungime, distanță, prețuri)

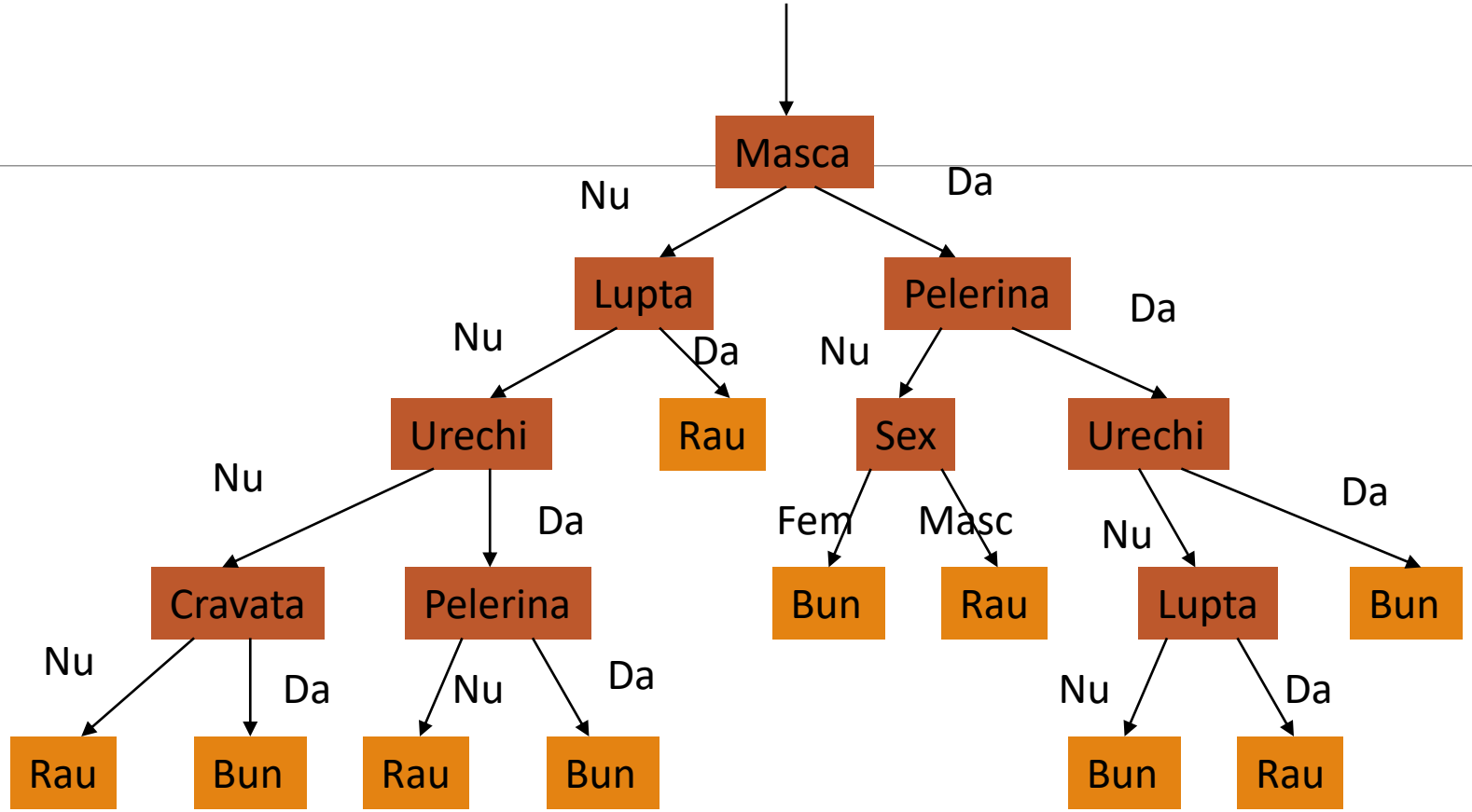
Exemplu de problema de clasificare

Atribute / Instante	Sex	Masca	Pelerina	Cravata	Urechi	Lupta	**Clasa**
	Set de invatare						
Batman	Masc	Da	Da	Nu	Da	Nu	**Bun**
Robin	Masc	Da	Da	Nu	Nu	Nu	**Bun**
Alfred	Masc	Nu	Nu	Da	Nu	Nu	**Bun**
Pinguin	Masc	Nu	Nu	Da	Nu	Da	**Rau**
Catwoman	Fem	Da	Nu	Nu	Da	Nu	**Rau**
Joker	Masc	Nu	Nu	Nu	Nu	Nu	**Rau**
	Date de test						
Batgirl	Fem	Da	Da	Nu	Da	Nu	**??**
Fred	Masc	Da	Nu	Nu	Nu	Nu	**??**

Conditii pentru o invatare "buna“
Arbore de decizie



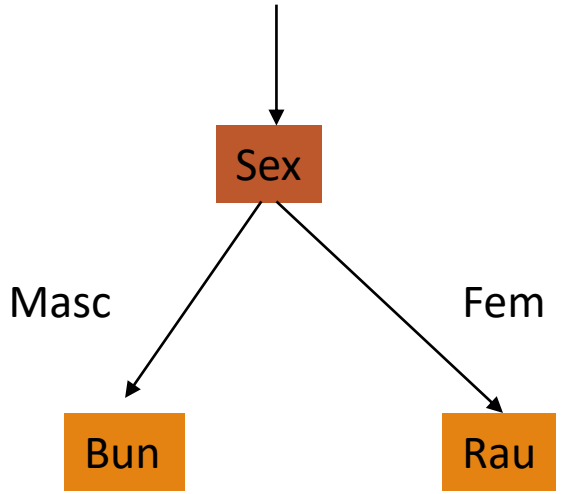
Conditii pentru o invatare "buna"



Clasifica datele dar complexitate prea mare (intuitiv)

Conditii pentru o invatare "buna"

Prea simplu, nu clasifica corect



Aleg prima varianta (cf. lamei lui Occam)

Conditii pentru o invatare "buna"

- Clasificatoarele trebuie sa fie suficient de "expresive" pentru a fi in concordanta cu setul de invatare
- Dar clasificatoarele care au o complexitate prea mare pot duce la fenomenul de "overfit" (overfitting) = include zgomot sau sabloane de date nerelevante

Occam Razor

Principiul lamei lui Occam

- ***prefer explicatiile simple celor complexe***

Wiliam of Occam, 1285 – 1349 (?)

filozof englez

"non sunt multiplicanda entia praeter necessitatem"

Invatarea inductiva prin AD

Vede invatarea ca achizitia cunostintelor structurate

Reprezentarea cunostintelor = **arbori de decizie** (AD)

Problema de invatare = **clasificare**

Invatare supervizata

Aplicatii posibile

Strategie = invatare batch (ne-incrementala)

AD se construieste pornind de la radacina spre frunze = ***Top Down Induction of Decision Tree***

Exemple

- Mediu – istorie a observatiilor
- Profesor – expert in domeniu

ID3 (Quinlan)

