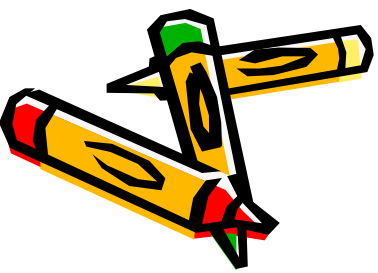
Clasificare

Marina Gorunescu
mgorun@inf.ucv.ro

Despre clasificare



clasificare

Prezentăm câteva abordări mai cunoscute a termenului de **clasificare**.

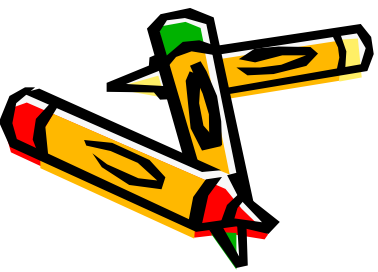
Clasificarea taxonomică reprezintă procesul plasării unui obiect sau concept într-o anumită categorie, pe baza proprietăților (caracteristicilor) acelui obiect.

Taxonomiştii sunt preocupați să construiască clasificări care însumează relații între unitățile taxonomice de diferite tipuri.

Aceste unități sunt incluse în clase care sunt disjuncte şi dispuse ierarhic, cum este spre exemplu sistemul lui Linne.

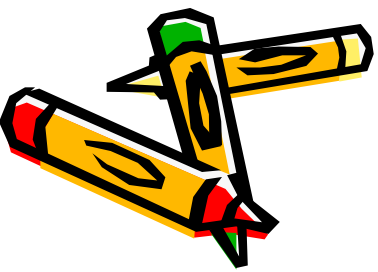
Clasificarea ştiinţifică (cunoscută şi ca *taxonomia ştiinţifică*), s-a bazat pe gruparea organismelor în diferite specii.

Clasificarea modernă are ca părinte pe omul de ştiința suedez **Carl von Linne** (Carolus Linnaeus, 1707-1778), care a grupat speciile în funcție de caracteristicile lor fizice.



probleme

- Arheologii sunt interesați de a găsi similarități în artefactele, cum ar fi ornamente sau unelte de piatră, găsite prin excavații, ceea ce le-ar permite să studieze distribuția spațială a tipurilor de artefacte.
(Hodson, Sneath, Doran,1966)

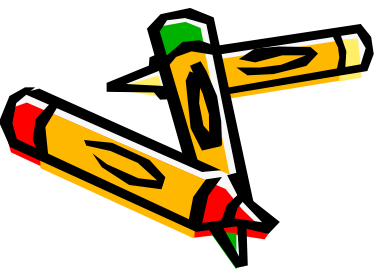


- Ecologiștii culeg informații asupra speciilor de plante ce se află într-o mulțime de parcele, inventariind speciile existente în fiecare parcelă și eventual înregistrând o măsură a bogăției fiecărei specii în parcelă.

Scop: a împărți aceste parcele în clase, astfel încât parcelele ce aparțin unei clase să aibă proprietăți distinctive față de celelalte clase.

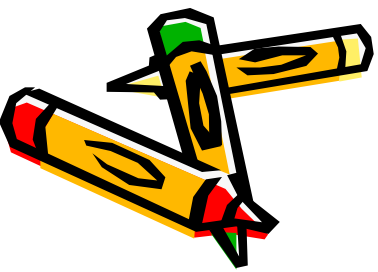
(Greig - Smith,1964)

- Analiştii social studiază interacţiunile existente într-o mulţime de indivizi şi sunt interesaţi de identificarea indivizilor ce au aceleaşi însuşiri (atribute).
(Arabie,Carrol,1989)



- Producători de whisky sunt interesați de o clasificare a distileriilor ce produc whisky, aceasta permițându-le să stabilească gama de proprietăți ale diverselor sortimente, cunoscând distileria care le produce.

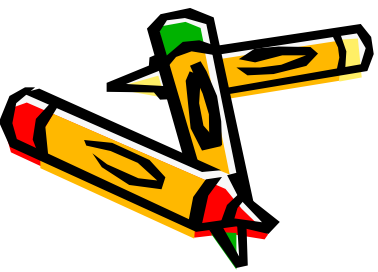
  Folosind aceste clasificări, producătorii își pot identifica concurența. (Lapointe, Legendre,1994)



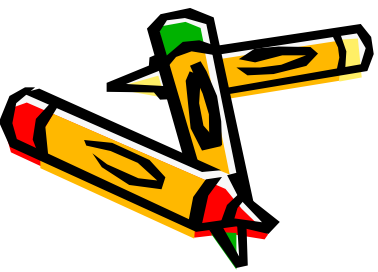
În aceste exemple un „obiect" este un artefact, o parcelă cu vegetație, un individ, o distilerie de whisky.

Obiectele sunt descrise printr-o mulțime de variabile, de exemplu:

- variabilă în cazul artefactelor este o proprietate fizică a acestora;
- în cazul parcelelor, o specie de plante este o variabilă.

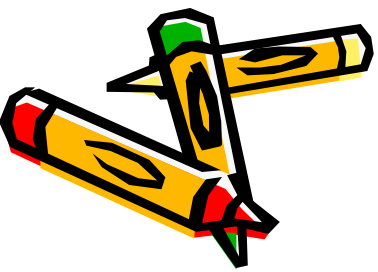


Rezultatul unei clasificări este o partiție a mulțimii obiectelor în clase disjuncte, obiectele aceleași clase fiind asemănătoare unele cu altele.



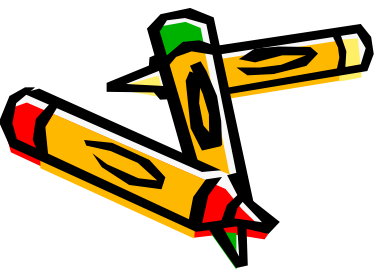
În sistemul lui Linne, o unitate taxonomică poate aparține unei specii, unui gen, unei familii, unui ordin.

Este interesant să obținem o clasificare ierarhică, care ar indica diferitele relații între clase.



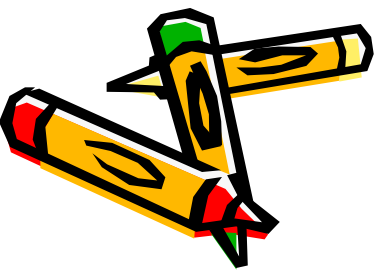
Clasificarea poate fi privită din puncte de vedere diferite:

- avem de determinat numărul de clase, trăsăturile caracteristice ale fiecărei clase şi obiectele ce constituie fiecare clasă;
- cunoscând clasele, determinăm apartenența fiecărui obiect la o anumită clasă;

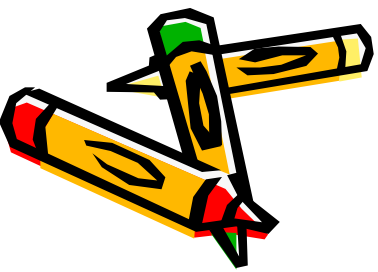


pattern recognition

- în „recunoaşterea formelor" (pattern recognition"), se presupune că fiecare obiect aparține unei anumite clase. Numărul claselor este cunoscut și cu ajutorul unei mulțimi de antrenament (**training set**) se determină proprietățile specifice fiecărei clase.
  Scopul este de a determina clasa căreia îi aparține un obiect .



Clasificarea statistică reprezintă o procedură statistică prin care obiecte individuale sunt plasate în diferite grupuri pe baza informației cantitative la dispoziție privind una sau mai multe caracteristici și utilizând o mulțime de antrenament la care se știe corespondența între fiecare obiect și categoria de care aparține.



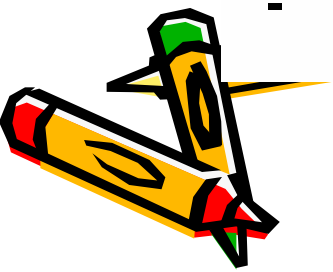
Formal, având la dispoziție o mulțime de antrenament

$$(X,Y)=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$$

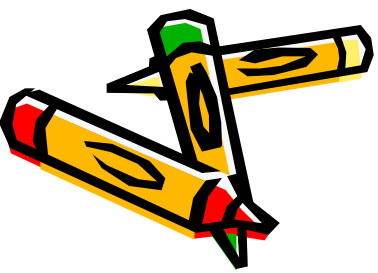
trebuie găsită o funcție de clasificare (***clasificator***)
$h: X \to Y$ pentru fiecare obiect x_i și clasă y_i
Odată funcția *h* estimată pe baza mulțimii de antrenament, utilizând corespondența $h(x_i)=y_i$, ea va fi utilizată pentru găsirea clasei corespunzătoare, adică $h(x)=?$ pentru un obiect nou, necunoscut x

aplicatii ale clasificarii

- Imagistica medicală;
- Recunoaşterea caracterelor optice;
- Geostatistica;
- Recunoaşterea vocală, a scrisului etc.;
- Biometria;
- Clasificarea documentelor;
- Căutarea pe Internet;
- Detectarea spam-urilor la poşta electronică.



Obiecte si caracteristici



clase (categorii)

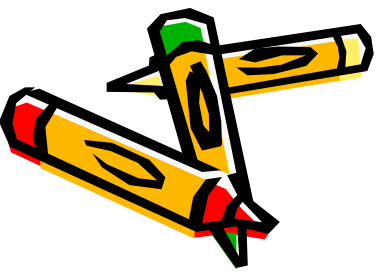
Obiectivul clasificării este de a grupa datele în *clase* (*categorii*):

- între elementele aceleiași clase avem un grad mare de **similaritate** (asemănare),
- între elementele din clase diferite gradul de similaritate este extrem de mic.

În raport cu scopul propus, se definește o măsură a similarității acestor elemente.

Fie o mulțime de elemente $X = \{x_1, \ldots, x_n\}$, pentru care este definită o măsură a similarității acestora;

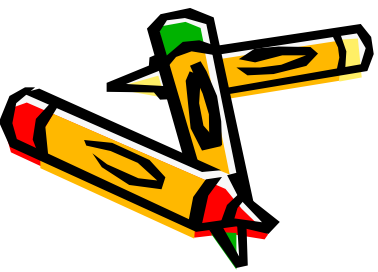
să determinăm clasele $\Omega_1, \ldots, \Omega_r$, clase ce formează o partiție a mulțimii X, astfel încât obiectele x_i care aparțin clasei Ω_j să fie cât mai asemănătoare între ele.



caracteristici (atribute)

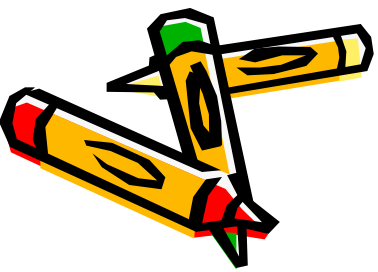
Pentru a putea decide într-o clasificare cărei categorii (clase) îi aparțin inputurile (obiectele $\mathbf{x_i}, i \in \{1, 2, \ldots, n\}$) este necesară cunoașterea mai multor caracteristici (*atribute*), provenite de obicei din măsurători.

Aceste caracteristici care pot fi considerate a fi componentele unui vector.

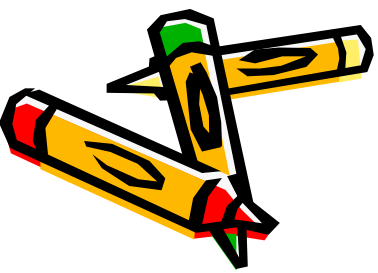


exemple actuale de clasificari automate

- recunoaşterea automată a vocii,
- recunoaşterea automată a amprentelor,
- recunoaşterea automată a lanțurilor ADN.

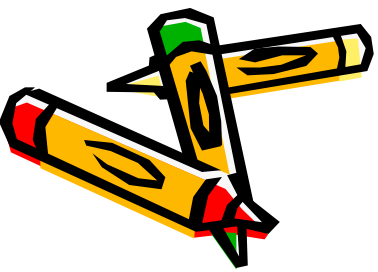


Alegerea caracteristicilor



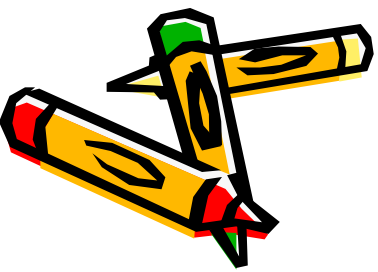
exemple clasice in literatura de specialitate

- Separarea a două specii de peşte: somon şi biban de mare, folosind senzori optici. (într-o fabrica de conserve de peşte)

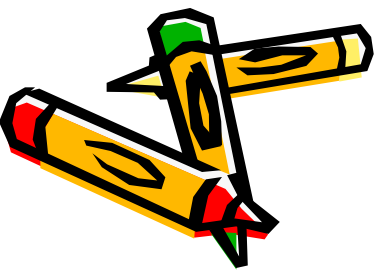


Prima etapă:
observarea diferențelor dintre cele două specii: lungime, luminozitate, lățime, numărul de configurații ale aripioarelor, etc.

Apar anumite variații de luminozitate, respectiv variații ale poziționării peștelui pe banda transportoare.

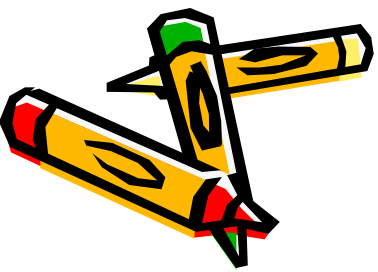


- *segmentarea*: imaginea unui peşte este izolată de a celorlalți;
- *extractorul de caracteristici* (se ocupă de reducerea bazei de date): păstrează doar o serie de caracteristici semnificative ale imaginii unui peşte.
- *clasificator*: ia decizia finală asupra speciei pe baza valorilor caracteristicilor



Tehnicile de clasificare, bazate în principal pe modele matematice, determină fiecare o anumită interpretare a structurii datelor.

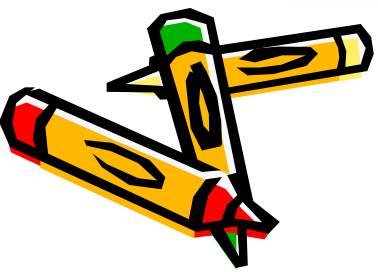
Se porneşte de obicei de la o metodă oarecare, care dă o prima clasificare ce ne furnizează o informaţie, care va fi ulterior îmbunătăţită prin aplicarea altor tehnici de clasificare.



alegerea caracteristicilor

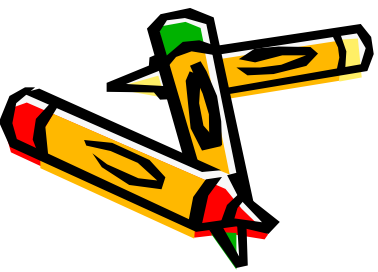
1. bibanul de mare este mai lung decât somonul;
   lungimea peștelui este o caracteristică care merită utilizată:
   dacă lungimea este mai mare decât o anumită valoare l ,
   peștele este biban de mare.

   Pentru alegerea acestui l vom face mai multe măsurători folosind o mulțime de antrenament.



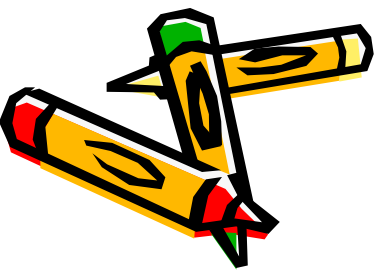
Mulțimea de antrenament este formată din 220 bibani de mare și 189 somoni.

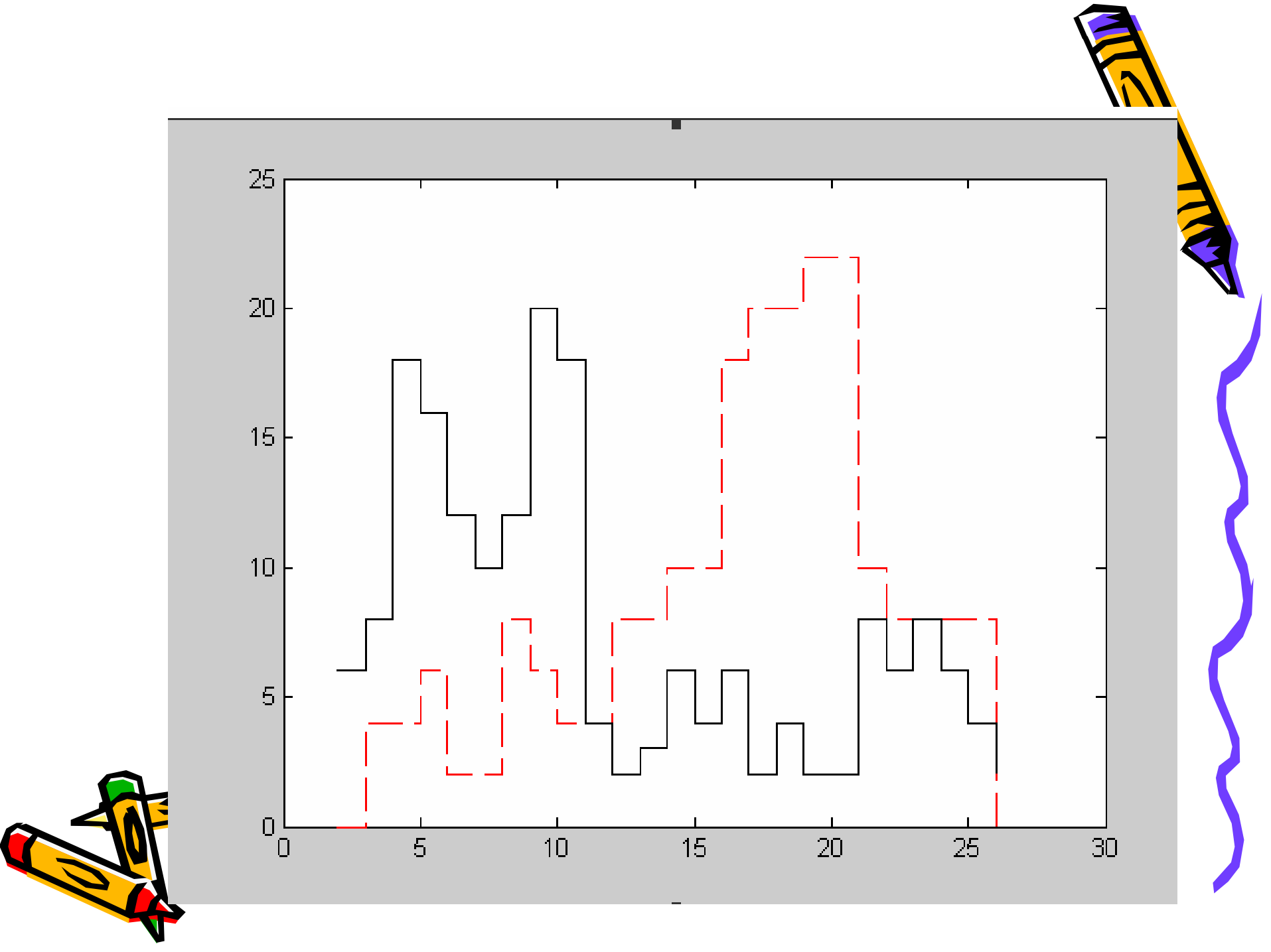
Vom reprezenta apoi grafic rezultatele obținute:



Lungime (unități)	Număr bibani de mare	Număr somoni		Lungime (unități)	Număr bibani de mare	Număr somoni
2	0	6		14	10	6
3	4	8		15	10	4
4	4	18		16	18	6
5	6	16		17	20	2
6	2	12		18	20	4
7	2	10		19	22	2
8	8	12		20	22	2
9	6	20		21	10	8
10	4	18		22	8	6
11	4	4		23	8	8
12	8	2		24	8	6
13	8	3		25	8	4

Vom desena histogramele corespunzătoare celor două specii de peşte, folosind MATLAB:

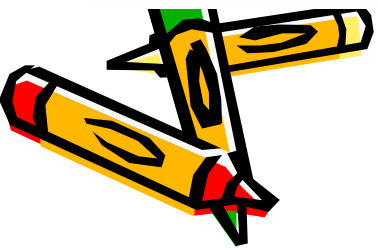



25
20
15
10
5
0
0
5
10
15
20
25
30

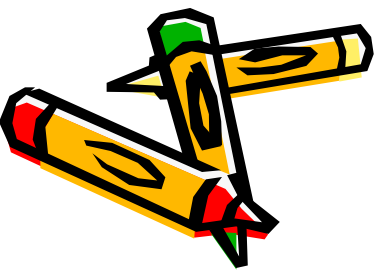
În acest caz este însă preferabil să folosim mediana lungimilor bibanului $med_B = 17$, respectiv mediana lungimilor somonului $med_S = 12$ (apar destul de multe valori extreme, care ar influența vădit valoarea mediei) .

$$l^* = \frac{med_B + med_S}{2} = 15 .$$

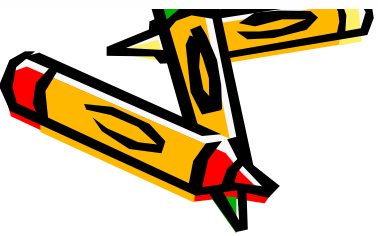
Criteriul de clasificare ales doar pe baza analizei lungimii peștilor este nepotrivit.

