Arbori de decizie ID3

Arborii de decizie sunt metode de clasificare ce presupun organizarea ierarhica a atributelor (caracteristicilor) unor date dintr-un anumit domeniu, la baza ierarhiei aflându-se posibilele decizii (clasele) care rezultă din parcurgerea acelor atribute. Arborii ID3 (Iterative Dichotomizer) ordonează ierarhic caracteristicile datelor în funcție de câștigul informațional pe care îl conferă luarea unei decizii în baza acelor caracteristici.

Arborele are următoarea structură:

- Rădăcina și nodurile intermediare conțin atributele, în ordinea descendentă a importanței lor decizionale.
- Muchiile reprezintă valorile (sau grupurile de valori ale) atributelor.
- Fiecare nod are un număr de descendenți egal cu numărul deciziilor posibile pornind de la acel nod. În cel mai simplu caz, pentru atribute cu valori ordinale, un nod are câte un descendent pentru fiecare valoare a atributelor sale.
- Nodurile frunză conțin valorile claselor.

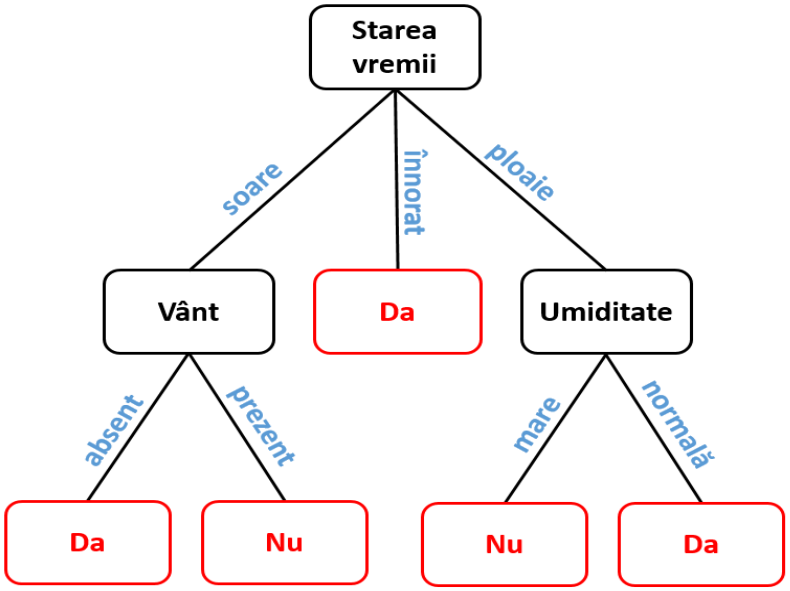
Clasificarea folosind arbori de decizie presupune:

- Generarea arborelui pornind de la un set de date de antrenare. Acest set de date conține o mulțime de instanțe ale căror clase se cunosc deja.
- Parcurgerea arborelui pentru o nouă instanță. Această parcurgere presupune divizarea repetată a soluțiilor posibile funcție de valorile atributelor instanțelor, până când se ajunge la un nod frunză care conține valoarea clasei ce i se atribuie instanței.

Fie următorul set de date de antrenare:

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

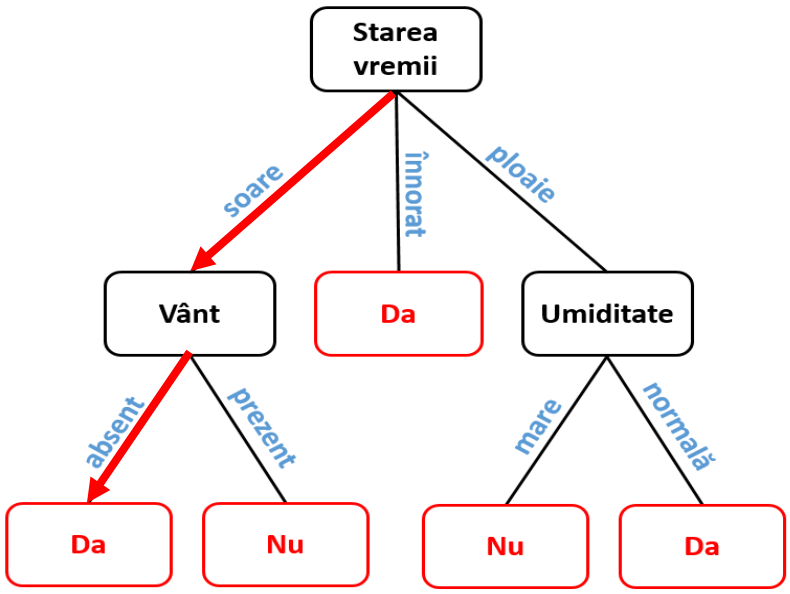
Arborele ID3 care se generează pe baza acestor date este:



