

4. Sailkapen-zuhaitzak. Ariketak

1. Tenisean jokatzeko zaletasuna duen kirolari batek 14 egunetako eguraldi-informazioa eta egun horietako bakoitzean hartutako erabakia (jokatu? bai ala ez) gorde ditu honako datubasean:

Eguna	X_1 (Zerua)	X_2 (Tenperatura)	X_3 (Hezetasuna)	X_4 (Haizea)	C Jokatu?
1	garbia	beroa	altua	ahula	ez
2	garbia	beroa	altua	zakarra	ez
3	hodeitsu	beroa	altua	ahula	bai
4	euritsu	epela	altua	ahula	bai
5	euritsu	hotza	normala	ahula	bai
6	euritsu	hotza	normala	zakarra	ez
7	hodeitsu	hotza	normala	zakarra	bai
8	garbia	epela	altua	ahula	ez
9	garbia	hotza	normala	ahula	bai
10	euritsu	epela	normala	ahula	bai
11	garbia	epela	normala	zakarra	bai
12	hodeitsu	epela	altua	zakarra	bai
13	hodeitsu	beroa	normala	ahula	bai
14	euritsu	epela	altua	zakarra	ez

Eguraldiari buruzko informazioa gordetzeko erabili dituen lau aldagai iragarleek honako informazioa ematen dute: zeruaren egoera (X_1), tenperatura (X_2), hezetasuna (X_3) eta haizea (X_4).

Kalkula ezazu *ID3* algoritmoa aplikatuz lortzen den sailkapen-zuhaitza.

2. Izan bedi 20 kasu dituen datubasea, non 4 aldagai iragarle eta C klase-aldagai hauek ditugun:

- X_1 : “Analisi Matematikoa”
- X_2 : “Kalkulua”
- X_3 : “Ikerketa Operatiboa”
- X_4 : “Probabilitatea eta Estatistika”
- C : “Datu-Meatzaritza” irakasgaia lehen deialdian gaindituko den, bai ala ez

X_i , $i = 1, \dots, 4$ aldagai iragarleak 1 balioa badu, orduan irakasgai hori lehen deialdian gainditua izan zela esan nahi du, eta 2 balioa badu, orduan deialdi bat baino gehiago behar izan zela irakasgaia gainditzeko.

Kalkula ezazu *ID3* algoritmoa erabiliz lortutako sailkapen-zuhaitza.

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	2	1	1	1
$\mathbf{x}_2$	2	2	1	1	1
$\mathbf{x}_3$	1	2	1	1	1
$\mathbf{x}_4$	2	2	1	1	1
$\mathbf{x}_5$	1	1	1	1	1
$\mathbf{x}_6$	2	2	1	1	1
$\mathbf{x}_7$	1	1	1	1	1
$\mathbf{x}_8$	2	2	1	1	1
$\mathbf{x}_9$	1	1	2	2	1
$\mathbf{x}_{10}$	2	1	2	2	1
$\mathbf{x}_{11}$	1	2	1	2	2
$\mathbf{x}_{12}$	2	2	1	2	2
$\mathbf{x}_{13}$	1	2	2	2	2
$\mathbf{x}_{14}$	2	2	2	2	2
$\mathbf{x}_{15}$	1	2	1	2	2
$\mathbf{x}_{16}$	2	2	2	2	2
$\mathbf{x}_{17}$	1	2	1	2	2
$\mathbf{x}_{18}$	2	1	2	1	2
$\mathbf{x}_{19}$	1	1	2	1	2
$\mathbf{x}_{20}$	2	2	2	1	2

3. 12 kasu dituen honako datubasetik abiatuz $ID3$ algoritmoa erabil ezazu sailkapen-zuhaitza eraikitzeko.

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	2	1	2	1
$\mathbf{x}_2$	2	2	1	1	1
$\mathbf{x}_3$	1	1	1	2	1
$\mathbf{x}_4$	2	1	2	1	1
$\mathbf{x}_5$	1	2	2	3	2
$\mathbf{x}_6$	1	2	2	3	2
$\mathbf{x}_7$	2	2	2	2	2
$\mathbf{x}_8$	3	2	3	2	2
$\mathbf{x}_9$	3	2	3	2	3
$\mathbf{x}_{10}$	2	2	3	2	3
$\mathbf{x}_{11}$	3	1	2	1	3
$\mathbf{x}_{12}$	3	2	2	1	3

4. Izan bedi 20 adibide dituen datubase hau:

Adibideak	X_1	X_2	C
$\mathbf{x}_1$	1	1	1
$\mathbf{x}_2$	1	1	1
$\mathbf{x}_3$	2	1	1
$\mathbf{x}_4$	2	1	1
$\mathbf{x}_5$	1	1	1
$\mathbf{x}_6$	3	1	1
$\mathbf{x}_7$	3	1	1
$\mathbf{x}_8$	1	1	1
$\mathbf{x}_9$	2	2	1
$\mathbf{x}_{10}$	2	2	1
$\mathbf{x}_{11}$	2	2	2
$\mathbf{x}_{12}$	4	2	2
$\mathbf{x}_{13}$	4	2	2
$\mathbf{x}_{14}$	2	2	2
$\mathbf{x}_{15}$	1	2	2
$\mathbf{x}_{16}$	2	2	2
$\mathbf{x}_{17}$	1	2	2
$\mathbf{x}_{18}$	2	1	2
$\mathbf{x}_{19}$	2	1	2
$\mathbf{x}_{20}$	2	1	2

Erabaki ezazu $ID3$ algoritmoak sailkapen-zuhaitzaren errorako X_1 edo X_2 aldagai iragarlea aukeratuko duen (informazio-irabazian oinarrituta). Har ezazu erabaki berbera $C4.5$ algoritmoa erabiliz (irabazi-ratioan oinarrituta).