Univers de obiecte **U** descrise in termenii unei colectii de atribute **{A}**

Fiecare **atribut** masoara o caracteristica importanta a unui obiect **o∈U**

Domeniul de valori atribute D_A= discret, simbolic (ulterior extins)

Fiecare obiect apartine unui **clase** dintr-o multime de clase mutual exclusive **{Cl}**

Se da **setul de invatare** (SI)

Problema = obtinerea unor **reguli de clasificare** / construirea unui **AD** care clasifica corect nu numai **∀o∈SI** dar si **∀o∈U**

ID3 (Quinlan)

Structura iterativa – fereastra din SI

S-au gasit AD corecti in cateva iteratii pt 30 000 obiecte cu 50 atribute

Empiric s-a aratat ca iterativ se obtin arbori mai buni decat daca s-ar construi din tot SI

Utilizare AD

Reguli de decizie

ID3 (Quinlan)

Metoda de constructie

C = multmea de obiecte / ex inv. din SI

A – atribut test cu valori / iesiri A_1, .. A_n

$[C_1, ..C_n]$, cu $C_i = \{o \in C \mid A = A_i\}$

"divide-and-conquer"

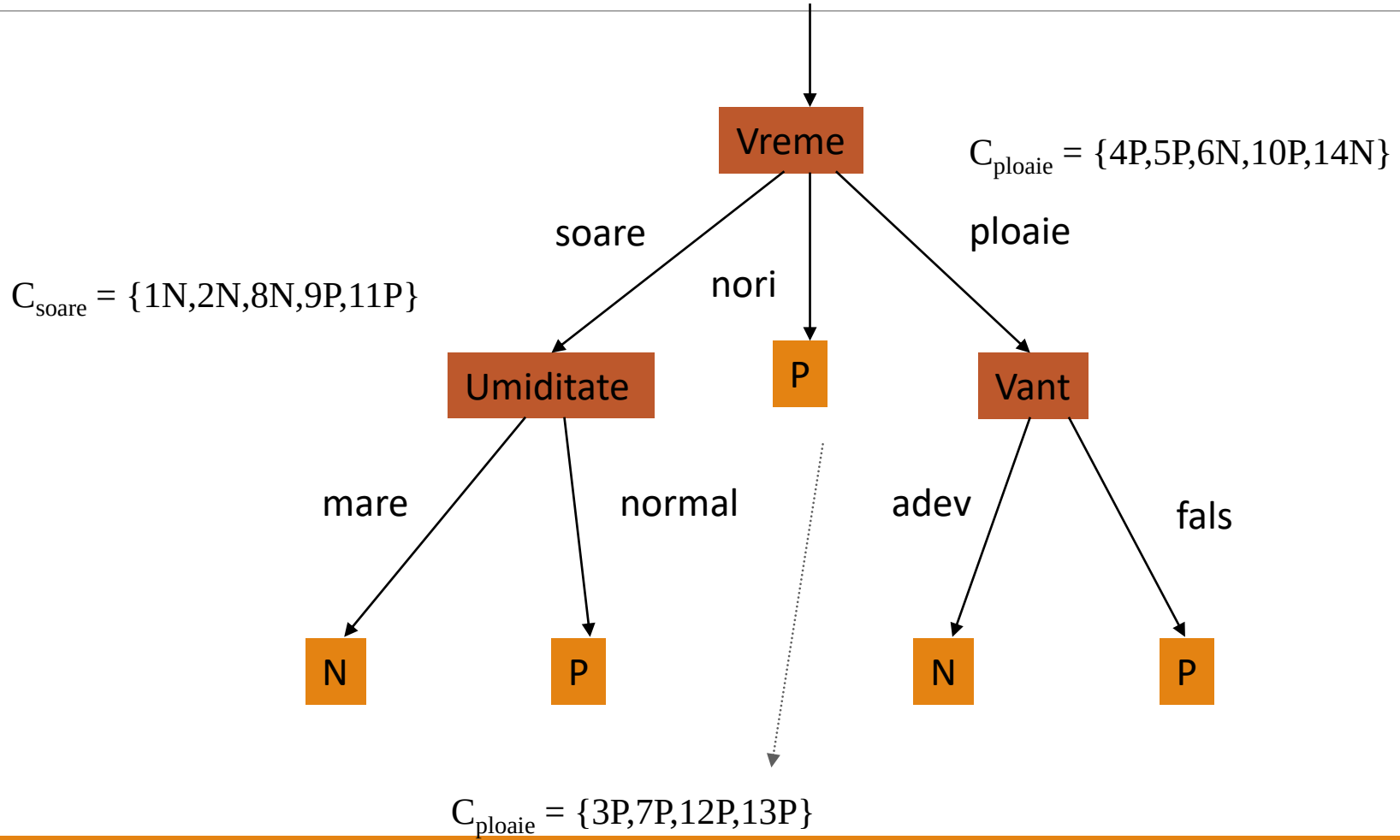
Impartirea/expandarea AD se opreste cand toate C_i apartin unei aceleiasi clase

Se termina intotdeauna (in cazul cel mai nefavorabil, cate un obiect in fiecare clasa)

ID3 – Exemplul 1

No.	Atribute				Clasa
	Vreme	Temperatura	Umiditate	Vant	
1	soare	cald	mare	fals	N
2	soare	cald	mare	adev	N
3	nori	cald	mare	fals	P
4	ploaie	placut	mare	fals	P
5	ploaie	racoare	normal	fals	P
6	ploaie	racoare	normal	adev	N
7	nori	racoare	normal	adev	P
8	soare	placut	mare	fals	N
9	soare	racoare	normal	fals	P
10	ploaie	placut	normal	fals	P
11	soare	placut	normal	adev	P
12	nori	placut	mare	adev	P
13	nori	cald	normal	fals	P
14	ploaie	placut	mare	adev	N