Presupunem că dorim să stabilim clasa următoarei instanțe:

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
15	Soare	Mare	Normală	Absent	???

Pentru aceasta, parcurgem arborele în funcție de valorile atributelor instanței, până când întâlnim un nod frunză, care conține o clasă:



Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
15	Soare	Mare	Normală	Absent	Da

Pentru generarea arborelui trebuie stabilită o ordine a atributelor (care este atributul cel mai important, funcție de care se ia prima decizie? În exemplul de mai sus, de ce atributul din nodul rădăcină este “Starea vremii”?). Pentru a determina gradul de importanță, pentru fiecare atribut se calculează *câștigul informațional* (IG – Information Gain) care rezultă din luarea deciziei pornind de la acel atribut. Pentru arborele complet și pentru oricare subarbore al său, nodul rădăcină va conține atributul care conferă cel mai mare IG.
În cazul arborilor ID3, IG se determină ca fiind măsura în care se reduce *entropia* mulțimii datelor de antrenare în urma împărțirii acestora pe baza valorilor atributelor.

Pentru o mulțime de date *S*, entropia se calculează conform cu Eq. 1.

$$\mathrm{H}(S) = \sum_{x \in X} -p(x) \log_2 p(x) \qquad (1)$$

unde:
X este mulțimea claselor - în exemplul de mai sus este mulțimea {Da, Nu}
x este o clasă din mulțimea *X*
p(x) este probabilitatea de apariție a clasei x în setul de date S
Entropia H(S) ia valori din [0, 1]. Valorile mari ale entropiei semnifică faptul că în S există un grad mare de incertitudine. Incertitudinea maximă înseamnă că cele două clase sunt prezente în mod egal în setul de date (50% din instanțe au clasa Da, 50% din instanțe au clasa Nu), caz în care H(S) = 1. La extrema cealaltă, toate clasele sunt fie Da, fie Nu, situație în care H(S) = 0. Câștigul informațional IG al unui atribut se referă la măsura în care scade entropia setului de date atunci când acesta este partajat de valorile acelui atribut.
De exemplu, atributul “Starea vremii” partiționează datele astfel:

Starea vremii = “Soare”:

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

Starea vremii = “Ploaie”:

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

Starea vremii = “Înnorat”:

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

În urma acestei partiționări, entropia setului de date se determină ca fiind suma ponderată a entropiilor individuale ale partițiilor (Eq. 2).

$$\mathrm{H}(S|A) = \sum_{t \in T} p(t)\mathrm{H}(t) \tag{2}$$

unde:

H(*S*|*A*) = entropia sistemului după partiționarea folosind valorile atributului A

T = mulțimea partițiilor reallizate folosind valorile atributului A

t = o partiție din mulțimea T

p(t) = probabilitatea valorii / valorilor atributului A care au generat partiția t

H(*t*) = entropia partiției t

Câștigul informațional care rezultă în urma partiționării se determină ca fiind diferența dintre entropiile setului de date înainte și după partiționare (Eq. 3).

$$IG(S, A) = \mathrm{H}(S) - \mathrm{H}(S|A) \tag{3}$$

Pentru arborele de decizie și pentru oricare subarbore al său, dorim ca nodul rădăcină să corespundă cu atributul care oferă cel mai mare IG. Întrucât entropia întregului set de date H(S) este aceeași indiferent de modalitatea de partiționare, este suficient să determinăm atributul A pentru care H(S|A) este minimă. Atfel spus, *atributul din nodul rădăcină este cel care minimizează entropia indusă în setul de date în urma partiționării pe baza valorilor acelui atribut*.