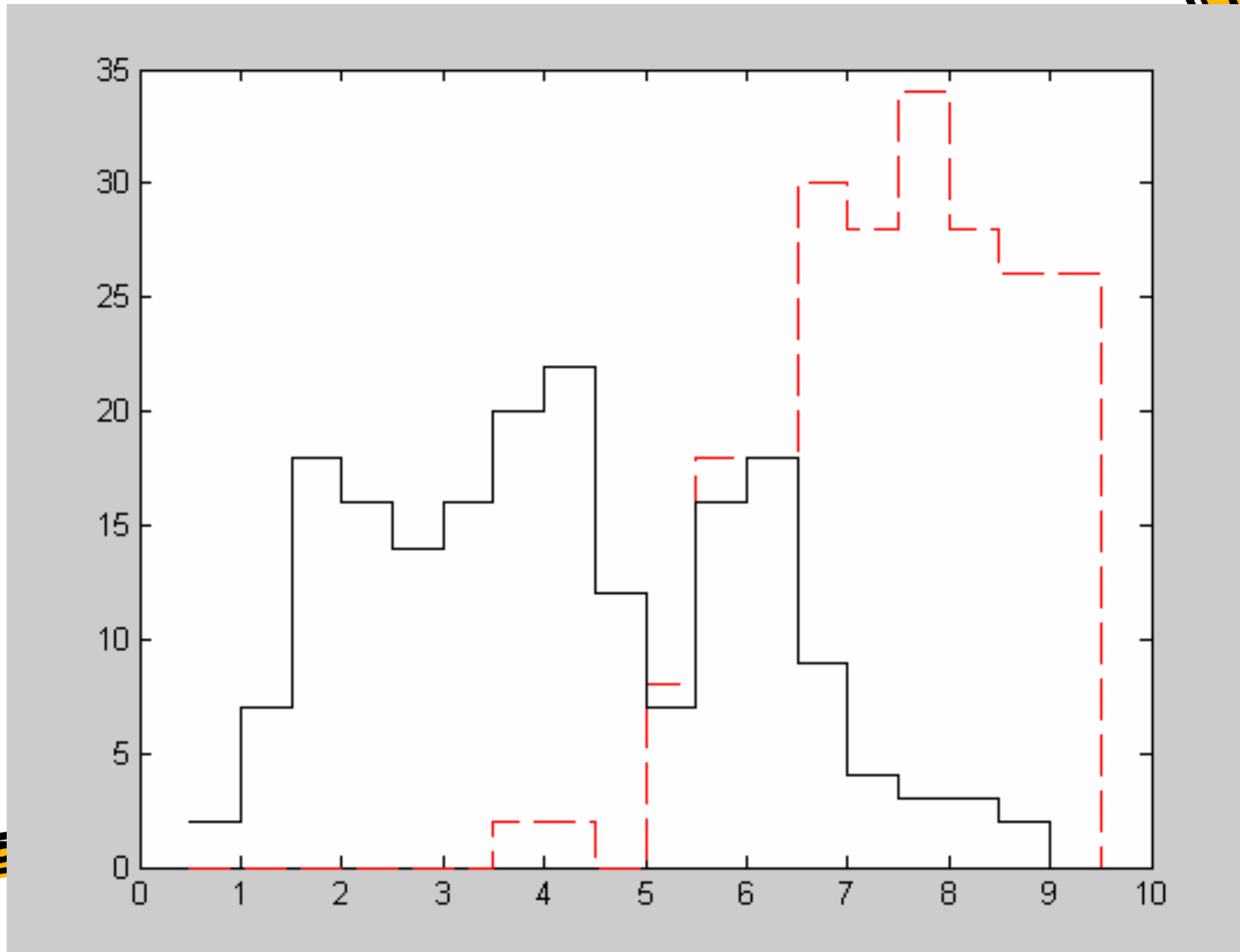


2. Pentru o clasificare mai precisă, vom alege drept caracteristică ***luminozitatea*** peștelui, după eliminarea „zgomotelor", adică a variațiilor de iluminare. Dacă luminozitatea va fi mai mică decât o anumită valoare x, peștele va fi somon, altfel este biban de mare.



Luminozitate (unități)	Număr bibani de mare	Număr somoni		Luminozitate (unități)	Număr bibani de mare	Număr somoni
0.5	0	2		5.5	18	16
1.0	0	7		6.0	18	18
1.5	0	18		6.5	30	9
2.0	0	16		7.0	28	4
2.5	0	14		7.5	34	3
3.0	0	16		8.0	28	3
3.5	2	20		8.5	26	2
4.0	2	22		9.0	26	0
4.5	0	12		9.5	0	0
5.0	8	7				



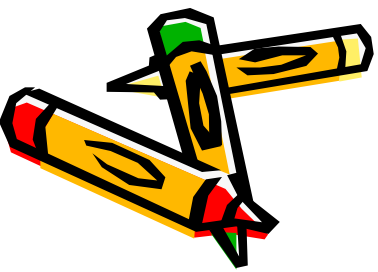

35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10

Mediile unităților ce măsoară luminozitatea sunt $\overline{m_B} = 7.21$, $\overline{m_S} = 3.93$ şi astfel $x = 5.57$.

Dacă lucrăm cu medianele, avem

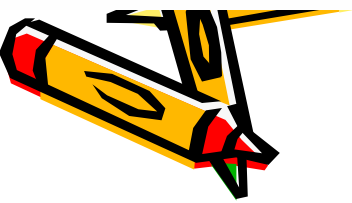
$med_B = 7.5$, $med_S = 4$ şi astfel $x^* = 5.75$.

Folosind luminozitatea peşteluio clasificare nesatisfăcătoare, reprezentând totuşi o abordare mai bună decât prima.



O clasificare greșită are evident, un anumit cost:

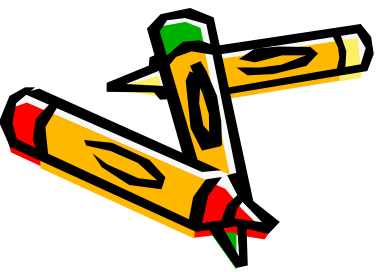
- dacă un somon este vândut ca biban de mare, fabrica este în pierdere;
- din punctul de vedere al cumpărătorului, dacă găsește într-o conservă etichetată biban de mare o bucată de somon, nu e deranjat, dar dacă într-o conservă etichetată somon găsește o bucată de biban de mare, sigur nu va mai cumpără acest tip de produs.



frontiera de decizie

Pentru a nu-și pierde clienții, fabrica va prefera prima variantă.

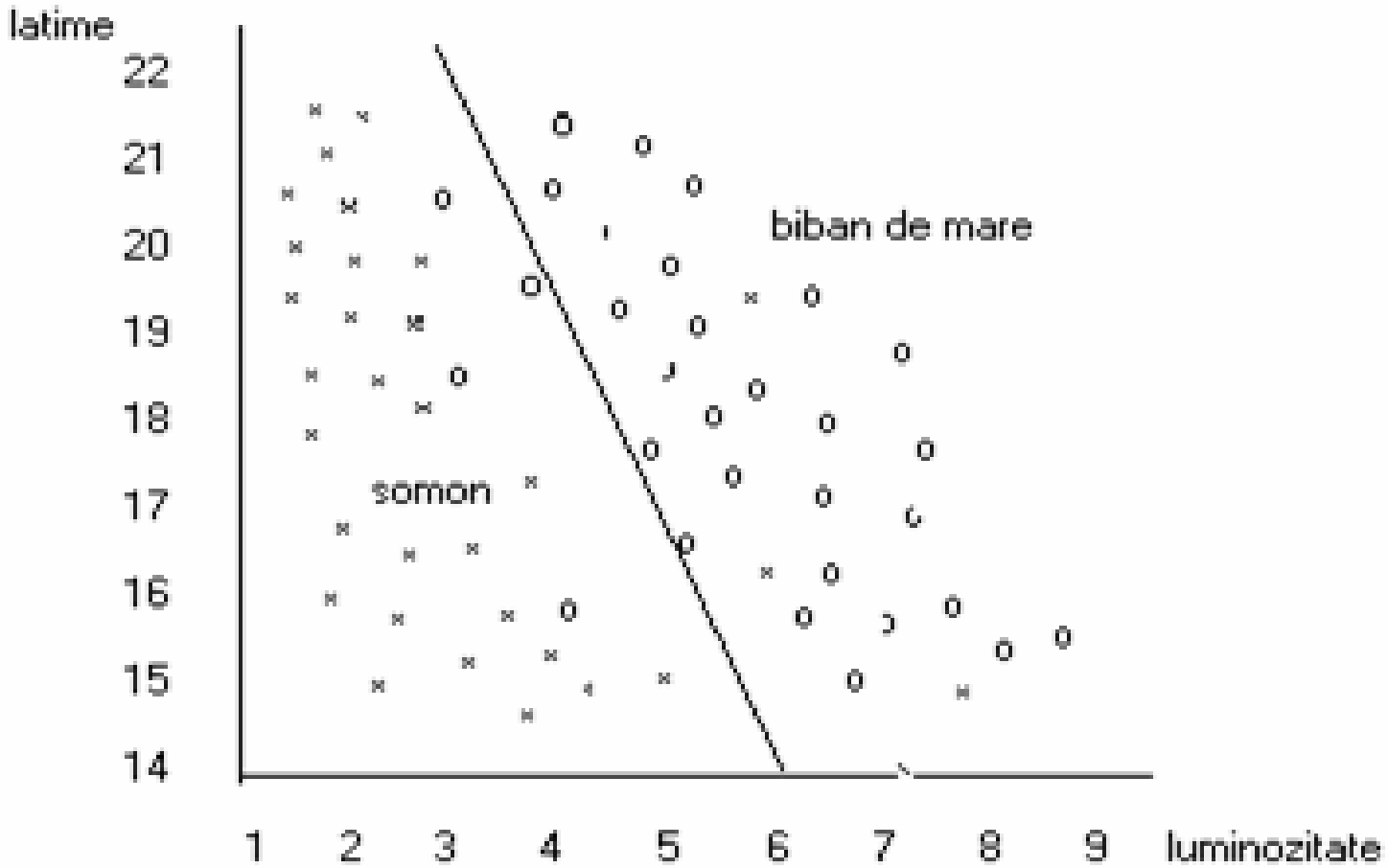
Obiectivul urmărit este alegerea ***frontierei*** (*suprafeței*) ***de decizie***, ceea ce înseamnă de fapt obținerea unei reguli de decizie, care să minimizeze costurile clasificării greșite (***misclassification cost***).



spatiul bidimensional al caracteristicilor

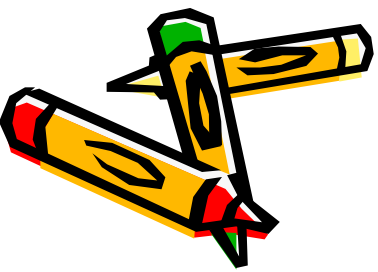
3. Pentru a îmbunătăți clasificarea, vom utiliza două caracteristici: luminozitatea peștelui x_1 și lățimea peștelui x_2, considerând vectorul $\mathbf{x} = (x_1, x_2)$, vector care reprezintă fiecare pește.

Exemplele de pești cunoscute vor fi reprezentate în spațiul bidimensional al caracteristicilor.
Construim o dreaptă, *frontiera de decizie*, care va separa planul în două regiuni de decizie, corespunzătoare celor două tipuri (*clase*) de pește.


latime
22
21
20
19
18
17
16
15
14
biban de mare
somon
1
2
3
4
5
6
7
8
9
luminozitate

merită să mărim dimensiunea spațiului caracteristicilor, luând în considerare și alte atribute, cum ar fi lungimea, culoarea sau poziționarea ochilor?

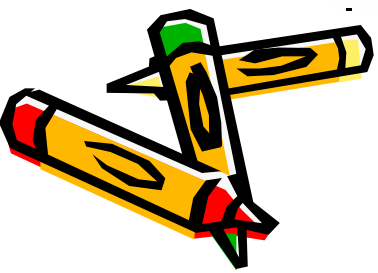
Unele caracteristici pot fi redundante, de exemplu culoarea ochilor peștilor este corelată cu lățimea lor, și astfel nu merită introdus acest nou atribut.



overfitting

Prea multe caracteristici pot influența negativ clasificarea, rezultând o suprafață de decizie particulară, specifică doar mulțimii de antrenament utilizate -fenomenul de ***overfitting***.

Obiectivul propus este de a determina, utilizând mulțimea de antrenament, o suprafață de decizie care să poată fi folosită cu succes în cazul unor exemple (inputuri) noi.



- Clasificarea celor trei tipuri de flori de Iris: *Iris Setosa*, *Virginica* şi *Versicolor* (există o bază de date foarte bună, datorată lui R.A. Fisher, 1936).

  Una dintre metode constă în măsurarea lungimii şi lăţimii petalelor. Notând

  x_1 = lungimea în cm a petalei şi

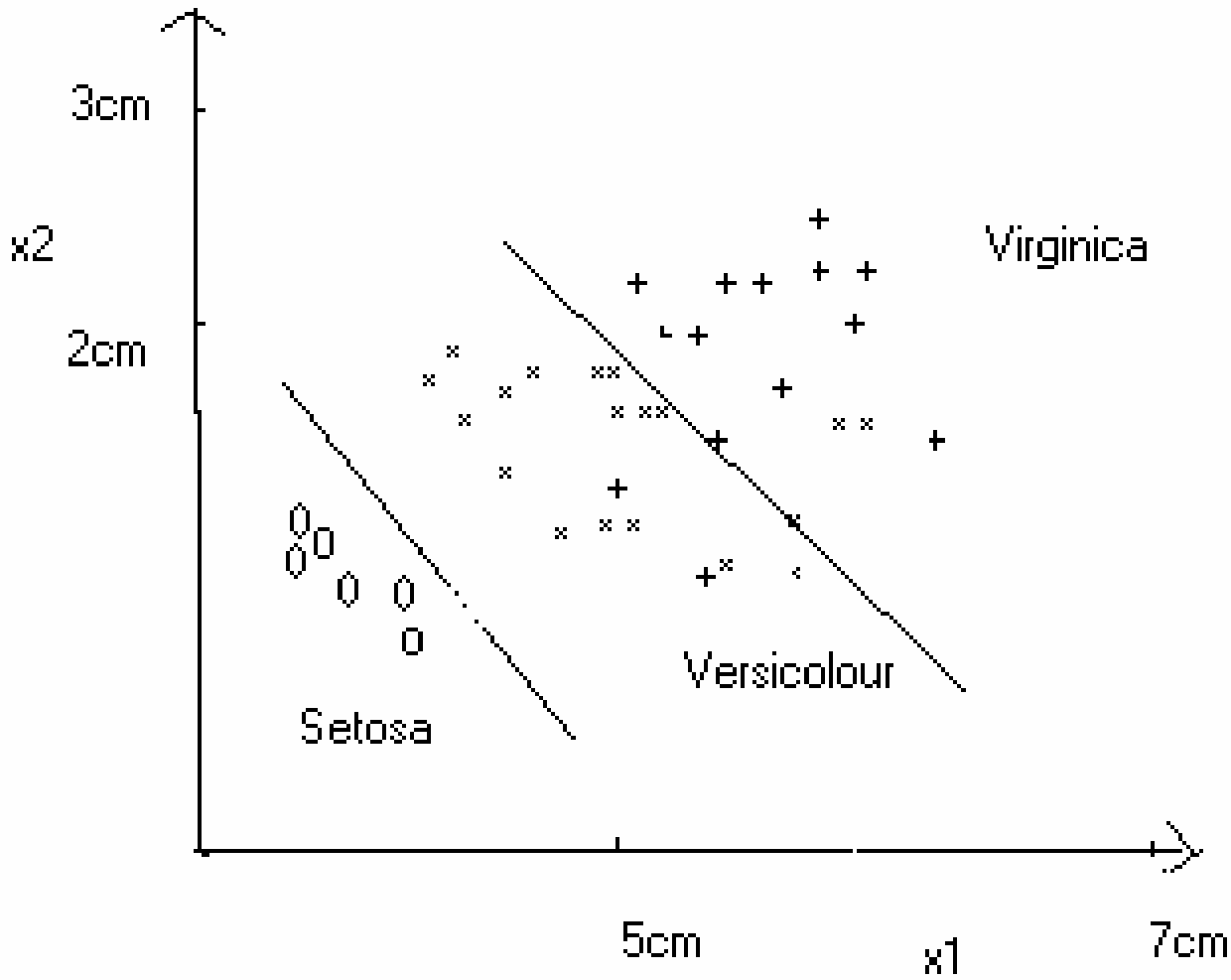
  x_2 = lăţimea în cm a petalei,

  construim vectorul $\mathbf{x} = (x_1, x_2)$, corespunzător unui iris.

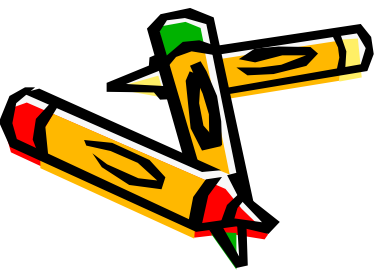
Iris setosa

Iris versicolor

Iris verginica

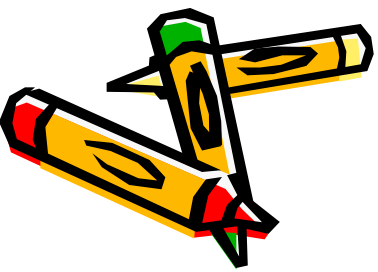

3cm
x2
2cm
Virginica
Versicolour
Setosa
5cm
x1
7cm

Fiecare punct din planul euclidian al caracteristicilor reprezintă un exemplu real de floare.
Cele două drepte, numite *suprafețe de decizie*, separă planul în trei regiuni, corespunzătoare celor trei clase (tipuri) de Iris.



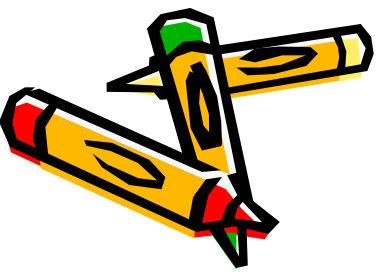
alegerea optima a caracteristicilor

În cazul unui număr mai mare de caracteristici (*curse of dimensionality*), este preferabilă reducerea dimensiunii vectorilor, studiind atent corelațiile existente între componente, folosind diverse metode statistice



În general, încercăm să lucrăm cu un număr mic de caracteristici, care sunt mai ușor de utilizat și cu ajutorul cărora vom obține regiuni de decizie mai simple.