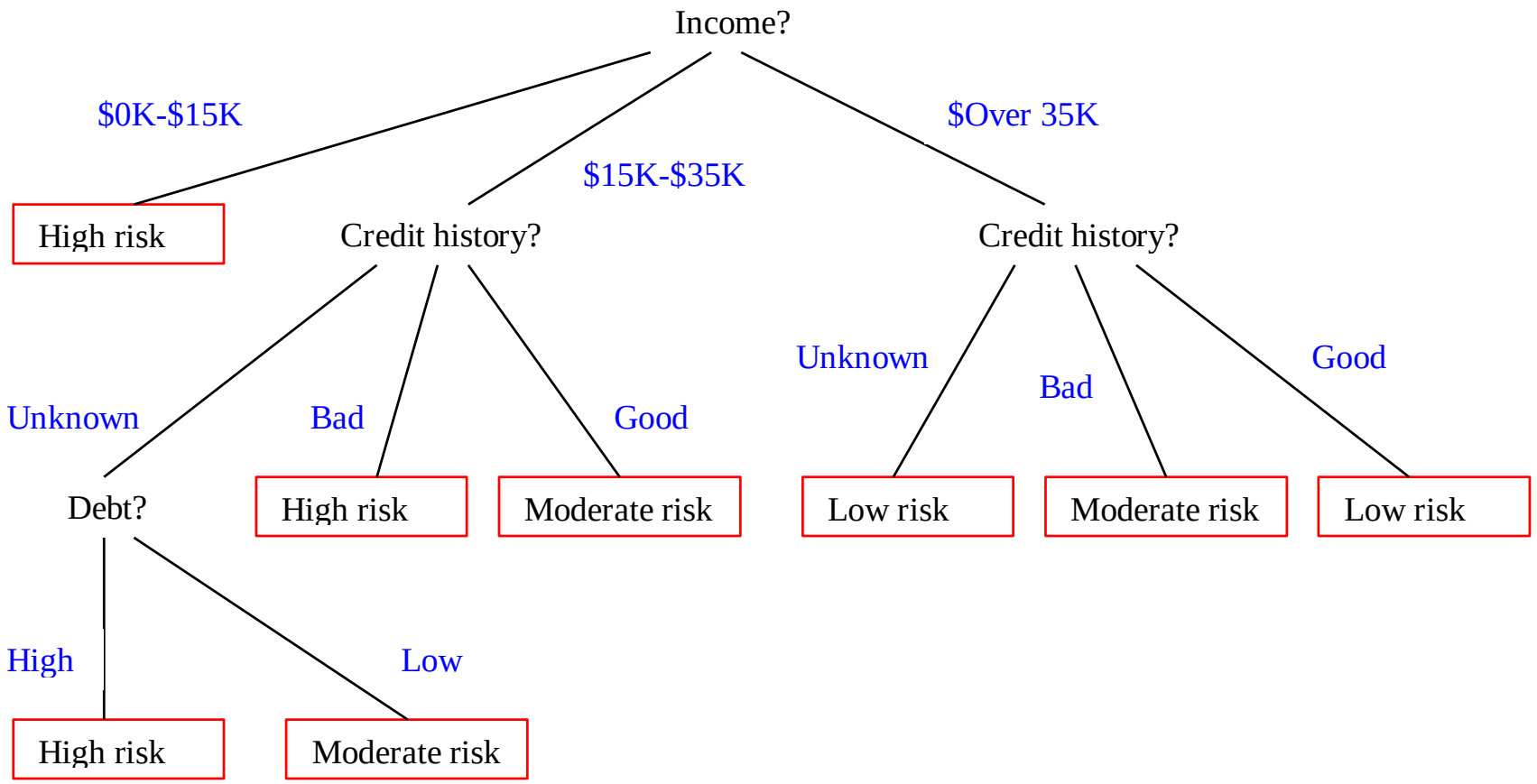
ID3 – Exemplul 1



ID3 – Exemplul 2 (mai multe clase)

No.	*Risk (Classification)*	*Credit History*	*Debt*	*Collateral*	**Income**
1	High	Bad	High	None	$0 to $15k
2	High	Unknown	High	None	$15 to $35k
3	Moderate	Unknown	Low	None	$15 to $35k
4	High	Unknown	Low	None	$0k to $15k
5	Low	Unknown	Low	None	Over $35k
6	Low	Unknown	Low	Adequate	Over $35k
7	High	Bad	Low	None	$0 to $15k
8	Moderate	Bad	Low	Adequate	Over $35k
9	Low	Good	Low	None	Over $35k
10	Low	Good	High	Adequate	Over $35k
11	High	Good	High	None	$0 to $15k
12	Moderate	Good	High	None	$15 to $35k
13	Low	Good	High	None	Over $35k
14	High	Bad	High	None	$15 to $35k

ID3 – Exemplu mai multe clase



ID3 – Arbore minim

Din acelasi SI se pot contrui diferiti AD

- Cum se poate obtine cel mai mic arbore (lama lui Occam) ?
- Cum selectez atributul din radacina unui arbore?

ID3 – Cum selectez A?

C cu $p \in$ P si $n \in$ N

Se presupune ca:

(1) Orice AD corect va clasifica obiectele proportional cu reprezentarea lor in C

Un obiect arbitrar $o \in$ C va fi clasificat:

- $\in$P cu probabilitatea *p/(p+n)*
- $\in$N cu probabilitatea *n/(p+n)*

(2) Cand un AD este utilizat pentru a clasifica obiecte, acesta intoarce o clasa $\Rightarrow$

AD poate fi vazut ca o sursa a unui mesaj 'P' sau 'N' avand informatia necesara pentru a genera acest mesaj

Teoria informatiei ofera criteriul

Teoria informatiei furnizeaza fundamentul matematic pentru masurarea continutului de informatie dintr-un mesaj

Un mesaj este privit ca o instanta dintr-un univers al tuturor mesajelor posibile

Transmiterea mesajului este echivalenta cu selectia unui anumit mesaj din acest univers

Teoria informatiei ofera criteriul

Pentru un univers de mesaje

$M = \{m_1, m_2, \ldots, m_n\}$

si o probabilitate $p(m_i)$ de aparitie a fiecarui mesaj, **continutul informational *I(M)*** al mesajelor din *M* se defineste astfel:

$$I(M) = \sum_{i=1}^{n} -p(m_i) \log_2(p(m_i))$$

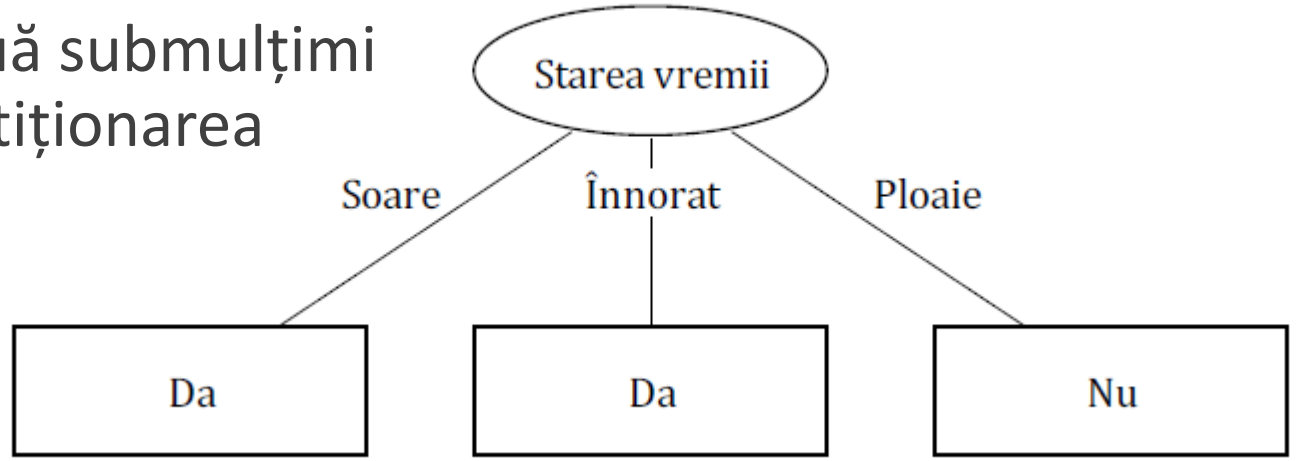
Testul de atribute

- Urmează o strategie greedy: se partiționează mulțimea de instanțe cu un test care maximizează un anumit criteriu
- Depinde de tipul atributului: nominal, ordinal sau continuu
- Depinde de numărul de posibilități de partiționare: binar sau multiplu