Pentru setul de date din exemplu, calculăm entropia astfel (Eq. 1):

Joc	
Da	**Nu**
9	**5**

$H(Joc) = - P_{(Joc = Da)} \log_2(P_{(Joc = Da)}) - P_{(Joc = Nu)} \log_2(P_{(Joc = Nu)}) =$
$= - 9/14 \log_2(9/14) - 5/14 \log_2(5/14) =$ **0.94**

Pentru un atributul *Starea vremii*, calculul se face astfel (Eq. 2):

		Joc	
		Da	**Nu**
Starea vremii	**Soare**	3	2
	Înnorat	4	0
	Ploaie	2	3

H(Joc, Starea vremii) = P(Soare) * H(Soare) + P(Înnorat) * H(Înnorat) + P(Ploaie) * H(Ploaie)

P(Soare) = 5/14
$H(Soare) = - 3/5 \log_2(3/5) - 2/5 \log_2(2/5) = 0.971$
P(Înnorat) = 4/14
$H(Înnorat) = - 4/4 \log_2(4/4) - 0 = 0$
P(Ploaie) = 5/14
$H(Ploaie) = - 2/5 \log_2(2/5) - 3/5 \log_2(3/5) = 0.971$

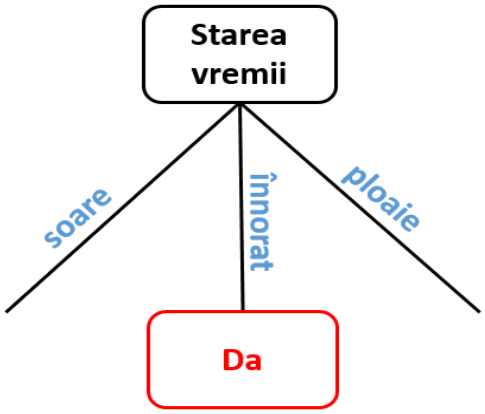
H(Joc, Starea Vremii) = 5/14 * 0.971 + 4/14 * 0 + 5/14 * 0.971 = **0.693**

În mod similar, determinăm entropia celorlalte atribute și obținem:

		Joc	
		Da	**Nu**
Starea vremii	**Soare**	3	2
	Înnorat	4	0
	Ploaie	2	3
	H = 0.693		
Temperatură	**Mare**	2	2
	Medie	4	2
	Mică	3	1
	H = 0.907		
Umiditate	**Mare**	3	4
	Normală	6	1
	H = 0.789		
Vânt	**Absent**	6	2
	Prezent	3	3
	H = 0.892		

Observăm faptul că entopia minimă rezultă prin partiționarea folosind atributul *Starea vremii*, așadar prima decizie se va lua funcție de valorile acestuia (acest atribut va constitui rădăcina arborelui).

Pornind de la rădăcina stabilită anterior, se generează trei ramuri ce corespund celor trei valori ale atributului (Soare, Ploaie, Înnorat). Observăm că partiția corespunzătoare valorii *Înnorat* are entropia = 0, ceea ce este în acord cu faptul că toate instanțele din acea partiție au aceeași clasă (Joc = Da). Prin urmare, nodul cu care se încheie această ramură va fi o frunză cu valoarea clasei Da.