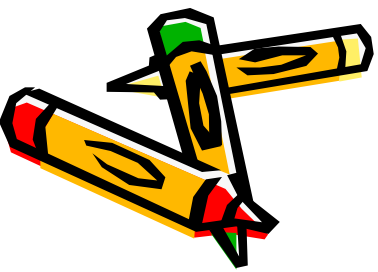
Suntem interesați de date „robuste", care nu sunt influențate de factorii externi (zgomot).



În cazul a *p* caracteristici, acestea vor fi componentele unui vector *p*-dimensional $\mathbf{x}_k = (x_1^k, \ldots, x_p^k)$.

Considerând *n* vectori de dimensiune *p*, notația x_i^k utilizată de obicei se referă la a *i*-a variabilă (caracteristică observată) a vectorului $\mathbf{x}_k$ (obiectul numărul *k* din mulțimea de antrenament).

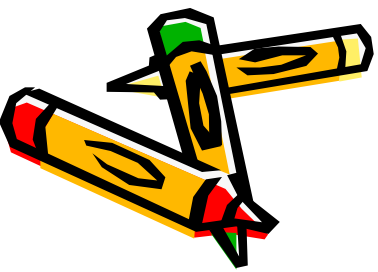
	Variabila 1	. .	Variabila *i*	. .	Variabila *p*
Obiectul 1	x_1^1		x_i^1		x_p^1
.....					
Obiectul *k*	x_1^k		x_i^k		x_p^k
.....					
Obiectul *n*	x_1^n		x_i^n		x_p^n



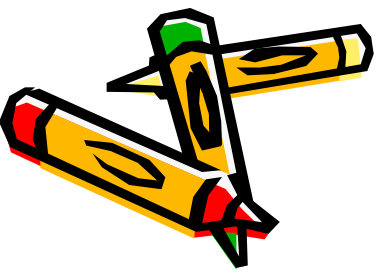
O alternativă mai utilă este aceea a reprezentării acestor date sub forma unei matrice **X**, cu *n* linii și *p* coloane:

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \dots & x_i^1 & \dots & x_p^1 \\ \dots & \dots & \dots & \dots & \dots \\ x_1^k & \dots & x_i^k & \dots & x_p^k \\ \dots & \dots & \dots & \dots & \dots \\ x_1^n & \dots & x_i^n & \dots & x_p^n \end{pmatrix}$$

Asemenea date vectoriale multidimensionale pot fi vizualizate până la un anumit punct, folosind software de vizualizare (de exemplu XGOBI). Aceste instrumente de vizualizare sunt utilizate pentru a vedea legăturile în spațiu dintre vectori, ca un ghid în alegerea caracteristicilor distinctive.



Recunoaşterea statistică a formelor se ocupă nu numai de optimizarea suprafeței de decizie, ci face prognoze asupra modului cum va reacționa clasificatorul construit la noile exemple.



- construirea clasificatorului folosind mulțimea de antrenament,
- verificarea performanței utilizând o mulțime de testare, în cazul căreia se cunoaşte cărei clase îi corespunde fiecare element.
- tabel cu numărul de clasificări corecte, respectiv incorecte, din fiecare clasă.
- din procentajul de clasificări corecte pentru fiecare clasă, se calculează procentajul clasificării corecte pentru toată mulțimea.

procentaj clasificare corecta

Presupunem că mulțimea **X** are *n* elemente, care aparțin la *r* clase $\Omega_1,\ldots,\Omega_r$.

Numărul elementelor din fiecare clasă Ω_i este n_i, în timp ce numărul elementelor corect clasificate în clasa Ω_i este $m_i \leq n_i$.

Procentaj clasificare corectă

$$\sum_{i=1}^{r} \frac{n_i}{n} \cdot \frac{100 m_i}{n_i} = \frac{\sum_{i=1}^{r} 100 m_i}{n}$$

misclassification matrix

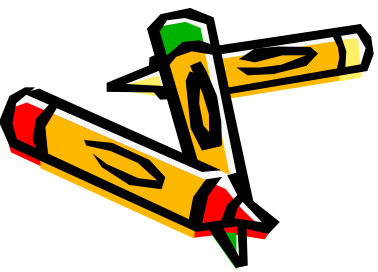
Pentru a vizualiza calitatea clasificării se foloseşte matricea clasificărilor greşite (*misclassification matrix*):

$$M_C = \begin{pmatrix} c_{11} & c_{12} & \dots & \dots & \dots & c_{1r} \\ c_{21} & c_{22} & \dots & \dots & \dots & c_{2r} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & c_{ij} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ c_{r1} & c_{r2} & \dots & \dots & \dots & c_{rr} \end{pmatrix}$$

c_{ij} reprezintă numărul vectorilor de testare care sunt din clasa Ω_i, dar au fost clasificați greşit ca aparținând clasei Ω_j:

etapele unei probleme de clasificare

- pre-procesarea datelor;
- alegerea caracteristicilor;
- stabilirea funcției de decizie (funcția discriminant);
- clasificarea.



pre-procesarea datelor asigură

- reducerea dimensiunii datelor;
- filtrarea factorilor externi (zgomot);
- suprimarea detaliilor nerelevante;
- sublinierea caracteristicilor importante;
- invarianța în raport cu translațiile și rotațiile;
- pregătirea datelor pentru procedeul de decizie prin scalarea sau normalizarea lor.

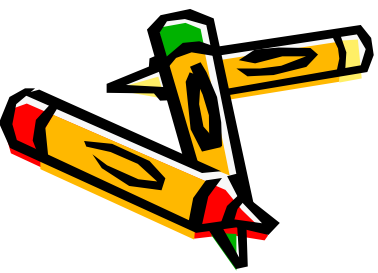
clasificarea datelor

Datele obținute prin măsurare pot fi clasificate în funcție de tipul de informație conținut.

- ***Datele categoriale*** sunt acele date care împart obiectele în diferite categorii.
- ***Datele numerice***

date categoriale

- date ***nominale***, ca de exemplu grupa sanguină (A/B/AB/O), culoarea ochilor, specia de Iris (Iris Setosa, Virginica și Versicolour).
- datele ***ordinale*** sunt date enumerative ordonate ca, de exemplu: gradul fumatului (nefumător, fost fumător, fumător ‚amator', fumător ‚înrăit'), ierarhizarea durerii (mică, medie, mare),



date nominale

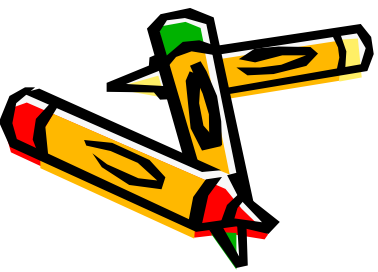
Datele nominale pot fi şi numerice, de exemplu codurile poştale

Datele nominale pot fi:

- binare (0 sau 1, da / nu, adevărat / fals)
- enumerative (date discrete pentru care nu este definită o ordine, cum ar fi categoriile socio-profesionale sau culoarea ochilor).

date numerice

- Datele ***discrete*** apar atunci când este vorba de observații numerice întregi, privitoare la un anumit proces de numărare ; de exemplu: numărul de copii ai unei familii, pulsul, codul numeric.
- Datele numerice ***continue*** se obțin de obicei în urma unor măsurători, de exemplu înălțimea, greutatea, tensiunea arterială, colesterolul unei anumite persoane, temperatura, viteza vântului, valoarea contului din bancă sau valoarea acțiunilor tranzacționate la Bursă etc.



remarca 1

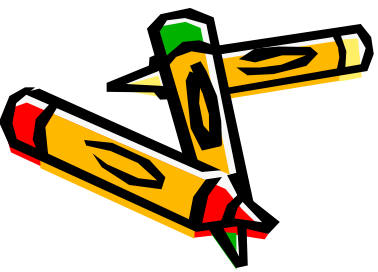
Din date continue se pot obține date discrete:

evaluarea venitului lunar:

venit lunar < 1000 lei,

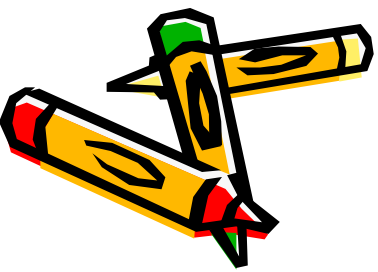
1000 lei < venit lunar < 2000 lei .

20 00 lei < venit lunar



remarca 2

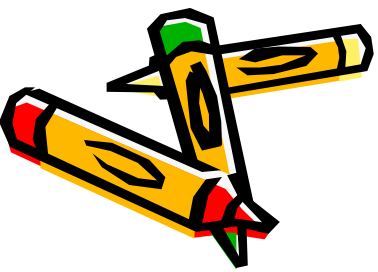
Datele numerice discrete sunt câteodată tratate ca date categoriale, de exemplu numărul de copii născuți de o femeie, 0, 1, 2, 3, 4, împart mamele în categoriile corespunzătoare numărului de copii.



remarca 3

nu este corect să interpretăm datele categoriale ordonate ca date numerice:

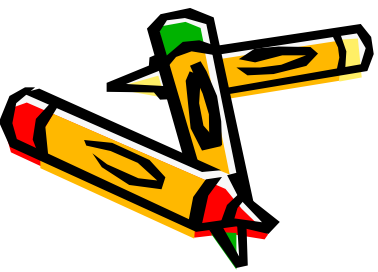
la stadiile în anumite boli, stadiul IV nu este de două ori mai rău decât stadiul II, ş.a.m.d.



clasificator

Un ***clasificator*** poate fi considerat a fi o aplicație între mulțimea de caracteristici și mulțimea claselor.

Aceasta se realizează folosind distanțe (metrici), definite pe spațiul caracteristicilor și elemente de teoria probabilităților.

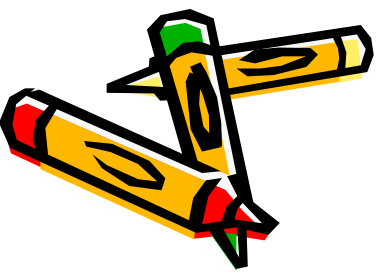


- *distanța euclidiană* în $\mathbb{R}^p$ între doi vectori $\mathbf{x}=(x_1,\ldots,x_p)$ și $\mathbf{y}=(y_1,\ldots,y_p)$ este definită prin:

$$d(\mathbf{x},\mathbf{y})=\sqrt{\sum_{k=1}^{p}(x_k-y_k)^2}$$

- *distanța Manhattan* $d_1(\mathbf{x},\mathbf{y})=\sum_{k=1}^{p}|x_k-y_k|$

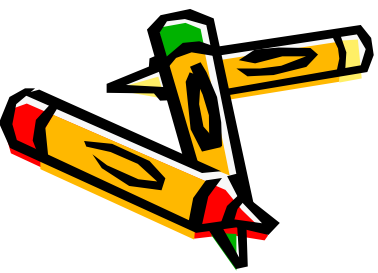
- distanța *Cebâșev* $d_\infty(\mathbf{x},\mathbf{y})=\max_{1\le k\le p}|x_k-y_k|$



Merita retinut!

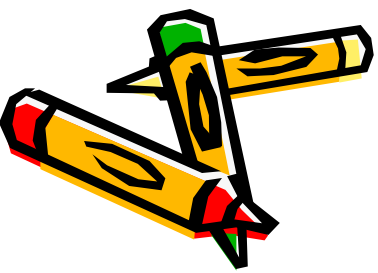
- pentru *datele binare*, alegem distanța definită prin:
  $d(0,0) = d(1,1) = 0$ și $d(1,0) = d(0,1) = 1$;

- pentru *datele enumerative*, distanța cea mai utilizată este $d(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$



clasificatorul liniar

Cel mai simplu tip de funcție de decizie este *clasificatorul liniar*, caz în care suprafețele de decizie sunt *hiperplane*.
Un hiperplan unidimensional este o valoare prag (*treshhold*), un hiperplan bidimensional este o dreaptă în timp ce un hiperplan în $\mathbf{R}^3$ este un plan în sensul cunoscut.

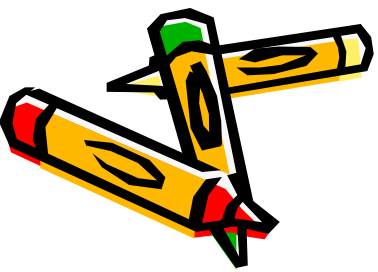


- Un exemplu simplu de funcție discriminant în cazul unidimensional este indicele de obezitate - IMC.

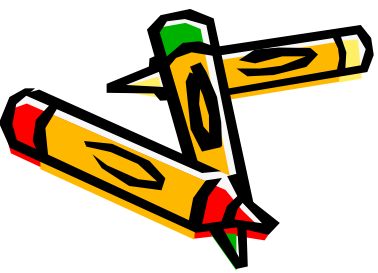
  Acesta se calculează raportând greutatea corporală (în kg) la pătratul înălțimii (în metri), conform formulei:

$$IMC = \frac{m}{h^2}$$

Un indice mai mare sau egal cu 30 clasifică persoana drept obeză, având nevoie de ajutor medical.

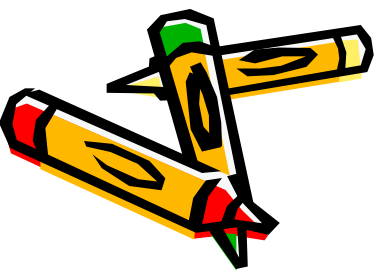


Clasificarea bayesiană



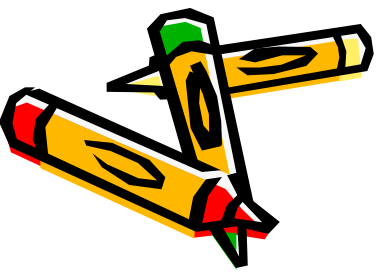
risc prognozat

Strategiile de decizie Bayes sunt folosite în scopul de a minimiza *„riscul prognozat"* .
Se aplică în probleme de clasificare cu număr mare de clase.



Termenul de *şansă* exprimă un grad subiectiv de încredere într-un rezultat particular.

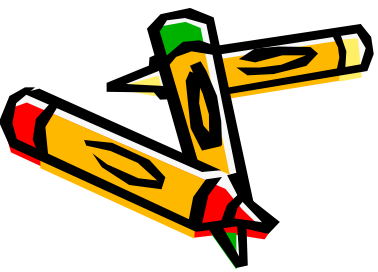
Abordarea matematică a noțiunii de şansă a dat naştere termenului de ***probabilitate***



probabilitatea conditionata

Probabilitatea condiționată este folosită pentru a măsura încrederea că un eveniment aleator va avea loc știind că alt eveniment aleator a avut loc.

Fie două evenimente A și B, probabilitatea condiționată ca evenimentul A să aibă loc știind ca evenimentul B a avut loc este $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$



- studenții promovează examenele A și B cu probabilitățile: $A \sim 70\%$, $B \sim 40\%$ şi $A \cap B \sim 35\%$.

- Probabilitatea condiționată ca studenții să promoveze cele două examene, dacă au promovat examenul A este

$$P(A \cap B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{1}{2}.$$

- Probabilitatea condiționată ca studenții să promoveze cele două examene, dacă au promovat examenul B este

$$P(A \cap B \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{7}{8}$$

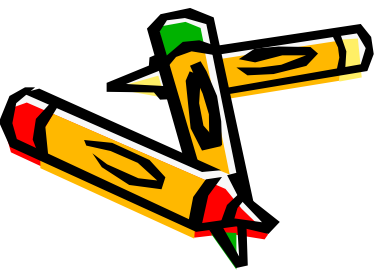
- Pentru a obține probabilitatea ca studenții să promoveze cele două examene, dacă au promovat cel puțin un examen calculăm

probabilitatea ca să promoveze cel puțin un examen:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7 + 0.4 - 0.35 = 0.75$$

și astfel

$$P(A \cap B \mid A \cup B) = \frac{P(A \cap B)}{P(A \cup B)} = \frac{7}{15}$$



- Fie $\mathbf{X}$ o bază de date medicale pentru pacienții cu diferite boli ale ficatului.
  - $\mathbf{x}_n$ este vectorul caracteristicilor medicale corespunzător pacientului cu numărul *n*, pacient ce este diagnosticat cu cancer hepatic,
  - $\mathbf{x}$ este vectorul caracteristicilor medicale al unui pacient ce nu a fost încă diagnosticat.

  Probabilitatea ca pacientul căruia nu i s-a pus încă un diagnostic să aibă cancer hepatic dacă pacientul cu numărul *n* are este $P(\mathbf{x}|\mathbf{x}_n)$.

formula Bayes

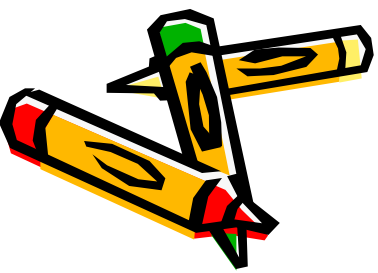
- Fie $(\Omega, \Sigma, \mathbf{P})$ un *spațiu de probabilitate*, B un eveniment arbitrar din Σ și $\{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$ o partiție a spațiului Ω. Atunci:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)} = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)},$$

$$P(B) > 0, \quad P(A_i) > 0, i = 1, .., n$$

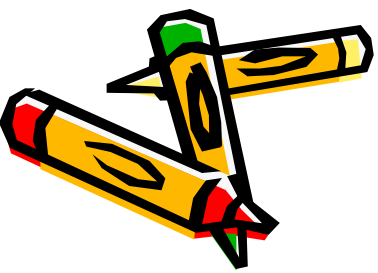
denumiri uzuale

- $P(A_i \mid B)$ - probabilitate *posterioară* (posterior probability),
- $P(A_i)$ - probabilitate *apriorică* (prior probability),
- $P(B \mid A_i)$ - *verosimilitate* (likelihood),
- iar $P(B)$ - *evidență/ dovadă* (evidence).



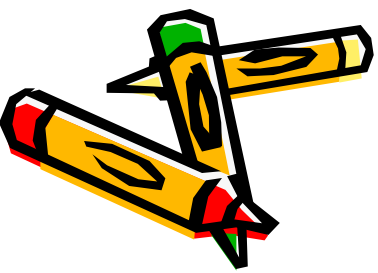
formula Bayes

$$probabilitate\ posterioara = \\ = \frac{verosimilitate \times probabilitate\ apriorica}{evidenta}$$



exemple

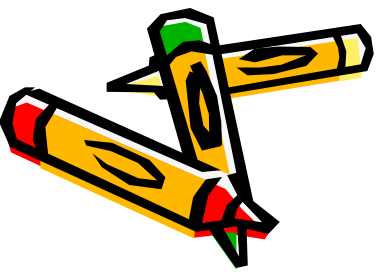
- In secția de gastroenterologie dintr-un spital sunt internați 49% bărbați și 51% femei.
  Din experiența anterioară estimăm că 2.5% dintre bărbați și 1.9% din femei au cancer hepatic.
  Notații:
  M- bărbați, F-femei
  A evenimentul ca diagnosticul să fie cancer hepatic



probabilitatea ca un individ, arbitrar ales, să fie diagnosticat cu cancer hepatic se calculează cu formula probabilității totale:

$$P(A) = P(A|M) \cdot P(M) + P(A|F) \cdot P(F) =$$
$$= 0.025 \cdot 0.49 + 0.09 \cdot 0.51 = 0.022 = 2.2\% ,$$

2.2% dintre bolnavii internați în secție au acest diagnostic



unde

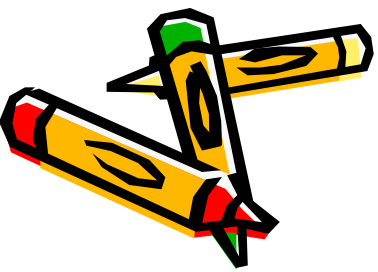
- $P(A|M)$ - probabilitatea ca o persoană să fie diagnosticată cancer hepatic, condiționată de faptul că este bărbat;
- $P(A|F)$ - probabilitatea ca o persoană să fie diagnosticată cancer hepatic, condiționată de faptul că este femeie;
- $P(M)$ proporția pacienților bărbați din totalul pacienților;
- $P(F)$ proporția pacienților femei din totalul pacienților.