Atribute nominale

- Partiționarea multiplă
  - Numărul de partiții = numărul de valori distincte
- Partiționarea binară
  - Se împart valorile în două submulțimi
  - Trebuie descoperită partiționarea optimă

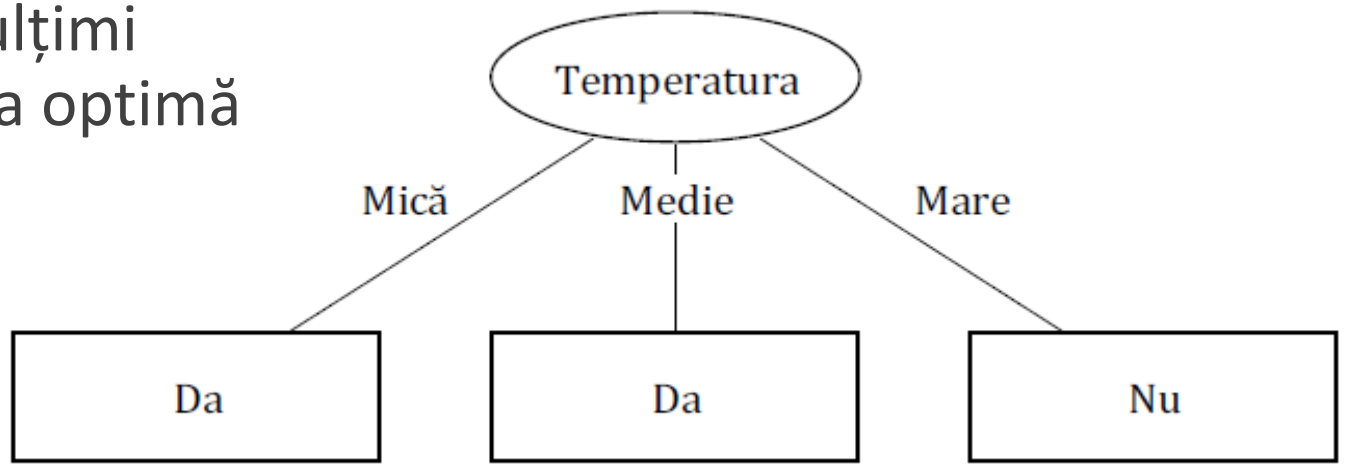
Starea vremii	Joc
Soare	Da
Înnorat	Da
Ploaie	Nu



Atribute ordonale

- Partiționarea multiplă
  - Numărul de partiții = numărul de valori distincte
- Partiționarea binară
  - Se împart valorile în două submulțimi
  - Trebuie descoperită partiționarea optimă

Temperatură	Joc
Mică	Da
Medie	Da
Mare	Nu



Atribute continue

- Se discretizează datele pentru a le transforma în atribute ordinale
  - Cu interval egal (histograma)
  - Cu frecvență egală (mulțimi cu numere egale de instanțe)
  - Grupare (clustering)
- Decizie binară: ($Ai < v$) sau ($Ai > v$)
  - Trebuie considerate toate partiționările posibile
  - Necesită un efort de calcul mai mare

Discretizarea

- Cu interval egal – de exemplu, 3 intervale
  - [65, 75], (75, 85], (85, 95]

Umiditate	Joc
65	Da
70	Da
72	Da
75	Da
80	Da
85	Da
86	Nu
90	Nu
90	Nu
91	Nu
93	Nu
95	Nu

Umiditate-Dis1	Joc
Mică	Da
Mică	Da
Mică	Da
Mică	Da
Medie	Da
Medie	Da
Mare	Nu
Mare	Nu
Mare	Nu
Mare	Nu
Mare	Nu
Mare	Nu

Discretizarea

- Cu frecvență egală
  - de exemplu, 3 intervale

Umiditate	Joc
65	Da
70	Da
72	Da
75	Da
80	Da
85	Da
86	Nu
90	Nu
90	Nu
91	Nu
93	Nu
95	Nu

Umiditate-Dis2	Joc
Mică	Da
Mică	Da
Mică	Da
Mică	Da
Medie	Da
Medie	Da
Medie	Nu
Medie	Nu
Mare	Nu
Mare	Nu
Mare	Nu
Mare	Nu

Discretizarea

- Binară
  - de exemplu, 85

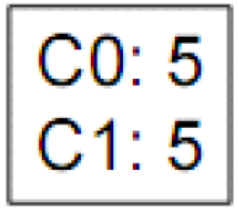
Umiditate	Joc
65	Da
70	Da
72	Da
75	Da
80	Da
85	Da
86	Nu
90	Nu
90	Nu
91	Nu
93	Nu
95	Nu

$(A_i \leq 85)?$

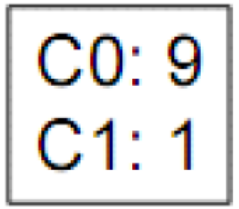
Umiditate	Umiditate-Bin	Joc
65	Da	Da
70	Da	Da
72	Da	Da
75	Da	Da
80	Da	Da
85	Da	Da
86	Nu	Nu
90	Nu	Nu
90	Nu	Nu
91	Nu	Nu
93	Nu	Nu
95	Nu	Nu

Partiţionarea optimă

- Euristică: se preferă nodurile cu cele mai omogene distribuții de clasă
  - Necesită o măsură a „impurității” nodurilor