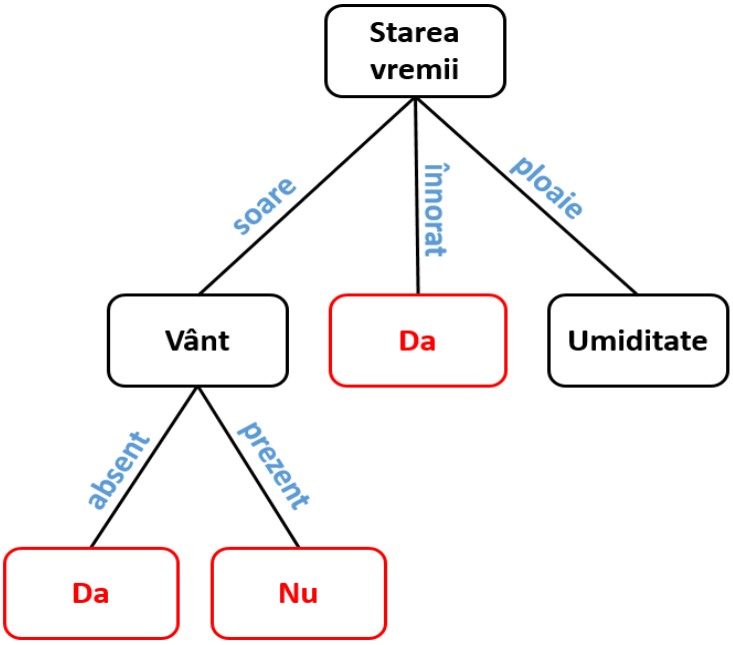


Celelalte ramuri au entropie nenulă, așadar nodurile de la capetele lor vor fi intermediare (se vor subdiviza la rândul lor).

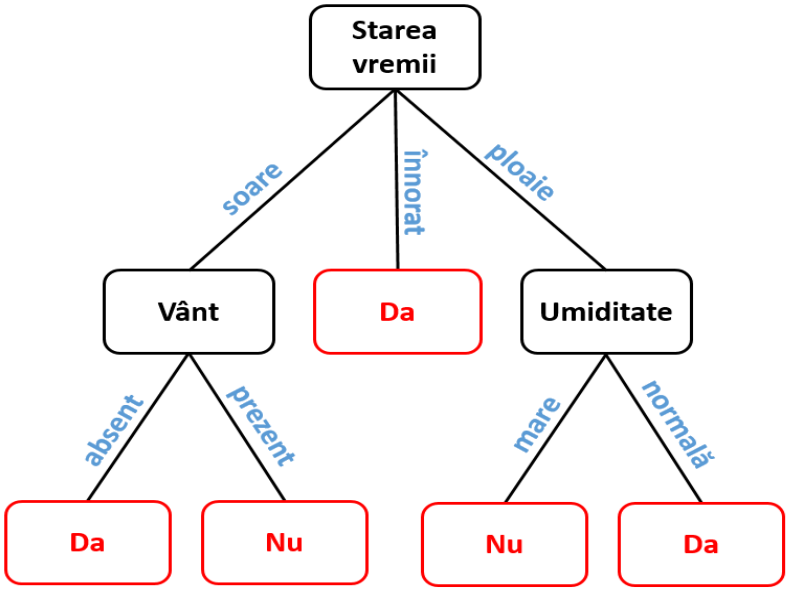
Pentru ramura *Soare*, setul de date devine:

Nr. instanţă	Starea vremii	Temperatură	Umiditate	Vânt	Joc
1	Soare	Medie	Mare	Absent	Da
2	Soare	Mică	Normală	Absent	Da
3	Soare	Mică	Normală	Prezent	Nu
4	Soare	Medie	Normală	Absent	Da
5	Soare	Medie	Mare	Prezent	Nu

Efectuând aceleași calcule pentru cele trei atribute rămase (atributul *Starea vremii* a fost deja prelucrat), rezultă faptul că entropia minimă corespunde atributului *Vânt*. Entropia este nulă în cazul valorilor *Absent* și *Prezent*, ceea ce înseamnă că ramurile corespunzătoare lor se vor încheia cu noduri frunză. Acest lucru se observă și din tabel, unde toate instanțele cu *Vânt* = *Absent* se încadrează în clasa *Da*, iar cele cu Vânt = Prezent corespund clasei *Nu*.