- Folosind formula Bayes, vom calcula probabilitatea ca un bolnav diagnosticat cu cancer hepatic să fie bărbat, respectiv femeie:

$$P(M \mid A) = \frac{P(A \mid M) \cdot P(M)}{P(A)} = 0.5583 = 55.83\%$$

$$P(F \mid A) = \frac{P(A \mid F) \cdot P(F)}{P(A)} = 0.4417 = 44.17\%$$

Dintre bolnavii cu acest diagnostic 55.83% sunt bărbați, respectiv 44.17% sunt femei.

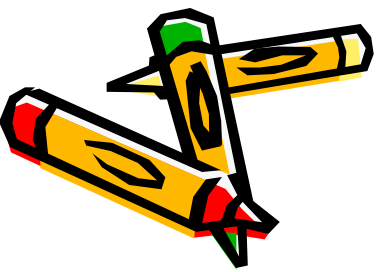


- Considerăm că într-o companie se produc aceleași produse în trei unități distincte U_1, U_2, U_3, ce au capacitățile de producție 60%, 30%, 10%, procente ce reprezintă probabilitățile ca un produs să provină de la una dintre cele trei unități.
  Fiecare unitate are rata de a produce obiecte cu defecțiuni de 6%, 3%, 5%.
  Probabilitatea ca un produs defect, arbitrar ales, să provină de unitatea U_1, respectiv U_2 sau U_3.

Notăm cu *A* evenimentul ca un produs ales la întâmplare să fie defect. Calculăm care este probabilitatea ca un produs arbitrar ales să fie defect.

$$P(A) = P(A|U_1) \cdot P(U_1) + P(A|U_2) \cdot P(U_2) +$$
$$+ P(A|U_3) \cdot P(U_3) = 0.06 \cdot .6 + 0.03 \cdot 0.3 + 0.05 \cdot 0.1 = 0.05$$

(formula probabilității totale)



calculăm probabilitatea ca un produs defect să provină din U_i:

$$P(U_1 \mid A) = \frac{P(A \mid U_1) \cdot P(U_1)}{P(A)} = \frac{0.06 \cdot 0.6}{0.05} = 0.72 = 72\%.$$

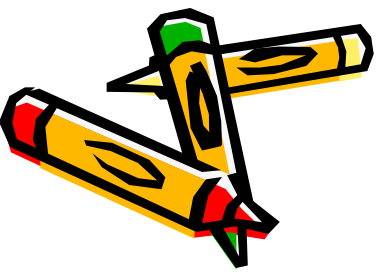
$$P(U_2 \mid A) = \frac{P(A \mid U_2) \cdot P(U_2)}{P(A)} = 0.18 = 18\%$$

$$P(U_3 \mid A) = \frac{P(A \mid U_3) \cdot P(U_3)}{P(A)} = 0.1 = 10\%$$

(formula Bayes)

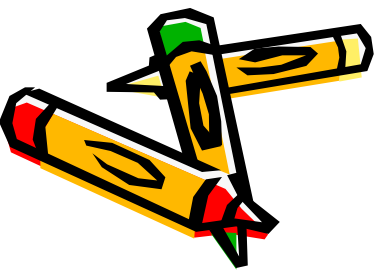
Prezentăm o problemă medicală simplificată:

- Testul unei boli rare este corect în proporție de 99%.
- O persoana din 100 000 prezintă această boală.
- Efectuați testul și răspunsul este pozitiv



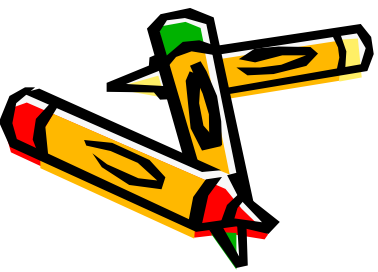
Se afirmă că din 1 000 000 de persoane, 10 000 (1%) vor fi considerate a fi posibili bolnavi, în timp ce doar 10 persoane (1 la 10 000) au intr-adevăr aceasta boală Așadar acest test, cu fiabilitatea de 99%, in cazul în care este pozitiv, dă 999 alerte false din 1000.

Nici un test nu este perfect și o problema serioasă o constituie rezultatele fals pozitive.



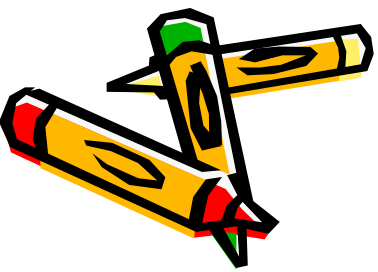
În general, în cazul in care o persoană este testată dacă suferă de o anumită boală, riscul ca rezultatul să fie pozitiv, dacă persoana este sănătoasă, este infim.

Problema este să determinăm în cazul unei boli rare probabilitatea ca un test pozitiv să fie greşit.



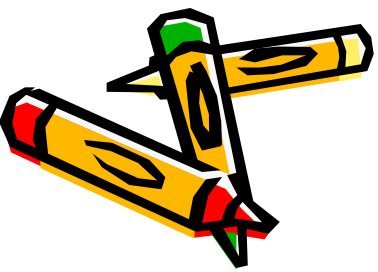
Considerăm cazul unei boli rare și aplicăm un test foarte fiabil:

- dacă pacientul a contractat boala, testul este pozitiv în 99% din cazuri, adică cu o probabilitate de 0.99.
- dacă pacientul este sănătos, testul este corect, adică negativ în 95% din cazuri, (cu o probabilitate de 0.95).



Boala atinge o persoană din 1000, deci are probabilitatea de 0.001 (pare mică dar în cazul unei boli mortale este considerabilă).

Avem toate informațiile pentru a determina probabilitatea ca testul să fie fals pozitiv

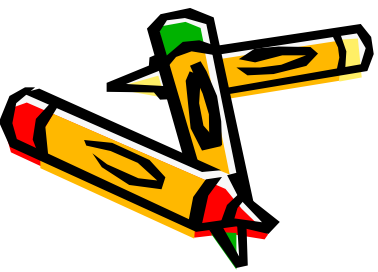


A evenimentul: "pacientul a contractat boala"

B evenimentul " testul este pozitiv"

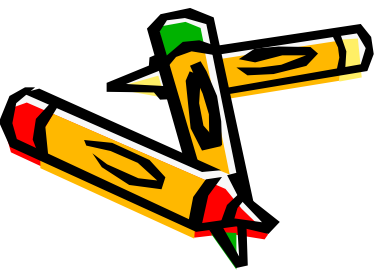
$$P(A \mid B) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} = 0.019$$

Ştiind că testul este pozitiv, probabilitatea ca pacientul să fie sănătos este $1 - 0.019 = 0.981$



Dacă tratamentul este complicat, costisitor sau periculos, pentru un pacient sănătos este nevoie de un test complementar, care va fi sigur mai precis şi mai costisitor.

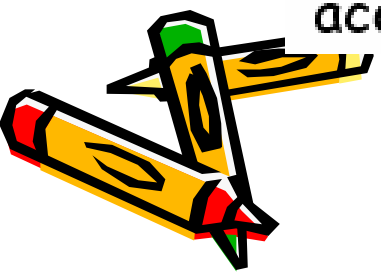
Primul test a eliminat cazurile cele mai evidente.



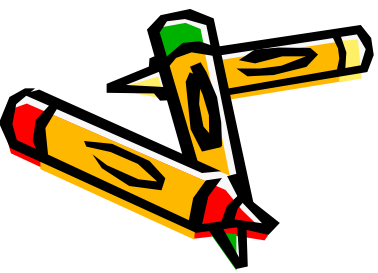
problema juridica si sociala

Dacă probabilitatea unui anumit tip de comportament, să zicem **delicvența** depinde de anumiți factori sociali, culturali sau ereditari, atunci:

- Aceasta presupune o reducere parțială a responsabilității morale şi juridice a delincventului, ceea ce antrenează o creştere a responsabilității societății, care nu a ştiut sau nu a reuşit să neutralizeze aceşti factori.

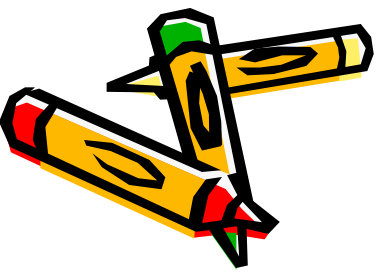


- Pe de altă parte această informație poate fi utilizată pentru ca politica de prevenție să fie orientată corespunzător şi trebuie văzut dacă interesul public sau morala se va acomoda la aceasta discriminare de facto a cetățenilor, chiar dacă este pozitivă.



riscul asteptat

Scopul, în teoria deciziilor, este de a minimiza probabilitatea de a greși sau *riscul așteptat*.



exemplu

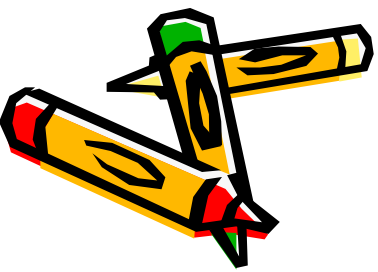
- Să considerăm un set de date medicale, corespunzătoare unei mulțimi $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ de pacienți.
  În urma analizelor avem două rezultate posibile: benign sau malign, ceea ce corespunde la două clase Ω_1 și Ω_2. Dacă mulțimea $\mathbf{X}$ este suficient de mare, definim probabilitățile apriorice $P(\Omega_1)$ și $P(\Omega_2)$.
  Dacă afirmația „pacientul aparține clasei Ω_1 (*benign*)" apare în 9 cazuri din 10, avem $P(\Omega_1) = 0.9$ și $P(\Omega_2) = 0.1$.

Încercând să clasificăm pacienții cunoscând doar probabilitățile apriorice, conform regulii:

$\mathbf{x} \in \Omega_1$ dacă $P(\Omega_1) > P(\Omega_2)$,

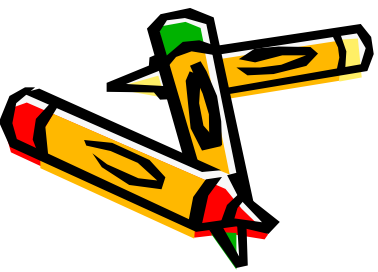
rezultatul nu este mulțumitor.

Conform regulii de mai sus orice nou pacient va fi clasificat în principiu ca fiind benign, cu toate că știm că un caz din 10 este malign.



regula de decizie bayesiana

- Fie D_i regula de decizie referitoare la clasa Ω_i .
- Fiind dat un vector $\mathbf{x}$, eroarea relativă la clasa Ω_i este definită de $P(eroare/\mathbf{x}) = 1 - P(\Omega_i|\mathbf{x})$.
- Se *minimizează* probabilitatea de a greşi.



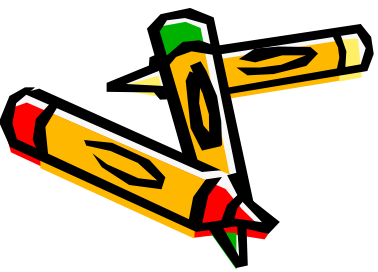
- Regula bayesiană de decizie este:

  Alege D_j dacă

  $$P(\Omega_j|\mathbf{x}) > P(\Omega_i|\mathbf{x}),\ i \in \{1,\ldots,j-1,j+1,\ldots r\}$$

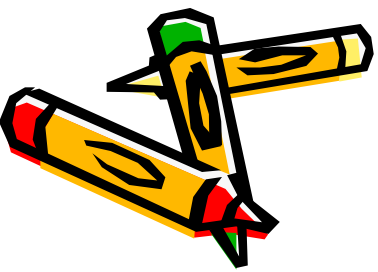
  sau echivalent

  $$P(\mathbf{x}|\Omega_j)\cdot P(\Omega_j) > P(\mathbf{x}|\Omega_i)\cdot P(\Omega_i)\ ,i \in \{1,\ldots,j-1,j+1,\ldots r\}.$$



Să considerăm un set de date care urmează a fi clasificate utilizând un clasificator bayesian; presupunem că fiecare atribut (inclusiv atributul corespunzător etichetei de clasă) este o variabilă aleatoare.

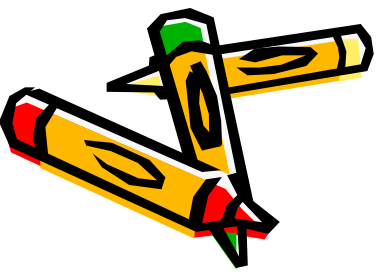
Fiind dat un obiect cu atributele $\{A_1, A_2, \ldots, A_p\}$ ne propunem clasificarea sa în clasa Ω_i.



Clasificarea este corectă atunci când probabilitatea condiționată:

$$P(\Omega_i \mid A_1, A_2, \ldots, A_p)$$

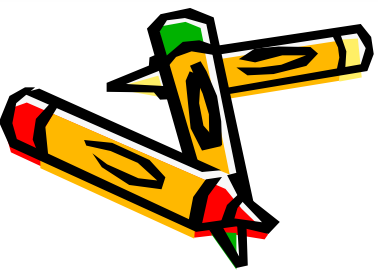
este maximă.



Problema concretă: a estima direct din date această probabilitate, în vederea maximizării sale.

- Se calculează probabilitățile posterioare $P(\Omega_i \mid A_1, A_2, \ldots, A_p)$ pentru toate clasele Ω_i, utilizând formula:

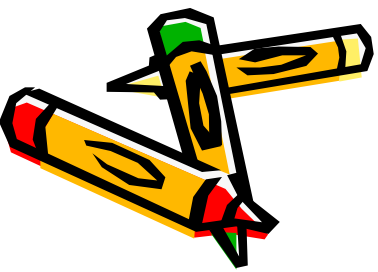
$$P(\Omega_i \mid A_1, A_2, \ldots, A_p) = \frac{P(A_1, A_2, \ldots A_p \mid \Omega_i) \cdot P(\Omega_i)}{P(A_1, A_2, \ldots A_p)}$$



Se alege apoi clasa Ω_j care maximizează

$P(\Omega_i \mid A_1, A_2, \ldots, A_p)$.

adică clasa Ω_j care maximizează $P(A_1 A_2 \ldots A_p \mid \Omega_i) \cdot P(\Omega_i)$.



clasificarea naiva Bayes

Naive Bayes presupune, de foarte multe ori fără niciun temei, independența evenimentelor.

În cazul de față, vom presupune independența reciprocă a atributelor. (ipoteză neadevărată de cele mai multe ori) pentru o anumită clasă Ω_i , adică:

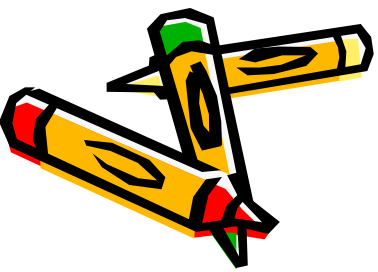
$$P(A_1, A_2, \ldots, A_p \mid \Omega_i) =$$

$$= P(A_1 \mid \Omega_i) \cdot P(A_2 \mid \Omega_i) \cdot \ldots \cdot P(A_p \mid \Omega_i)$$

Vom estima apoi probabilitățile $P(A_k \mid \Omega_i)$ pentru toate atributele A_k și clasele Ω_i, astfel încât un obiect nou, necunoscut, va fi clasificat în clasa Ω_j dacă probabilitatea corespunzătoare acestei clase:

$$P(\Omega_j) \cdot \prod_{k=1}^{p} P(A_k \mid \Omega_j),$$

este maximă față de celelalte.



exemple

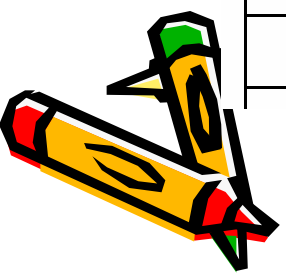
- Domeniul bancar, problema estimării riscului acordării unui credit unei anumite persoane

Folosim clasificarea bayesiană, având în vedere următoarele atribute:

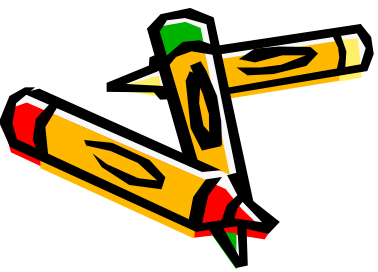
- datoria curentă,
- venit lunar
- garanții.

multime de antrenament

client	RISC	datorii	garanții	venit lunar
1	înalt	multe	nu există	850 RON
2	înalt	multe	nu există	1000 RON
3	înalt	puține	nu există	600 RON
4	înalt	puține	nu există	500 RON
5	scăzut	puține	nu există	1800 RON
6	înalt	puține	adecvate	500 RON
7	înalt	puține	nu există	700 RON
8	scăzut	puține	nu există	1600 RON
9	scăzut	puține	nu există	2800 RON



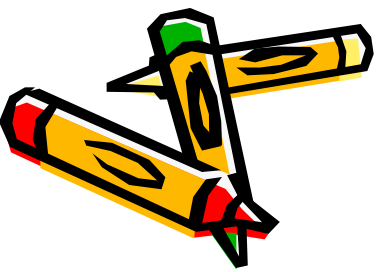
10	scăzut	multe	adecvate	1100 RON
11	înalt	multe	nu există	500 RON
12	înalt	multe	nu există	600 RON
13	scăzut	multe	nu există	1600 RON
14	înalt	multe	nu există	1400 RON
15	înalt	multe	adecvate	450 RON
16	înalt	puține	nu există	700 RON
17	scăzut	puține	adecvate	1200 RON
18	scăzut	puține	adecvate	3200 RON
19	scăzut	puține	adecvate	1100 RON
20	înalt	multe	nu există	400 RON



Există două clase distincte:
risc înalt și risc scăzut din punctul de vedere al riscului acordării unui credit
Probabilitățile celor două clase sunt:

$$P(\text{risc înalt}) = \frac{12}{20}$$

$$P(\text{risc scăzut}) = \frac{8}{20}.$$

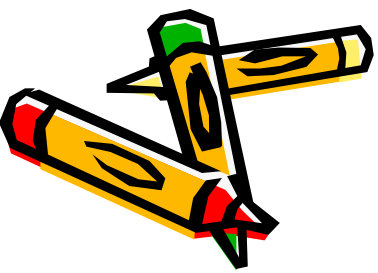


Probabilitățile condiționate de tipul $P(A_k \mid \Omega_i)$

- în cazul atributelor discrete, se vor calcula în mod natural după formula:

$$P(A_k \mid \Omega_i) = \frac{|A_{ki}|}{N_{\Omega_i}},$$

$|A_{ki}|$ reprezintă numărul instanțelor având atributul A_k și care aparțin clasei Ω_i.

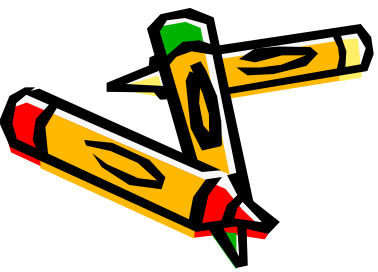


$$P(\text{datorii = puține } | \text{risc înalt}) = \frac{5}{12}$$

$$P(\text{datorii = puține } | \text{risc scăzut}) = \frac{6}{8}$$

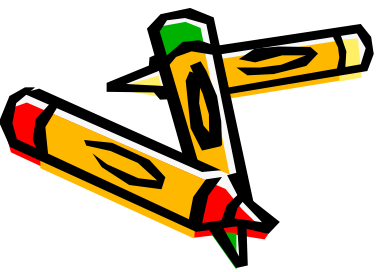
$$P(\text{garanții = adecvate} | \text{risc înalt}) = \frac{2}{12}$$

$$P(\text{garanții = adecvate} | \text{risc scăzut}) = \frac{4}{8}.$$



În cazul atributelor de tip continuu, pentru a evalua probabilitățile condiționate $P(A_k \mid \Omega_i)$, este nevoie de identificarea tipului de repartiție a atributului, privit ca variabilă aleatoare continuă.