Ne-omogene
Grad mare de impuritate

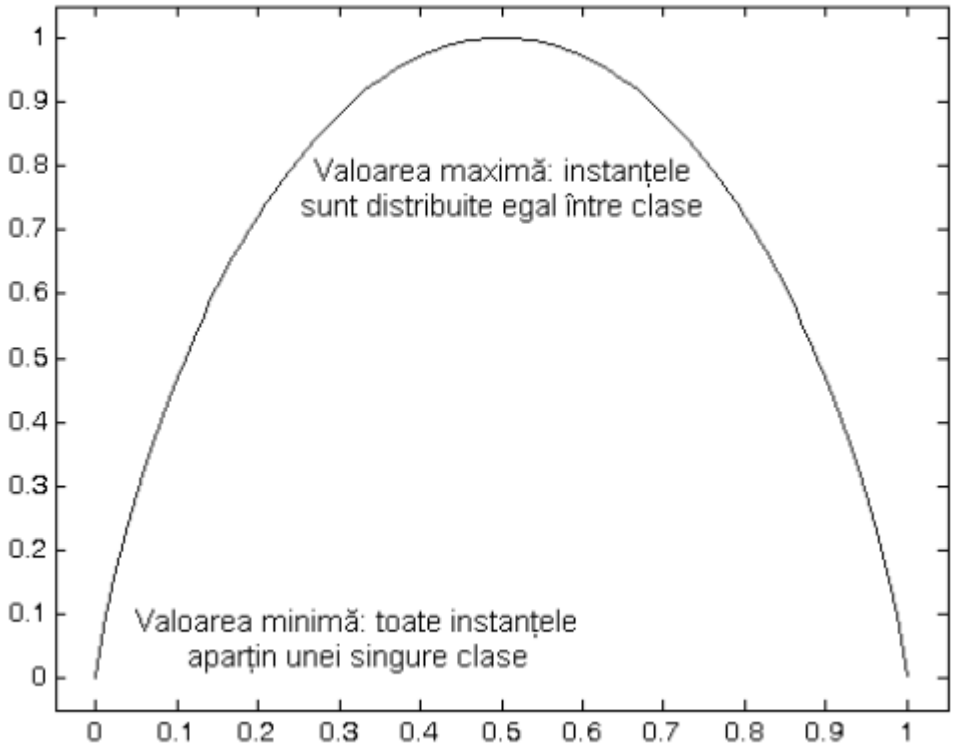


Omogene
Grad mic de impuritate

Măsuri de impuritate
Entropia

Shannon, 1948

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot \log_b P(x_i),$$

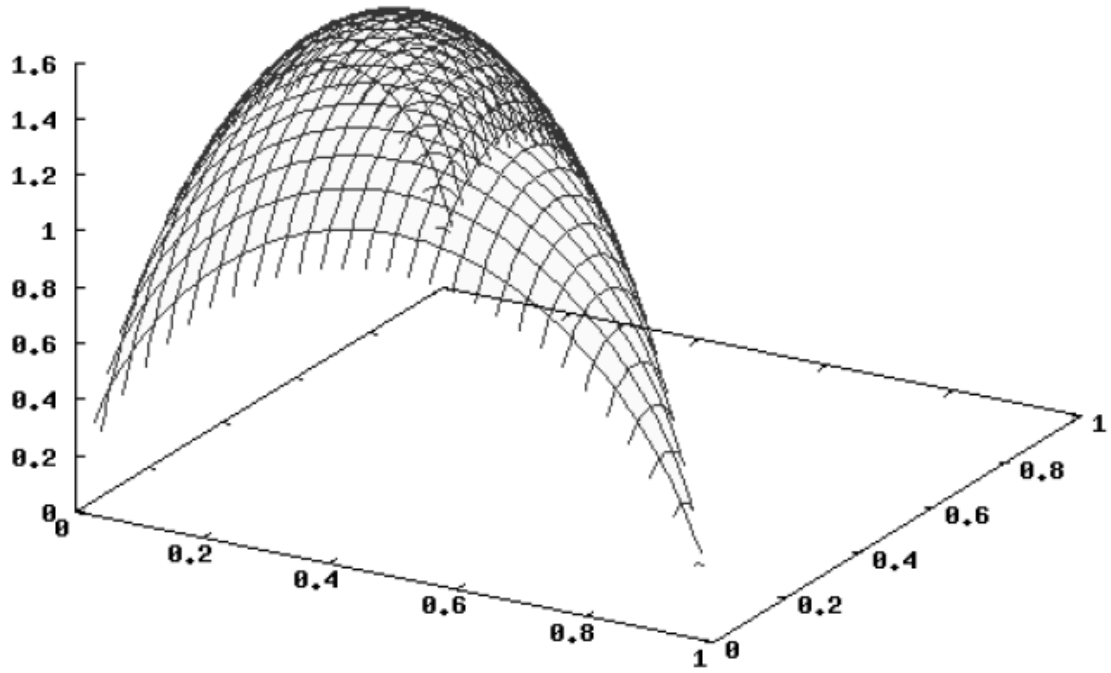


Graficul entropiei pentru 2 clase

Măsuri de impuritate
Entropia

Shannon, 1948

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot \log_b P(x_i),$$

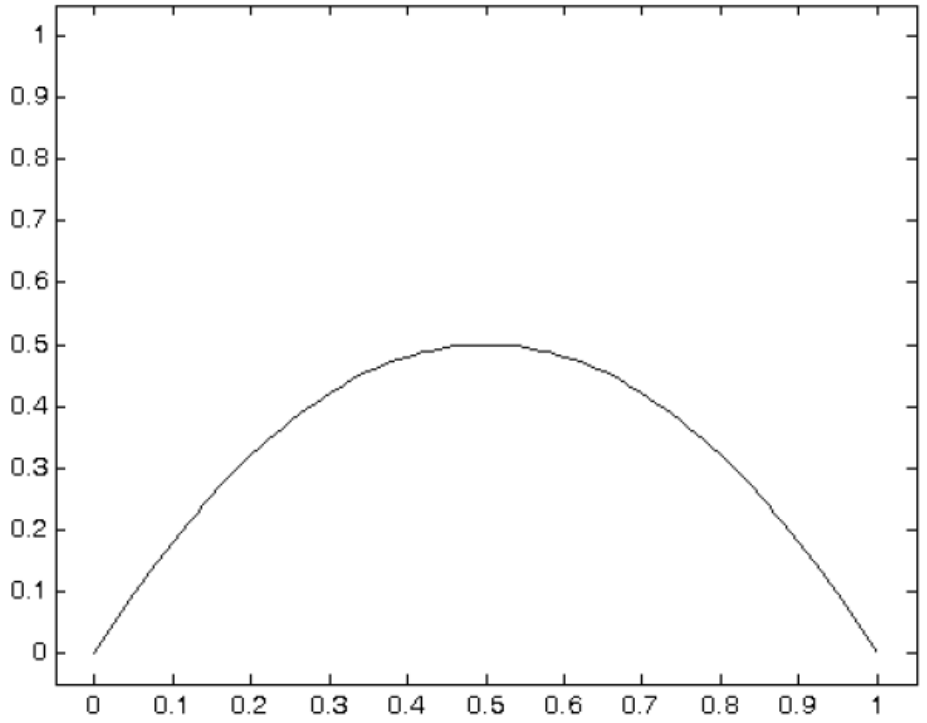


Graficul entropiei pentru 3 clase

Măsuri de impuritate
Indexul Gini

Breinman et al., 1984

$$G(X) = 1 - \sum_{i=1}^{n} (P(x_i))^2.$$



Graficul indexului Gini pentru 2 clase

Măsuri de impuritate Exemplu

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

Node N_1	Count
Class=0	0
Class=1	6

$$\text{Gini} = 1 - (0/6)^2 - (6/6)^2 = 0$$
$$\text{Entropy} = -(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$$

Node N_2	Count
Class=0	1
Class=1	5

$$\text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$$
$$\text{Entropy} = -(1/6)\log_2(1/6) - (5/6)\log_2(5/6) = 0.650$$

