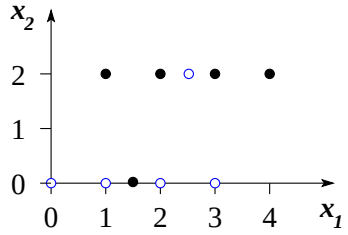
Învăţare automată

— Licenţă, anul III, 2021-2022, testul 1 —

Nume student: **Grupa:**

1. **(Extensii ale algoritmului ID3: variabile de intrare continue; "decision stumps"; eroarea la antrenare, eroarea la CVLOO; overfitting)**

Fie set de date de antrenament din figura alăturată. X_1 şi X_2 sunt atribute de intrare luând valori reale / continue, iar Y este variabilă de ieşire cu valori booleene. Pe acest set de date se foloseşte algoritmul ID3 pentru învăţare de arbori de decizie.



Notăm cu DT* arborele învăţat de algoritmul ID3, însă fără pruning, iar DT2 arborele de decizie care este obţinut din el prin pruning, are doar două noduri frunză (deci face o singură divizare de interval) şi în fiecare dintre aceste noduri-frunză ia decizia majoritară.

a. Care este eroarea produsă de DT* la antrenare?
(*Atenţie*! Nu este nevoie să aplicaţi efectiv algoritmul ID3. Este suficient să identificaţi zonele de decizie corespunzătoare arborelui ID3.)

b. Care este eroarea produsă de DT* la CVLOO (cross-validare cu metoda Leave-One-Out)?
În mod specific, veţi identifica instanţele de antrenament care produc eroare la CVLOO. Pentru fiecare dintre aceste cazuri veţi trasa pe câte un nou sistem de coordonate zonele de decizie corespunzătoare cazului respectiv.

c. Care este eroarea produsă de DT2 la antrenare?
În prealabil, veţi desena arborele DT2 şi veţi justifica în mod riguros cum a fost ales testul din nodul rădăcină al acestui arbore.

d. Care [credeţi că] este eroarea produsă de DT* la cross-validare cu metoda Leave-One-Out?
(În locul calculelor — care nu pot fi făcute efectiv în intervalul de timp limitat de care dispuneţi acum — vă cerem să daţi o *justificare informală*!)

e. Comparând rezultatele de la punctele b-d, este oare cazul să spunem că se produce overfitting? (Veţi da în prealabil *definiţia* riguroasă a overfitting-ului, cf. cărţii *Machine Learning* de Tom Mitchell.)

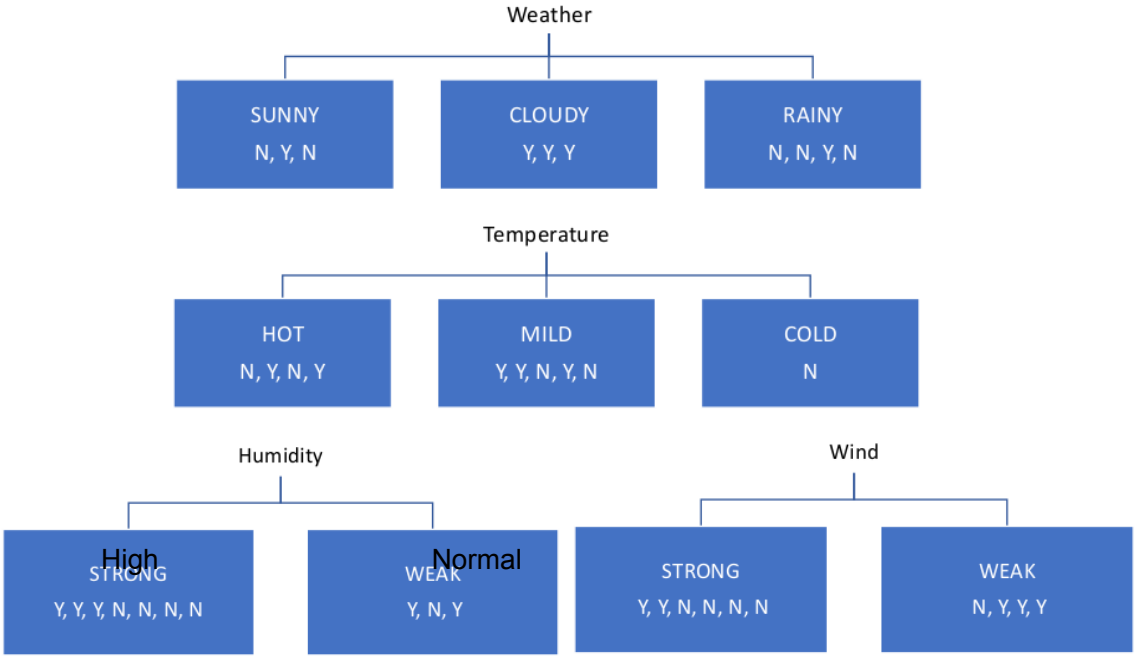
2. **(Algoritmul ID3: aplicare)**

Fie următorul set de date, relativ la când anume un copil iese afară să se joace:

Day	Weather	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

În acest exerciţiu vă cerem să elaboraţi arborele învăţat de algoritmul ID3 pe acest set de date.

a. Pentru a determina atributul de intrare corespunzător nivelului rădăcină, am alcătuit [eu în locul dumneavoastră] compaşii de decizie, sub o formă uşor diferită faţă de cea pe care am folosit-o la curs [şi în culegerea de exerciţii]:



Folosind aceşti compaşi de decizie, stabiliţi în mod riguros ce atribut va pune algoritmul ID3 în nodul rădăcină.

***Sugestie*: Puteţi folosi următorul tabel, în care se dau entropiile mai multor variabile Bernoulli, determinate de valoarea parametrului p:**

p	1/4	1/3	2/5	3/7
$H(p)$	0.8112	0.9182	0.9709	0.9852

b. Stabiliţi ce atribute pot fi puse în nodurile de test de pe nivelul 2 (adică, în descendenţii nodului rădăcină), iar apoi completaţi arborele ID3.[1] Este oare arborele [elaborat de algoritmul] ID3 unic? Care este eroarea pe setul de date de antrenare?

[1] *Atenţie*! Puteţi elabora raţionamente perfecte fără să faceţi calcule laborioase!
Explorând cu atenţie datele de antrenament, veţi vedea că puteţi să identificaţi atribute cu putere de predicţie / „discriminare“ maximă [a valorilor variabilei de ieşire]. Cât este entropia condiţională medie a acestor atribute [în raport cu variabila de ieşire]?