De obicei, se presupune că toate atributele continue urmează legea normală, urmând ca din date să se estimeze parametrii acesteia (media și dispersia).



Odată densitatea de repartiție estimată, putem evalua probabilitatea condiționată $P(A_k \mid \Omega_i)$ pentru fiecare clasă în parte.

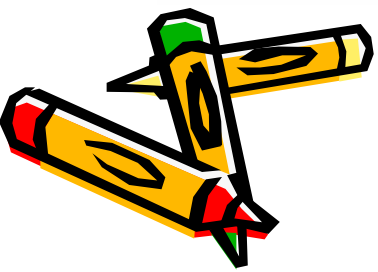
Atributul *Venit lunar* este considerat variabilă aleatoare continuă, de densitate:

$$P(A_k \mid \Omega_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{ki}} \cdot \exp\left(- \frac{(A_k - \mu_{ki})^2}{2 \cdot \sigma_{ki}^{\ 2}} \right)$$

Să analizăm acum modul de funcționare a clasificatorului astfel construit pe un caz nou:
unui individ care are următoarele atribute:

- datorii puține
- garanții adecvate
- venit lunar 2000 RON

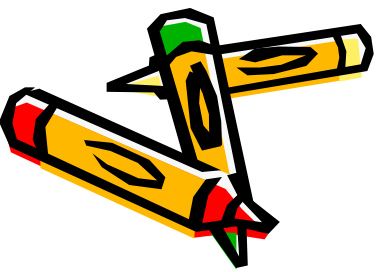
îi acordăm sau nu credit.



$$P(\text{venit lunar} = 2000 \mid \text{risc inalt})$$

$$= \frac{1}{\sqrt{2\pi} \cdot s1} \cdot \exp\left(- \frac{(2000 - m1)^2}{2 \cdot v1} \right) = 3.1803 \cdot 10^{-8}.$$

$$P(\text{venit anual} = 2000 \mid \text{risc scăzut}) = 4.8847 \cdot 10^{-4}$$



$$P(\text{datorii puține, garanții adecvate, venit} = 2000 | \text{risc înalt}) \times$$

$$\times P(\text{risc înalt}) = P(\text{datorii} = \text{puține} | \text{risc înalt}) \times$$

$$\times P(\text{garanții} = \text{adecvate} | \text{risc înalt}) \times$$

$$\times P(\text{venit anual} = 2000 | \text{risc inalt}) \times P(\text{risc înalt}) = 1.3251 \cdot 10^{-9}$$

$$P(\text{datorii puține, garanții adecvate, venit} = 2000 | \text{risc scăzut})$$

$$\times P(\text{risc scăzut}) = P(\text{datorii} = \text{puține} | \text{risc scăzut})$$

$$\times P(\text{garanții} = \text{adecvate} | \text{risc scăzut}) \times$$

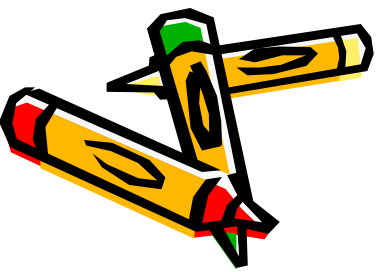
$$\times P(\text{venit anual} = 2000 | \text{risc scăzut}) \times P(\text{risc scăzut}) = 7.3271 \cdot 10^{-5}$$

Deoarece

P(datorii puține, garanții adecvate, venit = 2000 |risc scăzut)>

> P(datorii puține, garanții adecvate, venit = 2000 |risc înalt)

rezultă că banca poate să-i acorde împrumutul solicitat.



Considerând un individ care are următoarele atribute:

- datorii puține
- garanții adecvate
- venit lunar 600 RON

vom avea:

$$P(\text{datorii puține, garanții adecvate, venit} = 600 \mid \text{risc înalt}) \times$$

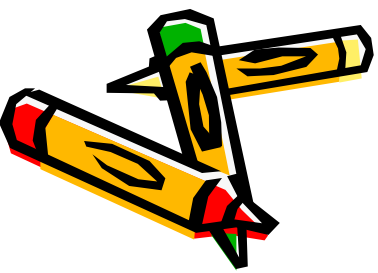
$$\times P(\text{risc înalt}) = 5.5936 \cdot 10^{-5}$$

$$P(\text{datorii puține, garanții adecvate, venit} = 600 \mid \text{risc scăzut}) \times$$

$$\times P(\text{risc scăzut}) = 2.3937 \cdot 10^{-5}$$

clasificarea naiva a textelor

Prin ipoteză, cuvintele ce apar în text se consideră a fi independente.



clasificarea documentelor

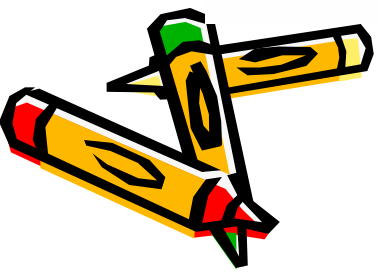
- ***Clasificarea ierarhică*** - folosită dacă se urmăreşte o analiză în detaliu a datelor. Această metodă corespunde aşa numitului „hard clustering", adică se acceptă o singură posibilitate de apartenenţă la o clasă (categorie).
- ***Clasificarea non- ierarhică*** - metodă mult mai rapidă, utilizată pentru baze mari de date. Această metodă este de tip „soft clustering", în sensul că în loc să accepte apartenenţa la o singură clasă, furnizează probabilitatea de apartenenţă a unui element la o anumită clasă (fiecare clasă va avea o anumită pondere de apartenenţă la ea).

clasificarea bayesiana naiva (soft clustering)

Să presupunem că avem r categorii (clase) de documente $\Omega = \{\Omega_1, \ldots, \Omega_r\}$.

A determina cărei categorii îi corespunde documentul D înseamnă a estima probabilitatea $P(\Omega_i \mid D)$ de apartenență a documentului D la clasa Ω_i, utilizând formula Bayes:

$$P(\Omega_i \mid D) = \frac{P(D \mid \Omega_i) \cdot P(\Omega_i)}{P(D)}.$$



$P(D \mid \Omega_i)$ este probabilitatea ca fiind dată clasa Ω_i, cuvintele din D să fie asociate cu această clasă.

Dacă documentul D este format din cuvintele $w_1,\ldots,w_m$, $P(D \mid \Omega_i)$ este probabilitatea apariției fiecărui cuvânt $w_j, 1 \le j \le m$ în clasa Ω_i, probabilitate ce poate fi calculată pe baza ipotezei de independență a variabilelor:

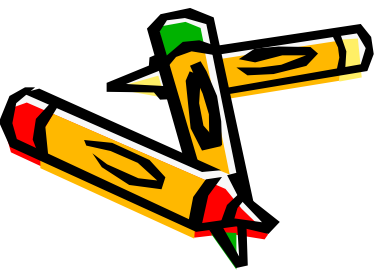
$$P(D \mid \Omega_i) = P(w_1 \mid \Omega_i) \cdot \ldots \cdot P(w_m \mid \Omega_i)$$

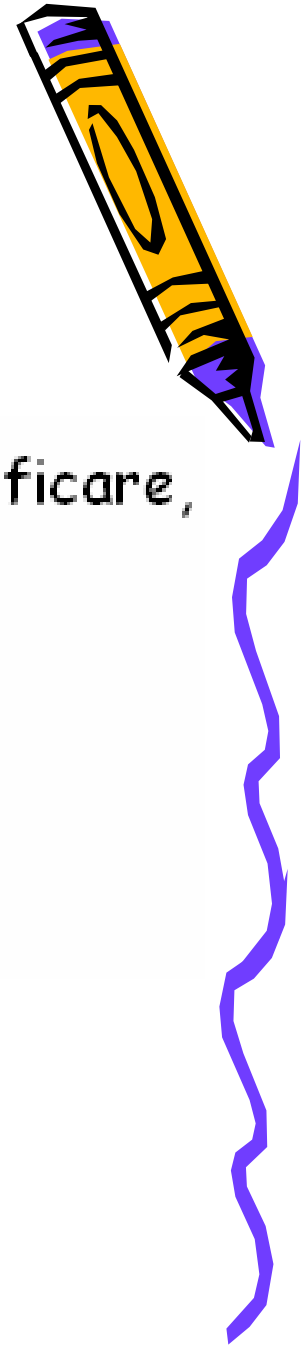
Notând cu n_{ij} numărul de apariții ale cuvântului ω_j în clasa Ω_i și cu n_i numărul de cuvinte din clasa Ω_i, calculăm:

$$P(\omega_j \mid \Omega_i) = \frac{n_{ij}}{n_i}$$

Notând cu n numărul total de cuvinte din Ω, avem:

$$P(\Omega_i) = \frac{n_i}{n}.$$



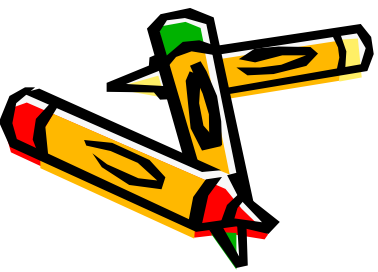


Există un program în MATLAB pentru această clasificare, program pentru se construiesc funcțiile:

parsefile,

addwords,

classify

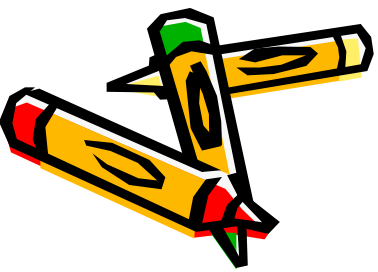


parsefile

- parsefile permite extragerea conținutului util dintr-un document.
  Funcția analizează textul, extrage cuvintele, ce vor fi puse în wordsfiles, însoțite de numărul lor de apariții în documentul *D*.
  Se obține un tabel cu cuvintele din document și cu numărul lor de apariții (valorile asociate lor).
  Funcția este folosită atât în timpul antrenamentului, cât și în clasificarea propriu-zisă.

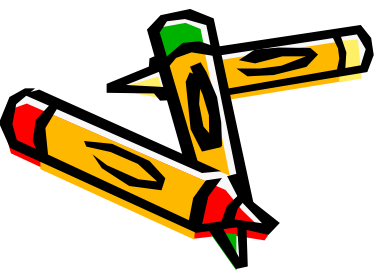
adwords

- addwords permite să adăugăm cuvinte şi valorile asociate lor. Este utilizată la antrenarea clasificatorului



classify

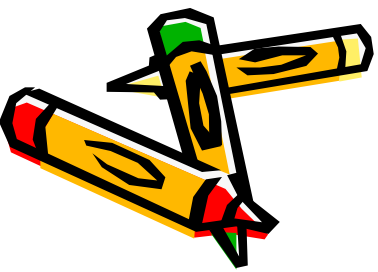
- classify face propriu-zis clasificarea, apelând la parsefile şi calculează toate probabilităţile ce apar în formula Bayes.
  Relaţia: $P(D \mid \Omega_i) = P(\omega_1 \mid \Omega_i) \cdot \ldots \cdot P(\omega_m \mid \Omega_i)$
  se logaritmează, transformând astfel produsul în sumă:
  $\log P(D \mid \Omega_i) = \log(P(\omega_1 \mid \Omega_i)) + \ldots + \log(P(\omega_m \mid \Omega_i))$
  Astfel se reduc calculele ce le va face clasificatorul şi se limitează erorile de "overflow".



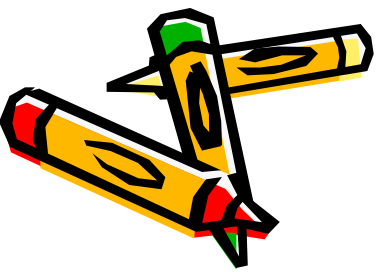
Într-o primă etapă cuvintele care nu apar în nicio clasă nu vor fi luate în considerare în calcule; unui asemenea cuvânt nu i se atribuie valoarea 0 (numărul de apariții) ci 0.1.

Antrenamentul este faza cea mai importantă, performanțele clasificatorului fiind determinate de gradul de învățare.

O temă interesantă ar fi adaptarea şi eventual îmbunătăţirea acestui program pentru texte în limba română, considerând clasele: politic, economic, administrativ, internaţional, sport, arte etc.



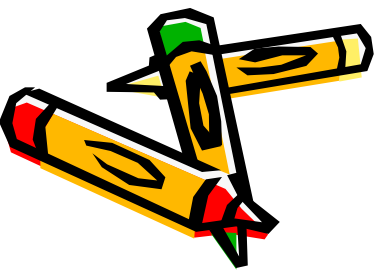
Această tehnică a fost utilizată pentru blocarea spam-urilor. Graham P., (http./www. paulgraham.com/better.html) folosind aceasta metodă, cu câteva modificări, a reuşit să stopeze 99.5% din spam-uri cu o eroare de clasificare mai mică de 0.03%.



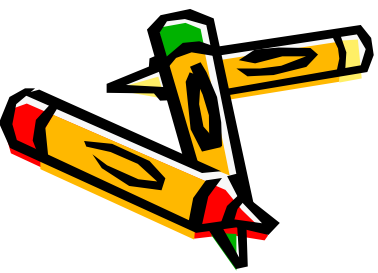
avantaje

Clasificarea bayesiană (naivă) prezintă o serie de avantaje:

- Este robustă în ceea ce privește izolarea zgomotului din date;
- În cazul valorilor lipsă, ignoră obiectul respectiv în timpul estimării probabilităților;
- Este robustă la atributele irelevante.

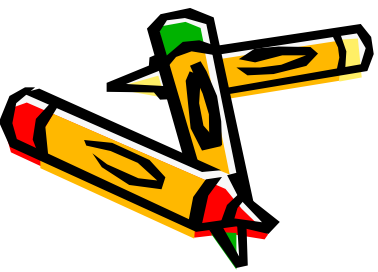


Arbori de clasificare şi decizie



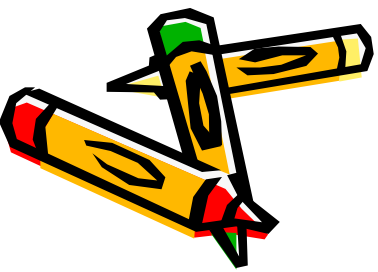
Putem face o clasificare, punând un șir de întrebări, șir în care fiecare întrebare este formulată în funcție de răspunsul primit la precedenta.

Acest procedeu merită a fi folosit în cazul datelor non-metrice, în sensul că de obicei răspunsurile la întrebări vor fi da/nu, adevărat/fals, proprietatea aparține sau nu unei mulțimi de proprietăți etc.



arbore de clasificare şi decizie

Setul de întrebări referitoare la atributelor obiectelor ce urmează a fi clasificate se reprezintă printr-un ***arbore de clasificare şi decizie*** , care este un arbore în sens informatic.

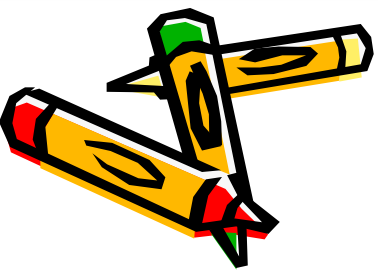


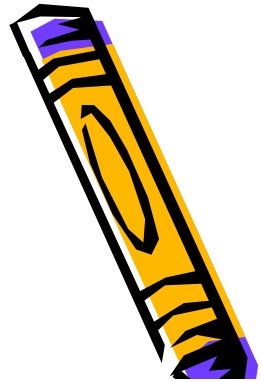
nodul radacina, ramuri, noduri de decizie

Prin convenție, primul nod -***nodul rădăcină***- se află în vârf, legat prin ***ramuri*** (links) de nodurile interne -***noduri de decizie***-.

Clasificarea unui anumit element (pattern) începe din nodul rădăcină, unde este pusă o anumită întrebare relativă la o proprietate specifică.

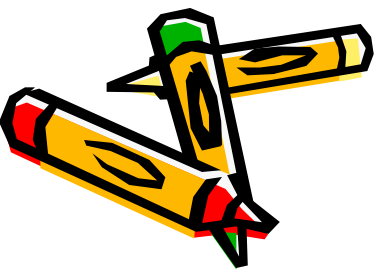
Răspunsurile posibile corespund etichetelor ramurilor.





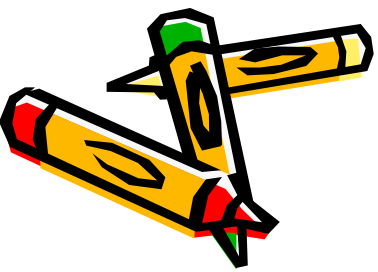
În cazul unor decizii binare, prin convenție, arcul din stânga corespunde unui răspuns afirmativ la test.

În funcție de răspunsuri, care trebuie să fie distincte şi exhausive, urmăm link-ul corespunzător spre un nod descendent, care ar putea fi considerat ca fiind rădăcina unui sub-arbore.



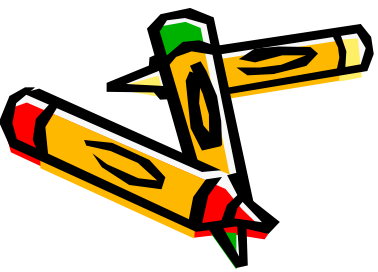
frunze

Continuăm astfel, până la nodurile terminale -***frunze***-, cărora nu le mai corespunde nici o întrebare, și care astfel nu mai au ramuri. Unui nod frunză îi corespunde o anumită categorie (clasă).



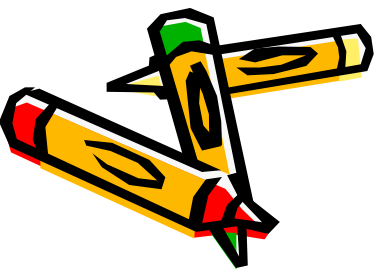
Un arbore de clasificare este utilizat în luarea unei decizii, motiv pentru care este folosită sintagma arbore de *clasificare și decizie*.

Acesta partiționează în mod recursiv mulțimea de antrenament până la obținerea nodurilor finale, care conțin fie numai elemente din aceeași categorie, fie elemente dintr-o categorie dominantă.



Putem interpreta decizia pentru orice clasificare ca fiind suma deciziilor de-a lungul drumului dintre nodul rădăcină și nodul frunză.

Cunoștințele experților umani au deosebită importanță în cazul unei mulțimi de antrenament ce are puține elemente.

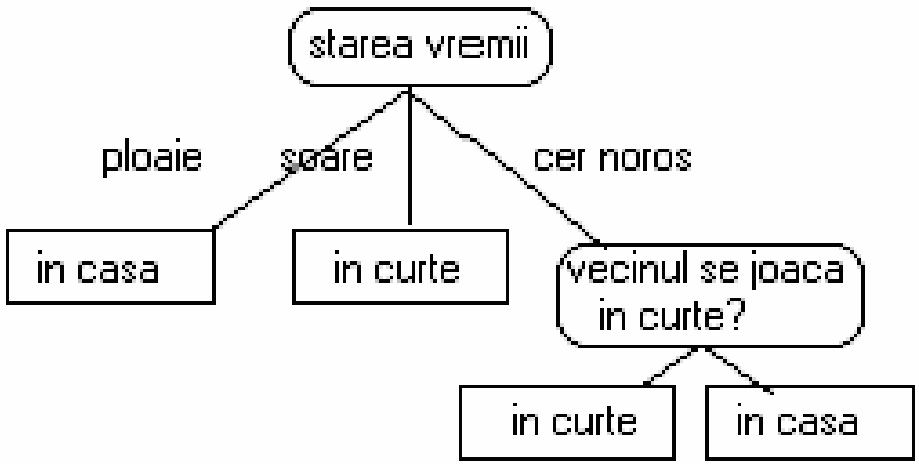


Arborele construit se foloseşte pentru a clasifica exemple necunoscute, în sensul de a decide dacă acestea aparţin sau nu unei anumite clase.