Node N_3	Count
Class=0	3
Class=1	3

$$\text{Gini} = 1 - (3/6)^2 - (3/6)^2 = 0.5$$
$$\text{Entropy} = -(3/6)\log_2(3/6) - (3/6)\log_2(3/6) = 1$$

Partiţionarea

- Când un nod părinte p este partiționat în k fii, calitatea partiționării (de exemplu, entropia) se calculează astfel:

$$H_s = \sum_{i=1}^{k} \frac{n_i}{n} \cdot H_i,$$

- unde ni este numărul de instanțe din fiul i și n este numărul de instanțe din nodul p
- s este o partiţionare (engl. “split”) din mulţimea tuturor partiţionărilor posibile
- Formulă similar pentru indexul Gini

Câştigul informaţional

- Calitatea unei partiționări este determinată de cresterea omogenității submultimilor rezultate
- Trebuie maximizat câștigul informațional:

$$\Delta_s = H_p - H_s = H_p - \sum_{i=1}^{k} \frac{n_i}{n} \cdot H_i.$$

- Deoarece nodului părinze este același pentru toți fiii se preferă valoarea minimă

$$s^* = \underset{s}{\operatorname{argmax}}\, \Delta_s = \underset{s}{\operatorname{argmin}} H_s = \underset{s}{\operatorname{argmin}} \sum_{i=1}^{k_s} \frac{n_i}{n} \cdot H_i.$$

- Termenul de „câștig informațional” se utilizează când se folosește entropia ca măsură de impuritate, dar principiul este același pentru indexul Gini sau orice altă măsură de impuritate

Inducţia unui arbore de decizie Algoritmul lui Hunt

- Fie *Dn* mulțimea instanțelor de antrenare care ajung la un nod *n*
- Algoritmul lui Hunt (Hunt 1962; Tan, Steinbach & Kumar, 2006):
  - Dacă *Dn* conține instanțe din aceeași clasă *yn* , atunci *n* este o frunză etichetată *yn*
  - Dacă *Dn* este o mulțime vidă, atunci *n* este o frunză etichetată cu clasa implicită (*default*) *yd*
  - Dacă *Dn* conține instanțe care aparțin mai multor clase, se utilizează un test de atribute pentru a partiționa datele în mulțimi mai mici
  - Se aplică recursiv procedura pentru fiecare submulțime

Exemplu
Clasificarea folosind un arbore de decizie

- Datele de antrenare
- Instanțele au atribute
- Fiecare instanță are atribute cu anumite valori
- Ultimul atribut este clasa

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
1	Soare	Mare	Mare	Absent	Nu
2	Soare	Mare	Mare	Prezent	Nu
3	Înnorat	Mare	Mare	Absent	Da
4	Ploaie	Medie	Mare	Absent	Da
5	Ploaie	Mică	Normală	Absent	Da
6	Ploaie	Mică	Normală	Prezent	Nu
7	Înnorat	Mică	Normală	Prezent	Da
8	Soare	Medie	Mare	Absent	Nu
9	Soare	Mică	Normală	Absent	Da
10	Ploaie	Medie	Normală	Absent	Da
11	Soare	Medie	Normală	Prezent	Da
12	Înnorat	Medie	Mare	Prezent	Da
13	Înnorat	Mare	Normală	Absent	Da
14	Ploaie	Medie	Mare	Prezent	Nu

Exemplu: Construirea unui arbore de decizie Partiţionare după atributul *Starea vremii*

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
1	Soare	Mare	Mare	Absent	Nu
2	Soare	Mare	Mare	Prezent	Nu
3	Înnorat	Mare	Mare	Absent	Da
4	Ploaie	Medie	Mare	Absent	Da
5	Ploaie	Mică	Normală	Absent	Da
6	Ploaie	Mică	Normală	Prezent	Nu
7	Înnorat	Mică	Normală	Prezent	Da
8	Soare	Medie	Mare	Absent	Nu
9	Soare	Mică	Normală	Absent	Da
10	Ploaie	Medie	Normală	Absent	Da
11	Soare	Medie	Normală	Prezent	Da
12	Înnorat	Medie	Mare	Prezent	Da
13	Înnorat	Mare	Normală	Absent	Da
14	Ploaie	Medie	Mare	Prezent	Nu

$$H_S = \frac{5}{14} \cdot 0{,}971 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0{,}971 = 0{,}694.$$

Nr. instanţă	Starea vremii	Joc
1	Soare	Nu
2	Soare	Nu
8	Soare	Nu
9	Soare	Da
11	Soare	Da

$$H_{S_S} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0{,}971.$$

Nr. instanţă	Starea vremii	Joc
3	Înnorat	Da
7	Înnorat	Da
12	Înnorat	Da
13	Înnorat	Da

$$H_{S_I} = 0.$$

Nr. instanţă	Starea vremii	Joc
4	Ploaie	Da
5	Ploaie	Da
6	Ploaie	Nu
10	Ploaie	Da
14	Ploaie	Nu

$$H_{S_P} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0{,}971.$$

Exemplu: Construirea unui arbore de decizie Partiţionare după atributul *Temperatură*

Nr. instanţă	Temperatură	Joc
5	Mică	Da
6	Mică	Nu
7	Mică	Da
9	Mică	Da
4	Medie	Da
8	Medie	Nu
10	Medie	Da
11	Medie	Da
12	Medie	Da
14	Medie	Nu
1	Mare	Nu
2	Mare	Nu
3	Mare	Da
13	Mare	Da

$$H_{T_L} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0{,}811$$

$$H_{T_M} = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0{,}918$$

$$H_{T_H} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.$$

$$H_T = \frac{4}{14}\cdot 0{,}811 + \frac{6}{14}\cdot 0{,}918 + \frac{4}{14}\cdot 1 = 0{,}911.$$

Exemplu: Construirea unui arbore de decizie Partiţionare după atributul *Umiditate*

Nr. instanţă	Umiditate	Joc
5	Normală	Da
6	Normală	Nu
7	Normală	Da
9	Normală	Da
10	Normală	Da
11	Normală	Da
13	Normală	Da
1	Mare	Nu
2	Mare	Nu
3	Mare	Da
4	Mare	Da
8	Mare	Nu
12	Mare	Da
14	Mare	Nu

$$H_{U_N} = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0{,}592$$

$$H_{T_M} = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0{,}985$$

$$\Rightarrow H_U = \frac{7}{14}\cdot 0{,}592 + \frac{7}{14}\cdot 0{,}985 = 0{,}789.$$

Exemplu: Construirea unui arbore de decizie Partiţionare după atributul *Vânt*

Nr. instanţă	Vânt	Joc
1	Absent	Nu
3	Absent	Da
4	Absent	Da
5	Absent	Da
8	Absent	Nu
9	Absent	Da
10	Absent	Da
13	Absent	Da
2	Prezent	Nu
6	Prezent	Nu
7	Prezent	Da
11	Prezent	Da
12	Prezent	Da
14	Prezent	Nu