În cazul ramurii Ploaie, se constată o situație asemănătoare pentru atributul Umiditate, iar subarborele corespunzător are o structură similară. Rezultă astfel arborele complet:



Cerințe:

1) Generați un arbore de decizie ID3 pornind de la setul de date din fișierul *data_vreme3.csv*:

Starea vremii	Temperatura	Umiditate	Vant	Joc
Ploaie	Mare	Mare	Absent	Nu
Ploaie	Mare	Mare	Prezent	Nu
Innorat	Mare	Mare	Absent	Da
Soare	Medie	Mare	Absent	Da
Soare	Mica	Normala	Absent	Da
Soare	Mica	Normala	Prezent	Nu
Innorat	Mica	Normala	Prezent	Da
Ploaie	Medie	Mare	Absent	Nu
Ploaie	Mica	Normala	Absent	Da
Soare	Medie	Normala	Absent	Da
Ploaie	Medie	Normala	Prezent	Da
Innorat	Medie	Mare	Prezent	Da
Innorat	Mare	Normala	Absent	Da
Soare	Medie	Mare	Prezent	Nu

O variantă de pseudocod pentru generarea arborelui:

```
ID3(Dataset, Lista_Atribute)
  - Generează un nod rădăcină
  - Dacă toate instanțele din setul de date aparțin aceleiași
    clase, returnează un arbore cu un singur nod, rădăcina, cu
    valoarea clasei
  - Dacă lista de atribute este goală, returnează un arbore cu un
    singur nod, rădăcina, cu valoarea clasei celei mai frecvente
  - Determină atributul cu entropia minimă (notat A)
  - Atributul nodului rădăcină este A
  - Pentru fiecare valoare v a atributului A
      o Adaugă o nouă ramură pentru rădăcină, corespunzătoare lui
        A = v
      o Determină submulțimea Dataset(v), care conține instanțele
        cu valoarea v a atributului A
      o Dacă Dataset(v):
          ▪ Nu conține nici o instanță, atunci ramura se încheie
            cu un nod frunză, cu valoarea celei mai frecvente
            clase din Dataset
          ▪ Conține instanțe, atunci ramura se continuă cu
            subarborele ID3(Dataset(v), Lista_Atribute – A)
```

Ca sugestie de implementare, se poate folosi codul sursă care însoțește laboratorul. La afișarea arborelui, se obține:

```
Starea vremii
              --Innorat--> Da
              --Ploaie--> Umiditate
                                  --Normala--> Da
                                  --Mare--> Nu
              --Soare--> Vant
                             --Absent--> Da
                             --Prezent--> Nu
```

2) Folosiți arborele pentru a clasifica instanțele din fișierul *data_vreme4.csv*. A clasifica o instanță înseamnă să parcurgem recursiv arborele pornind de la nodul rădăcină pe baza valorilor atributelor instanței, până când se ajunge într-un nod frunză ce conține valoarea unei clase.

Determinați eroarea de clasificare a arborelui pentru aceste instanțe.
Eroare de clasificare = nr. Instanțelor clasificate greșit / nr. tuturor instanțelor

Setul de date inițial, în care clasele sunt deja cunoscute:

Starea vremii	Temperatura	Umiditate	Vant	Joc
Soare	Mare	Normala	Absent	Da
Soare	Mare	Normala	Prezent	Nu
Ploaie	Medie	Mare	Prezent	Nu
Ploaie	Mare	Normala	Absent	Nu
Innorat	Mica	Mare	Prezent	Da

Folosind arborele ID3 generat anterior, aplicat individual pe fiecare instanță, rezultă clasele din ultima coloană:

Starea vremii	Temperatura	Umiditate	Vant	Joc	Joc – ID3
Soare	Mare	Normala	Absent	Da	Da
Soare	Mare	Normala	Prezent	Nu	Nu
Ploaie	Medie	Mare	Prezent	Nu	Nu
Ploaie	Mare	Normala	Absent	Nu	Da
Innorat	Mica	Mare	Prezent	Da	Da

Observăm faptul că, din cele 5 instanțe, una este clasificată greșit de către arborele ID3. Prin urmare, eroarea de clasificare este 1/5 = 0.2