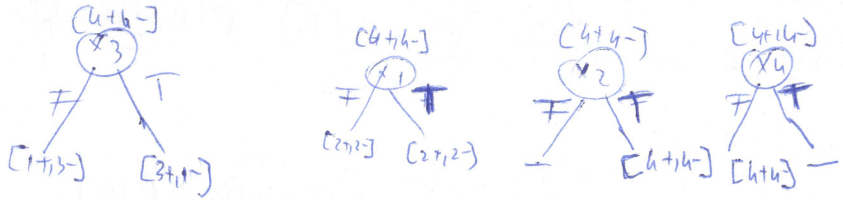
$\log_2 3 = \log_{10} 3 - \log_{10} 2$

$= 0{,}477 - 0{,}30$

$0{,}30$



2. **(Câştigul de informaţie — determinarea celui mai „bun" atribut; arbori de decizie consistenţi cu un set de date)**

Eticheta instanţei	X_1	X_2	X_3	X_4
1	T	T	T	F
1	T	T	T	F
1	F	T	T	F
1	F	T	F	F
0	T	T	F	F
0	T	T	F	F
0	F	T	F	F
0	F	T	T	F

Fie atributele binare X_1, X_2, X_3, X_4 (valorile lor pot fi doar T sau F), precum şi două tipuri de etichete, 0 şi 1. Veţi considera cele 8 instanţe din tabelul alăturat.

a. Vrem să învăţăm un arbore de decizie din acest set de exemple. În vederea selectării celui mai bun candidat pentru nodul rădăcină, calculaţi câştigul de informaţie pentru fiecare atribut X_i, cu $i = 1, \ldots, 4$. Ce atribut veţi selecta?

b. Există oare un arbore de decizie care poate clasifica în mod perfect instanţele date? În cazul afirmativ, desenaţi acel arbore de decizie. În caz contrar, daţi o explicaţie simplă.

a) Pt - calc $IG(Eticheta, X_i)$, ne folosim

$H(E) = H[4+,4-] = \frac{4}{8}\log_2 2 + \frac{4}{8}\log_2 2 = 1$ (cantg de inf e 0) ✓

$IG(E, X_1) = H[4+,4-] - \left(\frac{1}{2}H[2+,2-] + \frac{1}{2}H[2+,2-]\right) = 1 - \left(\frac{1}{2} + \frac{1}{2}\right) = 0$ ✓

$IG(E, X_2) = H[4+,4-] - \left(0 + 1 \cdot H[4+,4-]\right) = 1 - 1 = 0$ ✓

$IG(E, X_3) = H[4+,4-] - \left(\frac{1}{2}H[1+,3-] + \frac{1}{2}H[3+,1-]\right) =$

$= 1 - \left(\frac{1}{2}\left(\frac{1}{4}\log_2 4 + \frac{3}{4}\log_2 \frac{4}{3}\right) + \frac{1}{2}\left(\frac{3}{4}\log\frac{4}{3} + \frac{1}{4}\log_2 4\right)\right) =$

$= 1 - \left(\frac{1}{4}\log_2 4 + \frac{3}{4}\log_2 \frac{4}{3}\right) = 1 - \left(\frac{1}{2} + \frac{3}{4}\left(\log_2 4 - \log_2 3\right)\right) =$

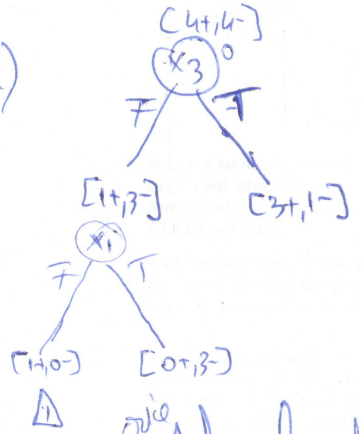
$= 1 - \left(\frac{1}{2} + \frac{3}{4}(2 - 1{,}57)\right) = 1 - \left(\frac{1}{2} + \frac{3}{4} \cdot 0{,}42\right) =$

$= 1 - (0{,}5 + 0{,}315) = 0{,}185$ ✓

$IG(E, X_4) = H[4+,4-] - (1 \cdot H[4+,4-] + 0) = 0$ ✓

Alegem X_3 ca rădăcină ✓

b)



⚠

— Arborele de decizie nu poate clasifica perfect datele de antrenament deoarece datele nu sunt consistente adică:

E	X_1	X_2	X_3	X_4	
0	F	T	F	F	} măcar una
1	F	T	F	F	}

din instanță va fi etichetată greșit

...