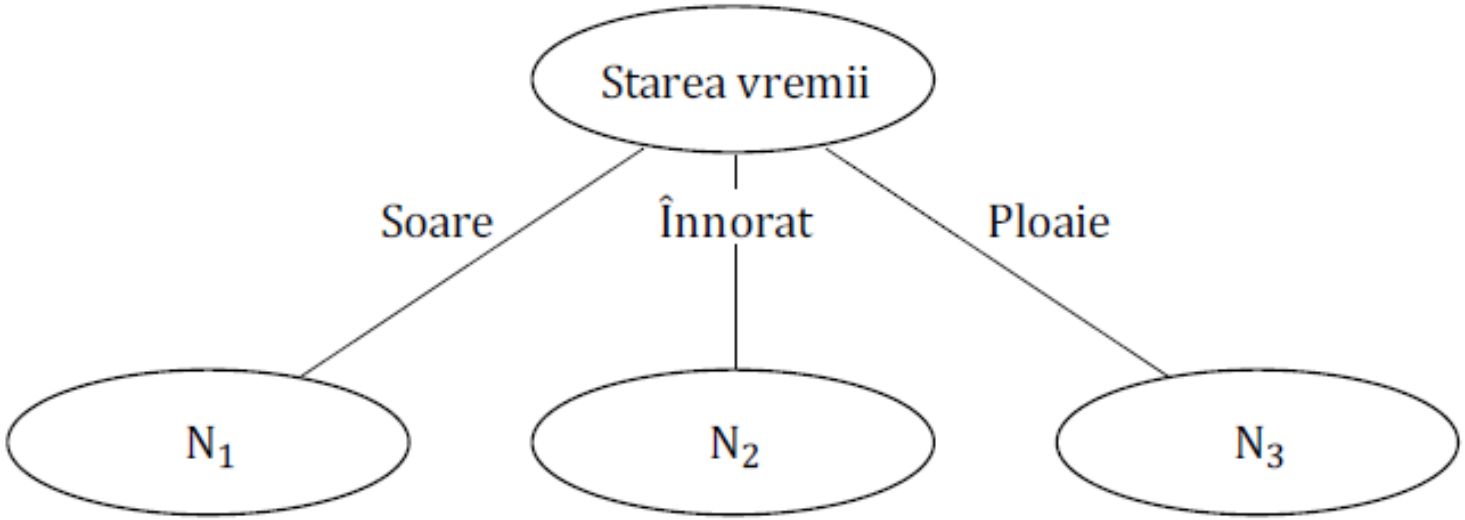
$$H_{V_A} = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0{,}811$$

$$H_{V_P} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

$$\Rightarrow H_V = \frac{8}{14}\cdot 0{,}811 + \frac{6}{14}\cdot 1 = 0{,}892.$$

Exemplu: Construirea unui arbore de decizie Partiţionare

- Valoarea maximă a câştigului informaţional este corespunzătoare minimului entropiei ponderate şi deci prima partiţionare se va face după atributul Starea vremii.



Exemplu: Construirea unui arbore de decizie

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
1	Soare	Mare	Mare	Absent	Nu
2	Soare	Mare	Mare	Prezent	Nu
3	Înnorat	Mare	Mare	Absent	Da
4	Ploaie	Medie	Mare	Absent	Da
5	Ploaie	Mică	Normală	Absent	Da
6	Ploaie	Mică	Normală	Prezent	Nu
7	Înnorat	Mică	Normală	Prezent	Da
8	Soare	Medie	Mare	Absent	Nu
9	Soare	Mică	Normală	Absent	Da
10	Ploaie	Medie	Normală	Absent	Da
11	Soare	Medie	Normală	Prezent	Da
12	Înnorat	Medie	Mare	Prezent	Da
13	Înnorat	Mare	Normală	Absent	Da
14	Ploaie	Medie	Mare	Prezent	Nu

- Pentru nodul *N*1 se repetă procedura, eliminând atributul *Starea vremii* şi păstrând doar instanţele care au ca valoare a acestuia *Soarele* (5 instanţe).

Exemplu: Construirea unui arbore de decizie Partiţionare

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
1	Soare	Mare	Mare	Absent	Nu
2	Soare	Mare	Mare	Prezent	Nu
3	Înnorat	Mare	Mare	Absent	Da
4	Ploaie	Medie	Mare	Absent	Da
5	Ploaie	Mică	Normală	Absent	Da
6	Ploaie	Mică	Normală	Prezent	Nu
7	Înnorat	Mică	Normală	Prezent	Da
8	Soare	Medie	Mare	Absent	Nu
9	Soare	Mică	Normală	Absent	Da
10	Ploaie	Medie	Normală	Absent	Da
11	Soare	Medie	Normală	Prezent	Da
12	Înnorat	Medie	Mare	Prezent	Da
13	Înnorat	Mare	Normală	Absent	Da
14	Ploaie	Medie	Mare	Prezent	Nu

$H_{T_L} = 0$ (1 instanță în clasa *Da*)

$H_{T_M} = 1$ (1 instanță în clasa *Da* şi 1 instanță în clasa *Nu*)

$H_{T_H} = 0$ (2 instanțe în clasa *Nu*)

$$\Rightarrow H_T = \frac{1}{5} \cdot 0 + \frac{2}{5} \cdot 1 + \frac{2}{5} \cdot 0 = 0{,}4.$$