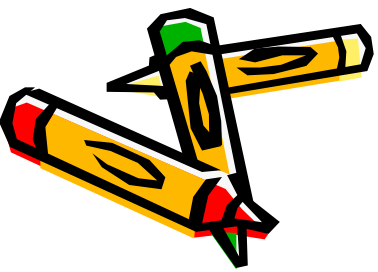
Un arbore de clasificare şi decizie poate fi interpretat ca o reprezentare grafică a unui procedeu de clasificare, nodurile interne fiind testele pentru atribute iar frunzele fiind clasele.

exemplu

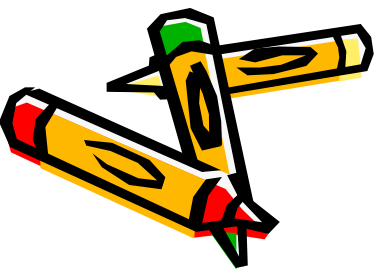
în ce condiții meteo ți se permite, copil fiind, să te joci în curte?



Utilizarea arborelui de clasificare și decizie este indicată nu numai pentru buna clasificare a rezultatelor ci și pentru luarea unor decizii optime prin obținerea unor reguli ușor de înțeles și explicat.

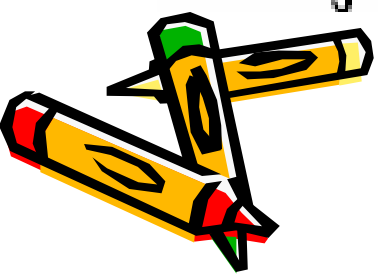


O regulă este creată coborând din vârf -nodul rădăcină- până la fiecare frunză şi este de tipul IF-THEN.
Orice pereche de valori ale unui atribut de-a lungul acestui traseu va forma o conjuncție în ipoteza regulii, iar frunza conținând clasa predictivă va forma consecința regulii .



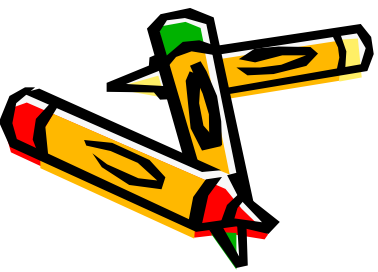
in exemplul dat avem:

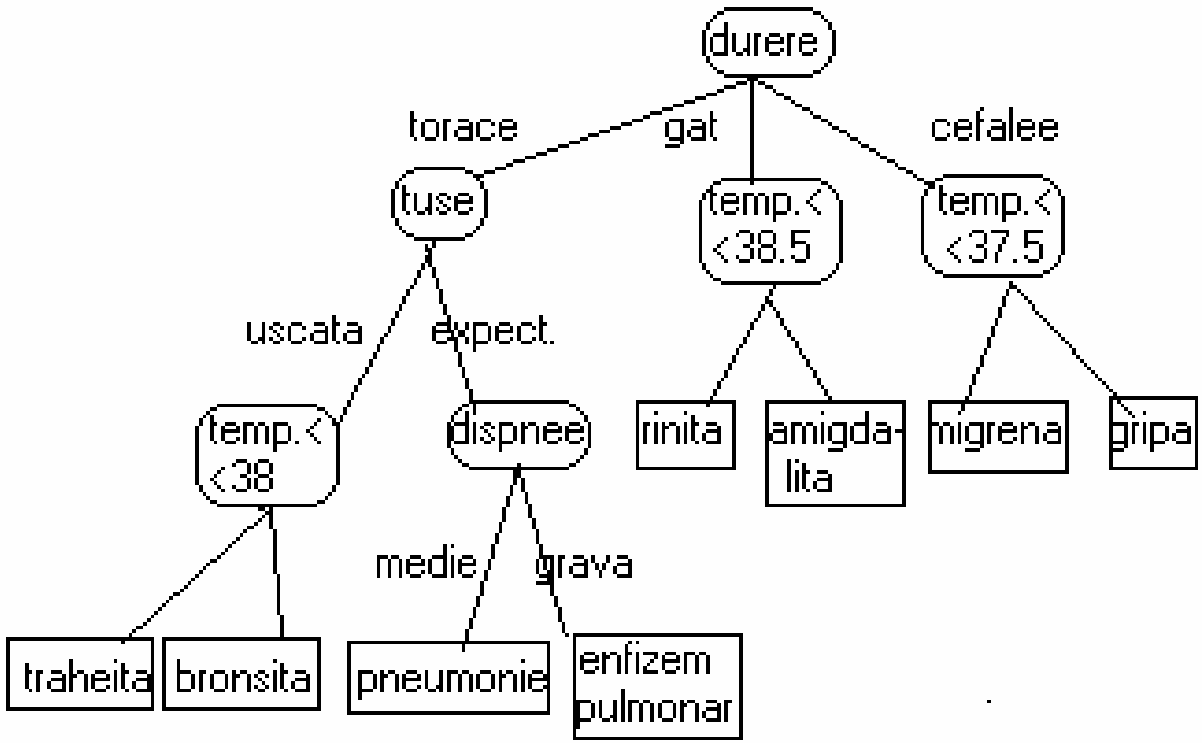
- dacă plouă, rămâi să te joci în casă;
- dacă e soare te joci în curte;
- dacă cerul e noros și copilul vecin se joacă în curte, te joci în curte;
- dacă cerul e noros și copilul vecin nu se joacă în curte, te joci în casă.



exemplu

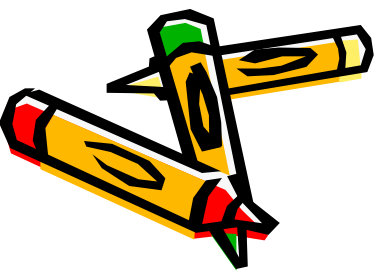
Un medic construieşte următorul arbore de clasificare pentru diagnosticarea unor boli.
Proprietățile, în acest caz simptoamele sunt: durere, tuse, stare febrilă, dispnee (respirație dificilă) .




durere
torace
gat
cefalee
tuse
temp.<
<38.5
temp.<
<37.5
uscata
expect.
temp.<
<38
dispnee
rinita
amigda-
lita
migrena
gripa
medie
grava
traheita
bronsita
pneumonie
enfizem
pulmonar

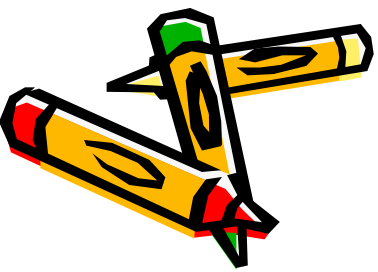
reguli de clasificare

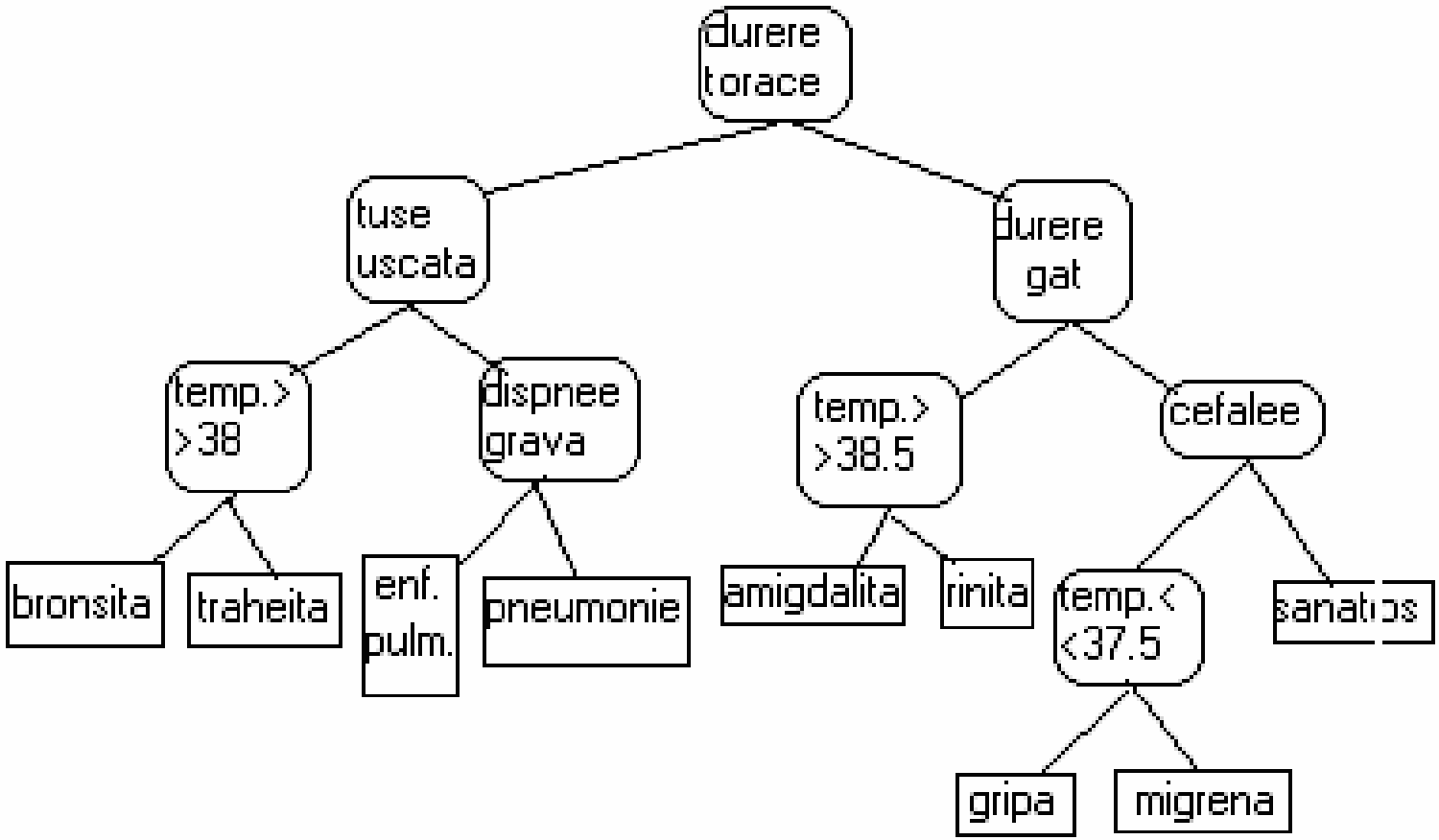
- Pattern-ul {*durere torace, tuse uscată, temperatură* >38} este clasificat ca fiind *bronşită* (traheo - bronşită)
- Pattern-ul {*durere torace, tuse expectorantă, dispnee gravă*) este clasificat ca fiind *enfizem pulmonar*
- Pattern-ul {*cefalee, temperatură* < 37,5} este clasificat ca fiind *migrenă*.



Un arbore oarecare poate fi reprezentat printr-un arbore binar echivalent, prin modificări minime ale întrebărilor test.

Prezentăm arborele binar corespunzător arborelui de diagnosticare construit anterior:

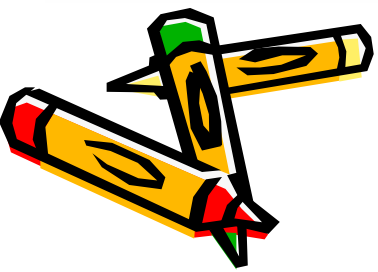



durere
torace
tuse
uscata
durere
gat
temp.>
>38
dispnee
grava
temp.>
>38.5
cefalee
bronsita
traheita
enf.
pulm.
pneumonie
amigdalita
rinita
temp.<
<37.5
gripa
migrena

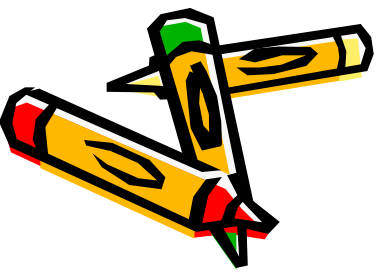
CART

CART (*Classification and Regression Trees, Breiman* 1984) este un algoritm ce constă în construcția arborelui folosind datele din mulțimea de antrenament.

Presupunem că avem o mulțime de date X, în care elementele sunt deja clasificate și cunoaștem proprietățile care fac distincția între clase.

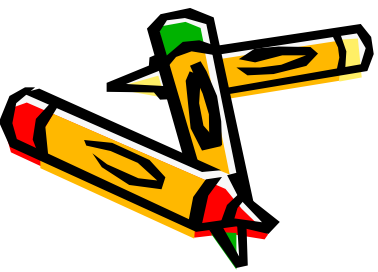


Construind acest arbore, mulțimea de antrenament va fi descompusă în submulțimi din ce în ce mai mici.
Fiecare ieșire (outcome) dintr-un nod se numește ***split*** și corespunde unei descompuneri a mulțimii de antrenament.
Rezultatul ideal ar fi ca să obținem o submulțime ce conține numai elemente ale aceleași clase, o submulțime pură, adică o ***frunză***.



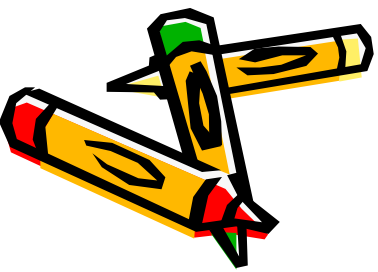
De cele mai multe ori submulțimile conțin și elemente din alte clase.

Ce facem:
acceptăm o decizie imperfectă sau
continuăm construcția arborelui, luând în considerare altă proprietate?

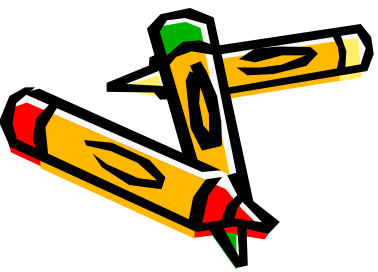


In construirea arborelui urmărim:

- Acuratețea clasificării ;
- Abilitatea explicării motivului luării unei decizii

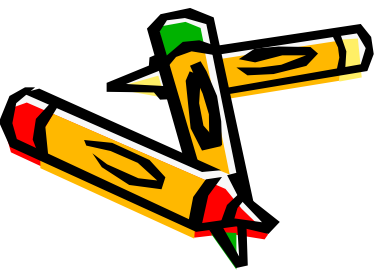


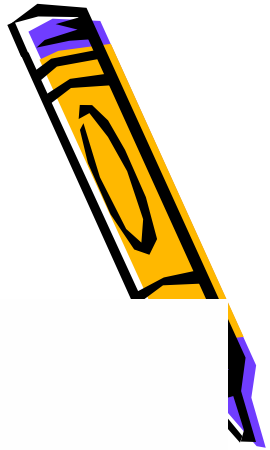
- se construieşte arborele de clasificare şi decizie pe baza mulțimii de antrenament (mulțime de date cunoscute) ;
- se utilizează arborele pentru a clasifica exemple necunoscute, în sensul de a decide cărei clase aparțin.



In construcția arborelui avem de rezolvat o serie de probleme:

- Câte ieşiri să avem la un nod?
- Ce proprietate să testăm la un nod?

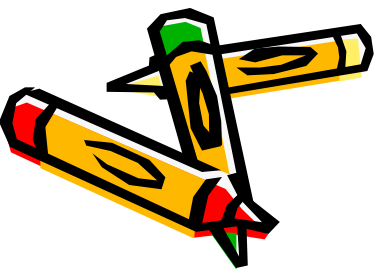




Ne vom ocupa mai mult de cazul arborilor binari.

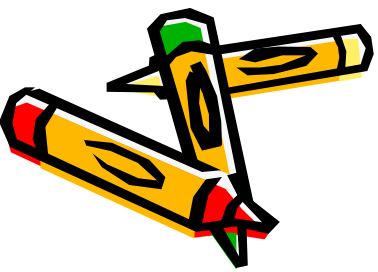
Vom studia doar cazul în care fiecare întrebare se referă la o singură proprietate, cazul arborelui „*monothetic*".

Ideal este să construim un arbore cât mai simplu, cu noduri puține, și în acest scop întrebarea din nodul N trebuie formulată în așa fel încât nodul $N+1$ să fie cât mai „pur".



masura de impuritate

Din punct de vedere matematic este mai simplu să definim o ***măsură de impuritate***, notată $i(N)$, care va fi nulă dacă toate elementele din nod aparțin aceleași clase și va fi maximă dacă în nod avem număr egal de elemente din fiecare clasă.

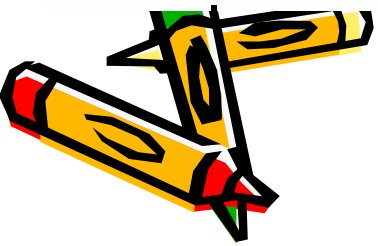


masura entropiei

Considerând că în mulțimea de antrenament există clasele $\Omega_1,\ldots,\Omega_r$, notăm cu $P(\Omega_j)$ raportul dintre numărul de elemente din nodul N aparținând clasei Ω_j și numărul de elemente din nodul N și putem defini:

măsura (*funcția*) ***entropiei*** în nodul N, dată de:

$$i(N) = -\sum_j P(\Omega_j).\log_2 P(\Omega_j);$$

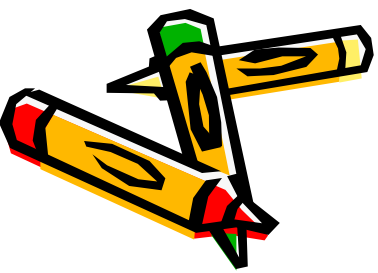


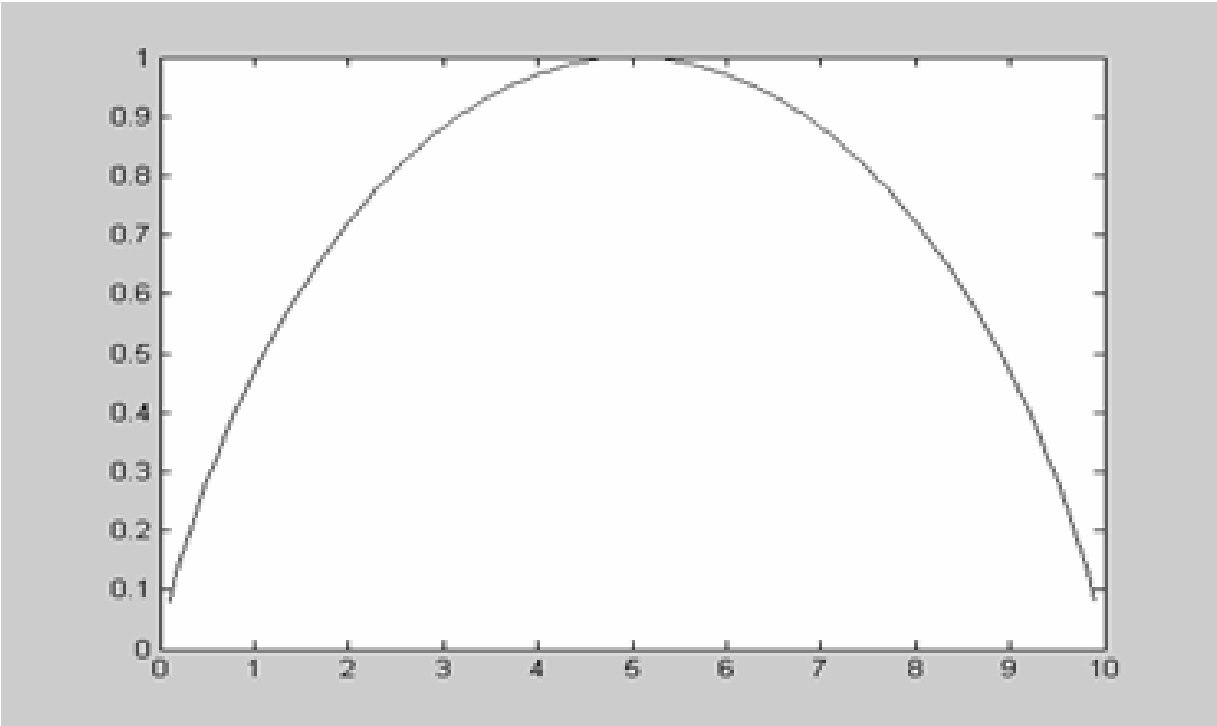
În cazul a două clase Ω_1 și Ω_2 avem:

$$i(N) = -P(\Omega_1)\log_2 P(\Omega_1) - P(\Omega_2)\log_2 P(\Omega_2)$$

dacă în nodul N avem n elemente aparținând claselor Ω_1 și Ω_2, dintre care x în clasa Ω_1, măsura entropiei în nod poate fi considerată a fi funcție de x:

$$f(x) = -\frac{x}{n}\log_2\frac{x}{n} - \left(1-\frac{x}{n}\right)\cdot\log_2\left(1-\frac{x}{n}\right)$$





graficul funcției în cazul $n = 10$.