5. 10 kasu dituen honako datubasetik abiatuz, kalkula ezazu $ID3$ algoritmoa erabiliz lortzen den sailkapen-zuhaitza. Kalkula ezazu baita $C4.5$ algoritmoa erabiliz lortzen dena.

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	1	1	1	1
$\mathbf{x}_2$	1	2	1	1	1
$\mathbf{x}_3$	2	3	1	2	1
$\mathbf{x}_4$	4	1	1	1	1
$\mathbf{x}_5$	4	2	2	1	1
$\mathbf{x}_6$	3	3	2	1	2
$\mathbf{x}_7$	3	1	2	1	2
$\mathbf{x}_8$	3	2	2	1	2
$\mathbf{x}_9$	2	3	2	2	2
$\mathbf{x}_{10}$	2	1	2	1	2

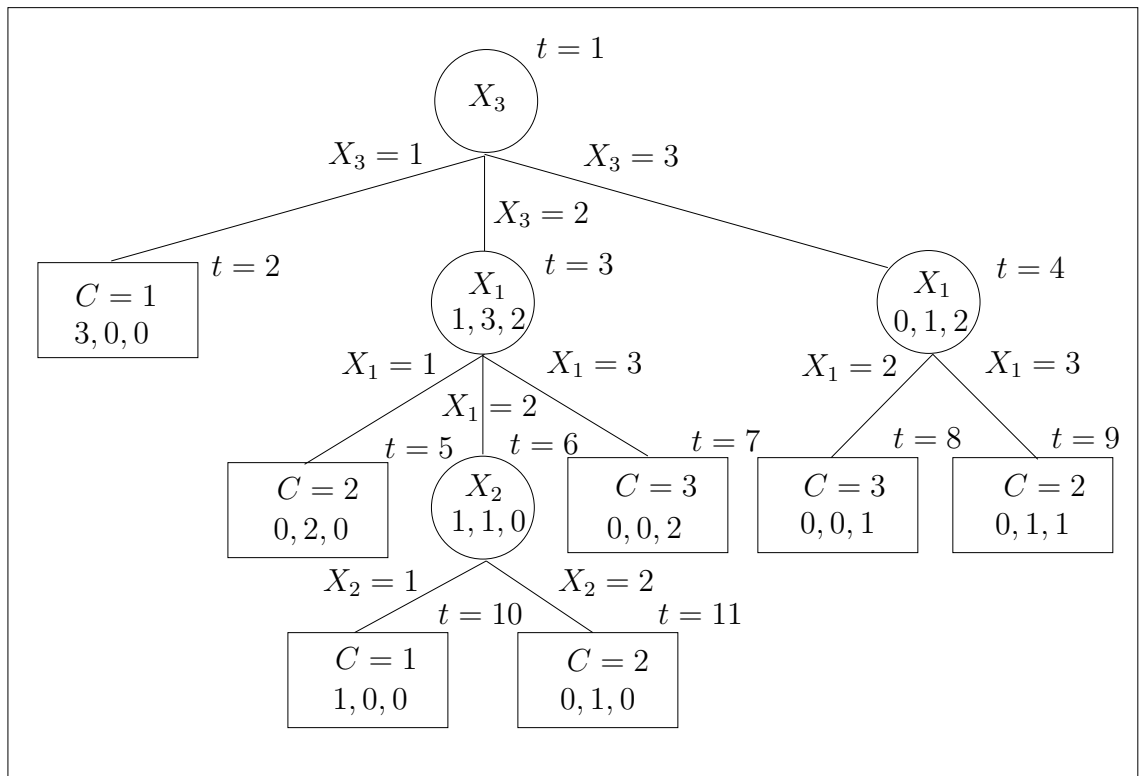
Sortu berri dituzun dituzun bi sailkapen-zuhaitz horietan oinarrituz, eman ezazu honako hiru kasu berrientzako C klasearen iragarpena:

$$\mathbf{x} = (3, 2, 1, 1), \quad \mathbf{x} = (1, 1, 2, 2), \quad \mathbf{x} = (2, 2, 2, 1)$$

6. Izan bedi 12 kasu dituen honako datubasea.

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	2	1	2	1
$\mathbf{x}_2$	2	2	1	1	1
$\mathbf{x}_3$	1	1	1	2	1
$\mathbf{x}_4$	2	1	2	1	1
$\mathbf{x}_5$	1	2	2	3	2
$\mathbf{x}_6$	1	2	2	3	2
$\mathbf{x}_7$	2	2	2	2	2
$\mathbf{x}_8$	3	2	3	2	2
$\mathbf{x}_9$	3	2	3	2	3
$\mathbf{x}_{10}$	2	2	3	2	3
$\mathbf{x}_{11}$	3	1	2	1	3
$\mathbf{x}_{12}$	3	2	2	1	3

Demagun dagokion sailkapen-zuhaitza eraiki dugula.



Zuhaitzaren nodoetan adierazten da klase bakoitzeko zenbat kasu pasako den nodo horretatik. $C4.5$ algoritmoaren inausketa prozesua aplika ezazu hostorik sakonenetatik hasita. Nola geratuko da sailkapen-zuhaitza inausketa-prozesua amaitu eta gero?