$H_{U_N} = 0$

$H_{U_M} = 0$

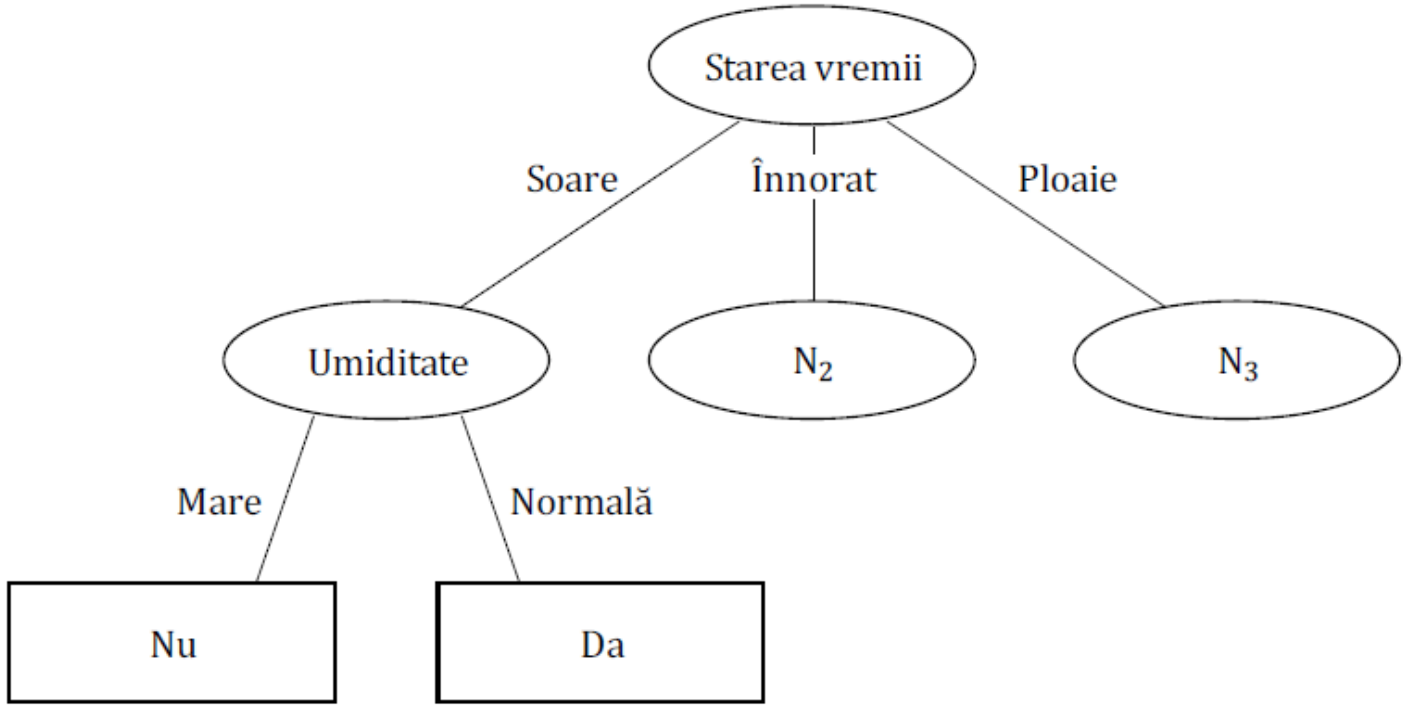
$\Rightarrow H_U = 0.$

$$H_{V_A} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0{,}918$$

$H_{V_P} = 1$

$$\Rightarrow H_V = \frac{3}{5} \cdot 0{,}918 + \frac{2}{5} \cdot 1 = 0{,}951.$$

Exemplu: Construirea unui arbore de decizie Partiţionare



Exemplu: Construirea unui arbore de decizie Partiţionare

Pentru nodul N2, avem următoarea mulţime de date

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
3	Înnorat	Mare	Mare	Absent	Da
7	Înnorat	Mică	Normală	Prezent	Da
12	Înnorat	Medie	Mare	Prezent	Da
13	Înnorat	Mare	Normală	Absent	Da

Nodul este omogen şi deci va fi la rândul său frunză, fără a mai trebui partiţionat

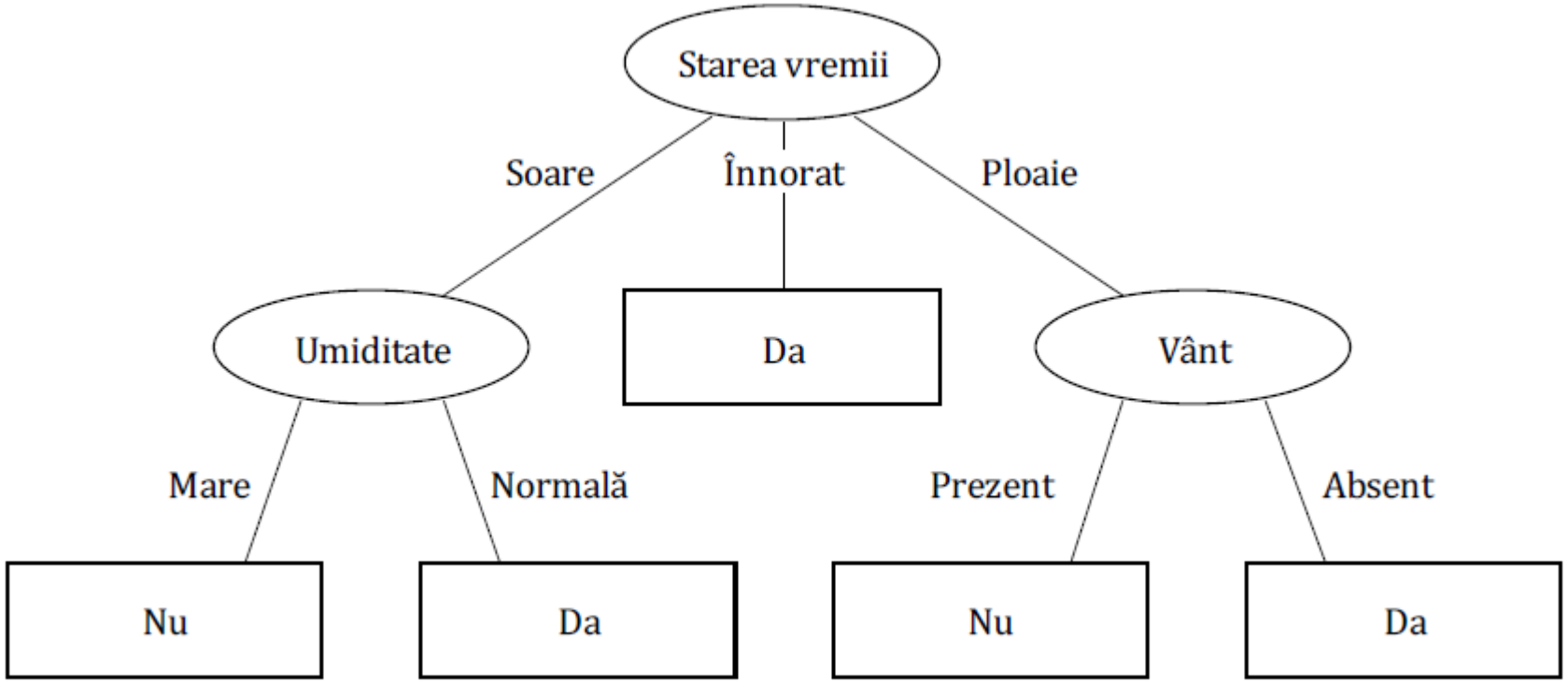
Exemplu: Construirea unui arbore de decizie Partiţionare

Pentru nodul N3, avem următoarea mulţime de date

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
4	Ploaie	Medie	Mare	Absent	Da
5	Ploaie	Mică	Normală	Absent	Da
6	Ploaie	Mică	Normală	Prezent	Nu
10	Ploaie	Medie	Normală	Absent	Da
14	Ploaie	Medie	Mare	Prezent	Nu

Aplicând aceeaşi procedură, vom obține: $H_T = 0{,}951$, $H_U = 0{,}951$ şi $H_V = 0$. Ultima valoare este minimă şi deci vom partiționa nodul N_3 după atributul *Vânt*, rezultând de asemenea două frunze.

Exemplu: Construirea unui arbore de decizie Arborele final



Temperatura este un atribut irelevant pentru această clasificare.

Bibliografie

- **Florin Leon** (2012). **Inteligenta artificiala**: rationament probabilistic, **tehnici de clasificare** Tehnopress, Iasi, ISBN 978-973-702-932-4, Capitolul 6 şi Capitolul 7