Valoarea maximă a măsurii entropiei în acest caz este egală cu 1.

informatia castigata

Problema de rezolvat: ce întrebare a testului punem în nodul N?
Considerăm cazul unui arbore binar.
Prin divizarea nodului N, ce are n elemente, folosind atributul A, obținem nodurile N_D și N_S ce au n_D și respectiv n_S elemente.
Informația câștigată prin partiționare este:

$$gain(A) = i(N) - \left(\frac{n_D}{n} \cdot i(N_D) + \frac{n_S}{n} \cdot i(N_S) \right)$$

informatie scontata

Termenul

$$E(A) = \frac{n_D}{n} \cdot i(N_D) + \frac{n_S}{n} \cdot i(N_S)$$

este cunoscut sub numele de *informația scontată*.

Suntem interesați ca informația câştigată să fie cât mai mare, acesta fiind criteriul alegerii atributului pentru partiționare.

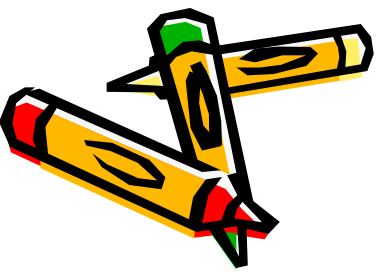
exemplu

- În perioada de vârf a virozelor respiratorii, elevii unei școli pot fi împărțiți în două categorii: bolnavi și sănătoși.

Descrierea va fi făcută pe baza a două atribute:

temperatura, atribut numeric

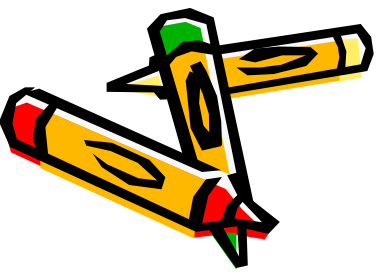
gât iritat, atribut nominal.



Considerăm un eşantion de 300 de pacienţi dintre care 200 sunt sănătoşi,

	gât iritat	gât normal
temperatura < 37.5	6 S , 37 B	191 S, 1 B
temperatura > 37.5	2 S, 21 B	1 S, 41 B

B (bolnavi) şi S (sănătoşi)



- calculăm entropia stării de sănătate:

$$i(stare) = -\frac{200}{300}\log_2\frac{200}{300} - \frac{100}{300}\log_2\frac{100}{300} = 0.9183\,;$$

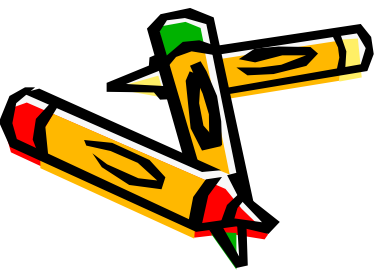
- calculăm entropia pentru temperatura > 37.5, respectiv temperatura < 37.5:

$$i(temp > 37,5) = -\frac{3}{65}\log_2\frac{3}{65} - \frac{62}{65}\log_2\frac{62}{65} = 0.2698\,,$$

$$i(temp < 37,5) = -\frac{197}{235}\log_2\frac{197}{235} - \frac{38}{235}\log_2\frac{38}{235} = 0.6384\,;$$

informația câștigată folosind atributul temperatură:

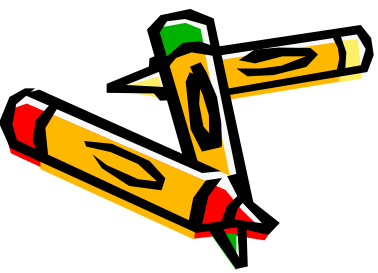
$$gain(temp) = 0.9183 - \frac{235}{300} \cdot 0.6384 - \frac{65}{300} \cdot 0.2698 = 0.3598$$



- calculăm entropia pentru gât iritat, respectiv gât neiritat:

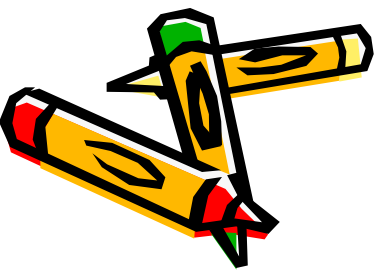
$$i(gat\ iritat) = -\frac{8}{66}\log_2\frac{8}{66} - \frac{58}{66}\log_2\frac{58}{66} = 0.5328\ ,$$

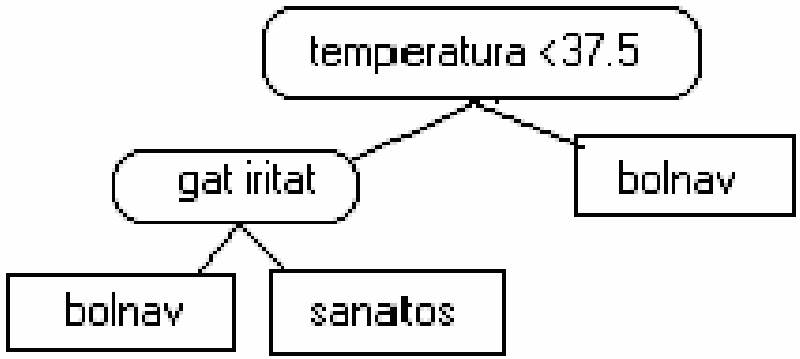
$$i(gat\ neiritat) = -\frac{192}{234}\log_2\frac{192}{234} - \frac{42}{234}\log_2\frac{42}{234} = 0.6790\ ;$$



- informația câștigată folosind atributul starea gâtului:

$$gain(stare\ gat) = 0.9183 - \frac{66}{300} \cdot 0.5328 - \frac{234}{300} \cdot 0.6790 = 0.2715$$

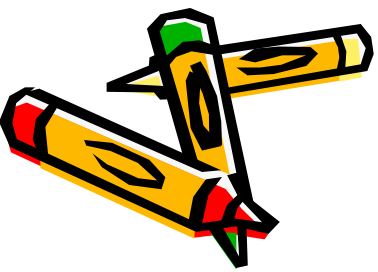




Frunzele nu sunt pure, dar au cea mai mică măsură de impuritate.

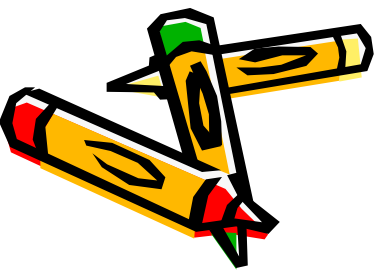
Putem avea impuritate din mai multe motive:

- date incorecte,
- date corecte dar atribute insuficiente,
- clasificarea implică un anumit grad de nedeterminare.



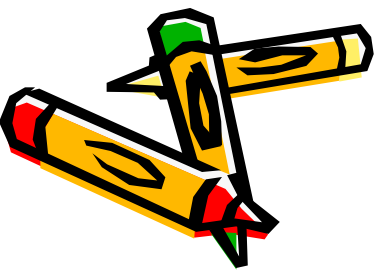
În cazul a trei clase Ω_1, Ω_2 și Ω_3 avem:

$$i(N) = -P(\Omega_1)\log_2 P(\Omega_1) - P(\Omega_2)\log_2 P(\Omega_2) -$$
$$- P(\Omega_3)\log_2 P(\Omega_3)$$

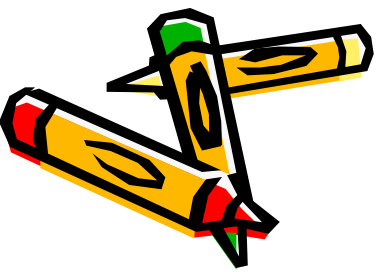


dacă în nodul N avem n elemente aparținând claselor Ω_1, Ω_2 și Ω_3, dintre care x în clasa Ω_1 și y în Ω_2, măsura entropiei în nodul N poate fi considerată a fi funcție de x și y:

$$f(x,y) = -\frac{x}{n}\log_2\frac{x}{n} - \frac{y}{n}\log_2\frac{y}{n} - \left(1-\frac{x+y}{n}\right)\cdot\log_2\left(1-\frac{x+y}{n}\right).$$

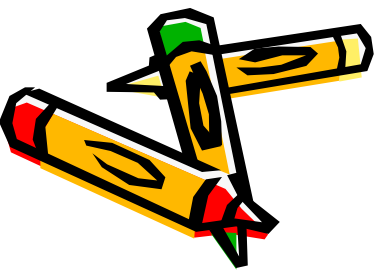


Calculăm extremele acestei funcții și obținem că $\left(\frac{n}{3}, \frac{n}{3}\right)$ este punct de maxim și astfel maximul impurității entropiei este 1.5850



Prin divizarea nodului N, ce are n elemente, folosind atributul A, obținem nodurile $N_1,\ldots,N_p$, ce au $n_1,\ldots,n_p$ elemente. Informația câștigată prin partiționare este:

$$gain(A) = i(N) - \sum_{k=1}^{p} \frac{n_k}{n} \cdot i(N_k)$$



exemplu

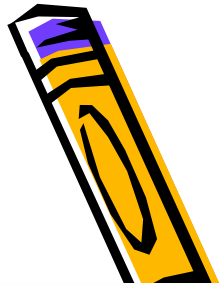
- În domeniul bancar, estimarea riscului acordării unui credit unei anumite persoane:

construim un arbore de clasificare și decizie, având în vedere următoarele proprietăți:

- comportamentul anterior al persoanei când a beneficiat de credite (istoria creditelor),
- datoria curentă,
- venit lunar
- garanții.

multimea de antrenament

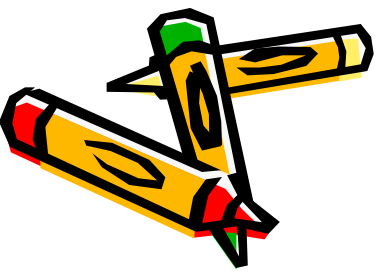
	RISC	Istoria creditelor	Datorii	Garanții	Venit lunar (lei)
1	înalt	proastă	multe	nu există	400-1000
2	înalt	necunoscută	multe	nu există	1000-2000
3	moderat	necunoscută	puține	nu există	1000-2000
4	înalt	necunoscută	puține	nu există	400-1000
5	scăzut	necunoscută	puține	nu există	peste 2000
6	scăzut	necunoscută	puține	adecvate	peste 2000
7	înalt	proastă	puține	nu există	400-1000
8	moderat	proastă	puține	nu există	peste 2000
9	scăzut	bună	puține	nu există	peste 2000



10	scăzut	bună	multe	adecvate	peste 2000
11	înalt	bună	multe	nu există	400-1000
12	moderat	bună	multe	nu există	1000-2000
13	scăzut	bună	multe	nu există	peste 2000
14	înalt	proastă	multe	nu există	1000-2000
15	înalt	necunoscută	multe	nu există	1000-2000
16	moderat	necunoscută	puține	nu există	1000-2000
17	moderat	proastă	puține	adecvate	1000-2000
18	scăzut	necunoscută	puține	adecvate	peste 2000
19	scăzut	bună	puține	adecvate	400-1000
20	înalt	proastă	multe	nu există	400-1000

- calculăm măsura entropiei:

$$i(risc) = -\frac{8}{20}\log_2\frac{8}{20} - \frac{5}{20}\log_2\frac{5}{20} - \frac{7}{20}\log_2\frac{7}{20} = 1.5589\ ;$$



- calculăm câștigul de informație, obținut prin utilizarea atributului *istoria creditelor*, pentru divizarea nodului:

$$i(ist.proasta) = -\frac{4}{6}\cdot\log_2\frac{4}{6} - \frac{2}{6}\cdot\log_2\frac{2}{6} = 0.9183\ ,$$

$$i(ist.necunoscuta) = -\frac{3}{8}\cdot\log_2\frac{3}{8} - \frac{2}{8}\cdot\log_2\frac{2}{8} - \frac{3}{8}\cdot\log_2\frac{3}{8} =$$

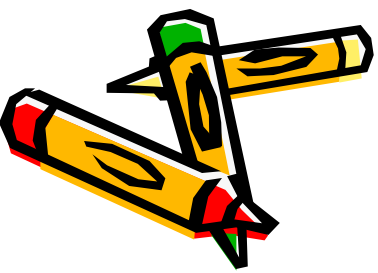
$$= 1.5613,$$

$$i(ist.buna) = -\frac{1}{6}\cdot\log_2\frac{1}{6} - \frac{1}{6}\cdot\log_2\frac{1}{6} - \frac{4}{6}\cdot\log_2\frac{4}{6} = 1.2516\ ,$$

$$gain(ist.credite) = i(risc) - \frac{6}{20} \cdot i(ist.proasta) -$$

$$- \frac{8}{20} \cdot i(ist.necunoscuta) - \frac{6}{20} \cdot i(ist\ buna) =$$

$$= 1.5589 - \left(\frac{6}{20} \cdot 0.9183 + \frac{8}{20} \cdot 1.5613 + \frac{6}{20} \cdot 1.2516 \right) = 0.2834\ ;$$



- calculăm câştigul de informaţie, obţinut prin utilizarea atributului *datorii*:

$$i(\textit{datorii multe}) = -\frac{6}{9}\cdot\log_2\frac{6}{9} - \frac{1}{9}\cdot\log_2\frac{1}{9} - \frac{2}{9}\cdot\log_2\frac{2}{9} = 1.2244,$$

$$i(\textit{datorii putine}) = -\frac{2}{11}\cdot\log_2\frac{2}{11} - \frac{4}{11}\cdot\log_2\frac{4}{11} - \frac{5}{11}\cdot\log_2\frac{5}{11} =$$

$$= 1.4949\ ,$$

$$gain(\textit{datorii}) = 1.5589 - \frac{9}{20}\cdot 1.2244 - \frac{11}{20}\cdot 1.4949 = 0.1857\ ;$$

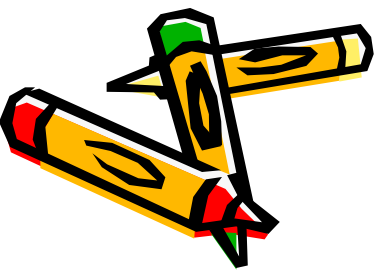
- câștigul de informație pe baza atributului *garanții* :

  $gain(garantii) = 0.1966$;

- câștigul de informație pe baza atributului *venit*:

  $gain(venit) = 0.8120$.

In nodul rădăcină vom avea atributul ***venit***, pentru care am obținut cel mai mare câștig de informație.



Studiem ce atribut utilizăm în subnodul corespunzător celor cu venit între 400-1000 RON, calculând câştigurile de informaţie corespunzătoare:

$$i(venit\ 400-1000)=0.65\ ,$$

$$i(ist\ proasta)=0\ ,\ i(ist.necunoscuta)=0\ ,\ i(ist\ buna)=1\ ,$$

$$gain(ist\ credite)=0.65-\frac{1}{3}=0.3167\ ;$$

Avem $gain(datorii)=0.1909$ şi $gain(garantii)=0.65$ şi astfel în acest subnod atributul ales va fi ***garanţii***.

Pentru cei ce au un venit lunar cuprins între 1000-2000 RON:

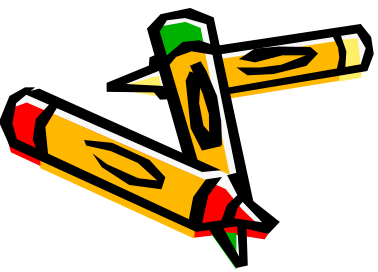
$$i(venit\,1000-2000)=0.9852\ ,$$

$$i(ist\ proasta)=1\ ,\ i(ist.necunoscuta)=0\ ,\ i(ist\ buna)=1\ ,$$

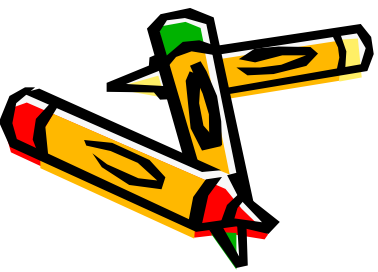
şi astfel $gain(ist\ credite)=0.9852-\frac{4}{7}-\frac{2}{7}=0.1281$.

Având $gain(datorii)=0.5216$ şi $gain(garantii)=0.1281$,

îi vom asocia subnodului atributul *datorii*.



Cei cu datorii puține prezintă un risc moderat pentru bancă, în schimb în cazul celor cu datorii multe, creăm un subnod căruia îi atribuim întrebarea legată de ***istoria creditelor*** avute.

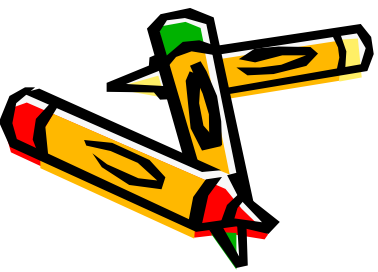


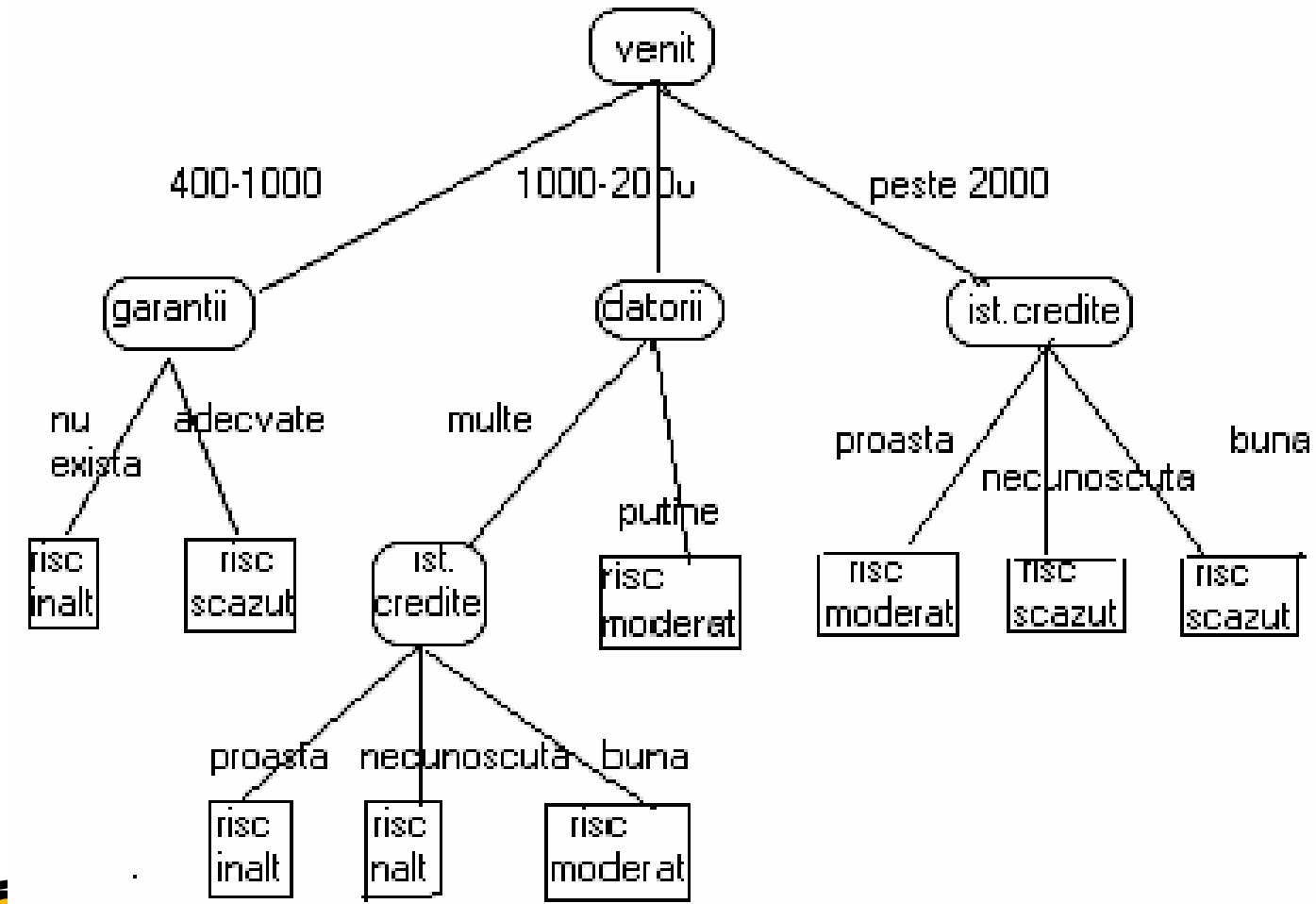
Pentru cei cu venit mai mare de 2000 RON avem următoarele câştiguri de informaţie:

$$gain(ist\ credite) = 0.8120\ , gain(datorii) = 0.4184\ ,$$

$$gain(garantii) = 0.3484\ ,$$

atributul fiind astfel ***istoria creditelor***.




venit
400-1000
1000-2000
peste 2000
garantii
datorii
ist.credite
nu exista
adecvate
multe
putine
proasta
necunoscuta
buna
risc inalt
risc scazut
ist. credite
risc moderat
risc moderat
risc scazut
risc scazut
proasta
necunoscuta
buna
risc inalt
risc nalt
risc moderat

masura de impuritate Gini

Altă metodă de definire a impurității unui nod, "***măsura de impuritate Gini***", dată de:

$$i_G(N) = \sum_{i \neq j} P(\Omega_i) \cdot P(\Omega_j) = \left(\sum_{i=1}^{n} P(\Omega_i) \right)^2 - \sum_{i=1}^{n} P^2(\Omega_i) =$$

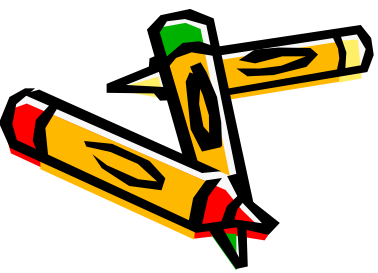
$$= 1 - \sum_{i=1}^{n} P^2(\Omega_i)$$

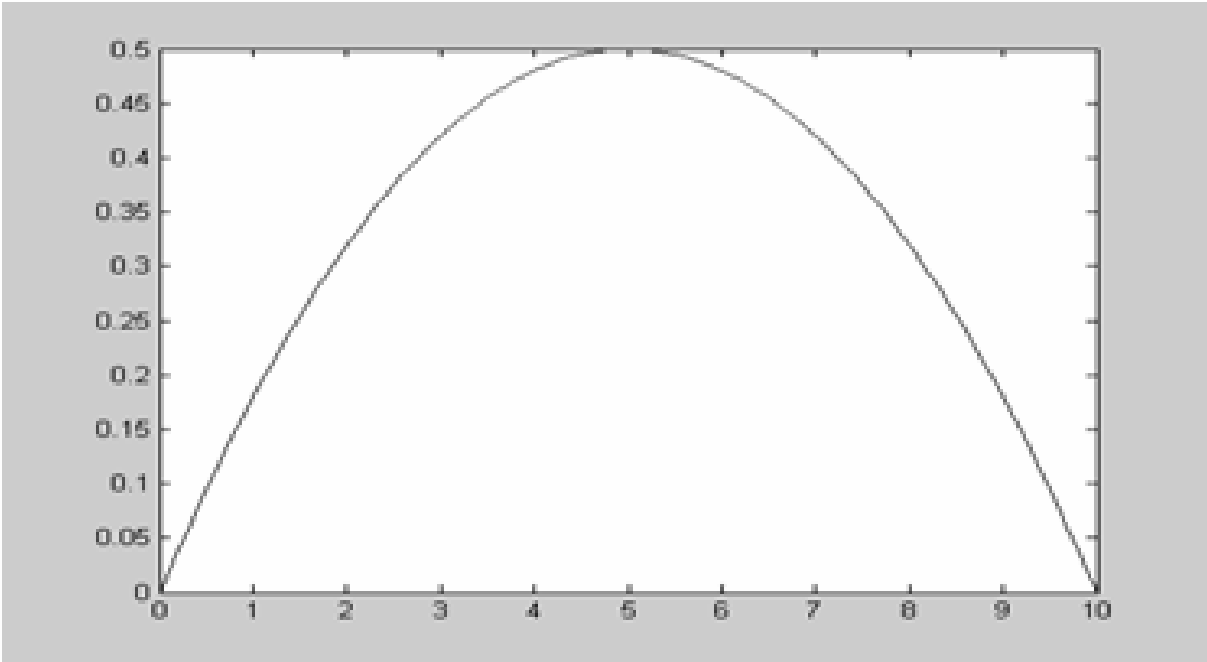
Dacă în mulțimea de antrenament există doar două clase Ω_1 și Ω_2, avem:

$$i_G(N) = 1 - P^2(\Omega_1) - P^2(\Omega_2),$$

și anume dacă în nodul N avem n elemente aparținând claselor Ω_1 și Ω_2, dintre care x din clasa Ω_1, impuritatea Gini în nodul N poate fi considerată a fi funcție de x

$$f(x) = 1 - \left(\frac{x}{n}\right)^2 - \left(1 - \frac{x}{n}\right)^2$$



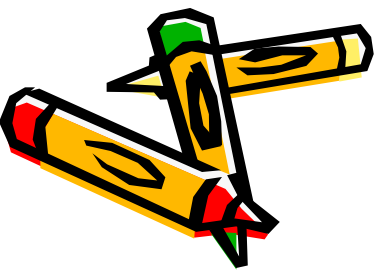


Valoarea maximă a măsurii Gini în acest caz este 1.

Prin divizarea nodului N, ce are n elemente, folosind atributul A, obținem nodurile N_D și N_S ce au n_D și respectiv n_S elemente. Indexul Gini de partiționare este:

$$Gini_{split}(N) = \left(\frac{n_D}{n} \cdot i_G(N_D) + \frac{n_S}{n} \cdot i_G(N_S) \right)$$

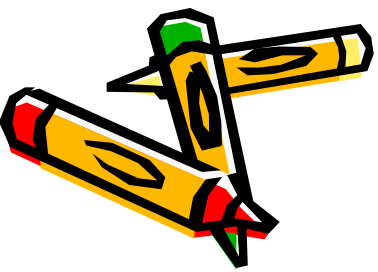
Cea mai mică valoare a indexului Gini de partiționare ne dă acel atribut care minimizează impuritatea divizării,



Considerăm un eşantion de 300 de pacienţi dintre care 200 sunt sănătoşi,

	gât iritat	gât normal
temperatura < 37.5	6 S , 37 B	191 S, 1 B
temperatura > 37.5	2 S, 21 B	1 S, 41 B

B (bolnavi) şi S (sănătoşi)



exemplu

- În exemplul anterior, cu viroze respiratorii, avem:

$$i_G(temp < 37{,}5) = 1 - \left(\frac{197}{235}\right)^2 - \left(\frac{38}{235}\right)^2 = 0.2711\ ,$$

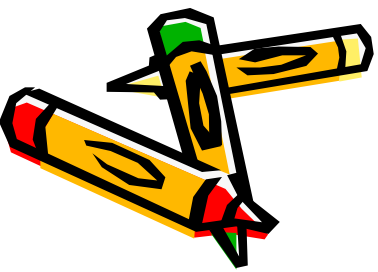
$$i_G(temp > 37{,}5) = 1 - \left(\frac{3}{65}\right)^2 - \left(\frac{62}{65}\right)^2 = 0.0880\ ,$$

$$Gini_{split}(temperatura) = \frac{235}{300} \cdot 0.2711 + \frac{65}{300} \cdot 0.0889 = 0.2314$$

indicele de partiționare Gini pentru *starea de iritare* a gâtului:

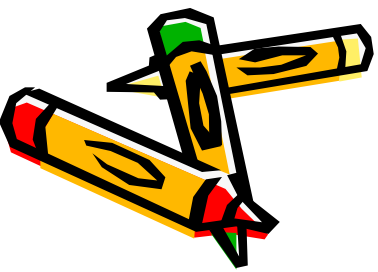
$$Gini_{split}\,(stare\ gat) = \frac{66}{300}\cdot 0.2130 + \frac{234}{300}\cdot 0..2945 = 0.3649\,.$$

utilizând această măsură de impuritate, decidem să facem prima partiționare cu atributul *temperatura*, la fel ca în cazul precedent.



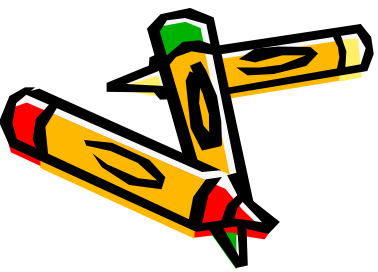
Dacă prin divizarea nodului N, ce are n elemente, folosind atributul A, obținem nodurile $N_1,\ldots,N_p$, ce au $n_1,\ldots,n_p$ elemente, indicele de partiționare Gini pentru atributul A din nodul N este:

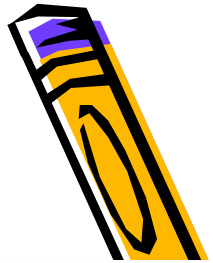
$$Gini_{split}(A) = \sum_{k=1}^{p} \frac{n_k}{n} \cdot i_G(N_k).$$



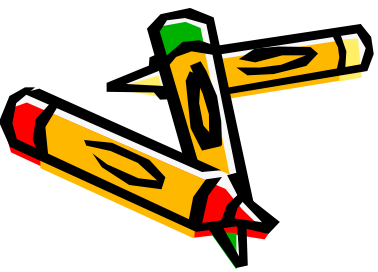
exemplu

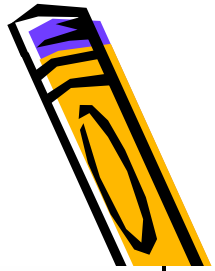
O agenție de turism dorește să realizeze profilul clientului ce alege să petreacă prin intermediul ei concediul în țară sau în străinătate, folosind ca atribute: *vârsta*, *starea civilă* (căsătorit/necăsătorit), *venitul lunar* (sub 1500 lei/peste 1500 lei), *studiile* (superioare/medii).
Ca mulțime de antrenament vom utiliza datele următoare:



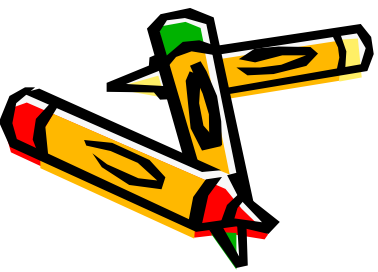


	Destinația	Vârsta	Stare civilă	Venit	Studii
1	țară	27	căsătorit	<1500	medii
2	străinătate	29	necăsătorit	>1500	superioare
3	țară	52	căsătorit	<1500	medii
4	străinătate	58	necăsătorit	>1500	superioare
5	țară	30	necăsătorit	<1500	medii
6	țară	39	căsătorit	<1500	medii
7	țară	60	căsătorit	<1500	medii
8	țară	51	căsătorit	>1500	superioare
9	străinătate	24	necăsătorit	<1500	superioare
10	țară	22	necăsătorit	< 1500	medii





11	străinătate	64	căsătorit	>1500	superioare
12	străinătate	61	căsătorit	> 1500	superioare
13	înstrăinătate	29	căsătorit	> 1500	medii
14	țară	65	căsătorit	<1500	medii
15	țară	45	necăsătorit	< 1500	medii
16	străinătate	32	necăsătorit	>1500	medii
17	străinătate	34	căsătorit	< 1500	superioare
18	străinătate	38	necăsătorit	<1500	medii
19	țară	49	căsătorit	<1500	medii
20	țară	32	necăsătorit	< 1500	medii
21	țară	48	căsătorit	> 1500	superioare



Vom împărți mulțimea pe categorii de vârstă: mai mică, respectiv mai mare decât 25, 35, 45, 55, căutând valoarea optimă de partiționare prin utilizarea indexului Gini.

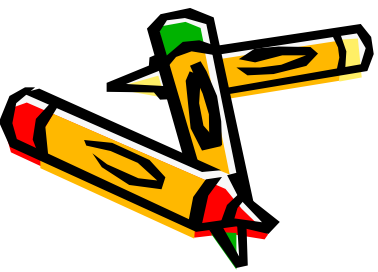
	concediu în țară	concediu în străinătate
vârsta < 25	1	1
vârsta > 25	11	8

$$i_G(virsta < 25) = 0.5\ ;\ i_G(virsta > 25) = 0.4875\ ;$$

$$Gini_{split}(virsta = 25) = \frac{2}{21} \cdot 0.5 + \frac{19}{21} \cdot 0.4875 = 0.4887$$

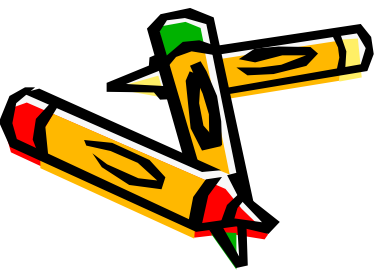
	concediu în țară	concediu în străinătate
vârsta < 35	4	5
vârsta > 35	8	4

$$Gini_{split}\,(virsta = 35) = \frac{9}{21} \cdot 0.4936 + \frac{12}{21} \cdot 0.4444 = 0.4656$$



	concediu în țară	concediu în străinătate
vârsta < 45	6	6
vârsta > 45	6	3

$$Gini_{split}\,(virsta = 45) = \frac{12}{21} \cdot 0.5 + \frac{9}{21} \cdot 0.4444 = 0.4762$$

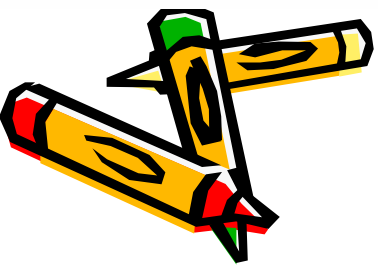


	concediu în țară	concediu în străinătate
vârsta < 55	7	6
vârsta > 55	3	3

$$Gini_{split}\,(virsta = 55) = \frac{13}{21} \cdot 0.4970 + \frac{9}{21} \cdot 0.4688 = 0.4863$$

Cea mai mică valoare a indicelui Gini de partiționare este pentru 35 de ani.
Se ia media aritmetică dintre această valoare și altă valoare de vârstă apropiată (adică 25 ani sau 45 ani), și anume pe cea care are indicele de partiționare mai mic, în cazul nostru 45.

În nodul rădăcină atributul va fi " *vârstă* < 40 ani".



În cazul clienților în vârstă mai mică de 40 ani avem:

	concediu în țară	concediu în străinătate
căsătorit	2	2
necăsătorit	3	4

$$Gini_{split}(stare\ \ civila) = \frac{4}{11} \cdot 0.5 + \frac{7}{11} \cdot 0.4898 = 0.4935\ ;$$

	concediu în țară	concediu în străinătate
peste 1500 RON	0	3
sub 1500 RON	5	3

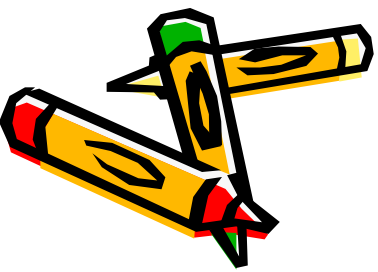
$$Gini_{split}(venit) = 0.3409;$$

	concediu în țară	concediu în străinătate
superioare	0	3
medii	5	3

$$Gini_{split}(studii) = 0.3409$$

Având aceeaşi valoare a indicelui de partiționare Gini pentru atributele *venit*, respectiv *studii*, alegem aleator, să zicem *studii superioare*.

Calculăm indicii de partiționare pentru a decide ce subnod alegem în cazul celor ce au vârsta sub 40 ani şi studii medii:



	concediu în țară	concediu în străinătate
peste 1500 RON	0	2
sub 1500 RON	5	1

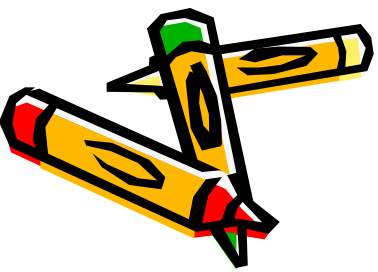
$Gini_{split}\ (venit) = 0.2084\ ;$

	concediu în țară	concediu în străinătate
căsătorit	2	1
necăsătorit	3	2

$Gini_{split}\ (stare\ civila) = 0.4666\ .$

Evident următorul atribut este *venit* < 1500 RON.

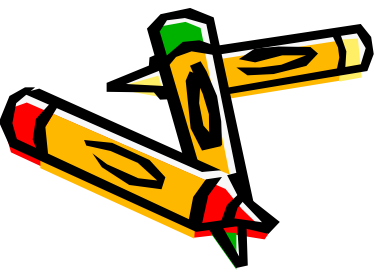
Să vedem dacă atributul *stare civilă* influențează puritatea frunzei "concediu în țară" în cazul clienților cu vârsta sub 40 ani., studii medii și venitul lunar sub 1500 RON.



	concediu în țară	concediu în străinătate
căsătorit	2	1
necăsătorit	3	0

Concluzie (pe baza mulțimii de antrenament):

Persoanele necăsătorite cu vârsta sub 40 ani, studii medii și venitul lunar sub1500 RON aleg să-și petreacă concediul în țară, în schimb doar 66% dintre cei căsătoriți din aceiași categorie aleg aceeași destinație.



- în cazul clienților în vârstă mai mare de 40 ani avem:

	concediu în țară	concediu în străinătate
căsătorit	7	2
necăsătorit	0	1

$$Gini_{split}\ (stare\ civila) = 0.3111\ ;$$

	concediu în țară	concediu în străinătate
peste 1500 RON	2	3
sub 1500 RON	5	0

$$Gini_{split}\ (venit) = 0.24\ ;$$

	concediu în țară	concediu în străinătate
superioare	2	3
medii	5	0

$$Gini_{split}\ (studii) = 0.2400\ .$$

Din nou avem aceeaşi valoare a indicelui de partiționare pentru atributele venit, respectiv studii, să alegem acum atributul venit < 1500 RON.

Din datele prezentate rezultă că aceia cu un asemenea venit lunar aleg să-şi petreacă concediul în țară.

vom calcula indicii de partiționare pentru clienții peste 40 ani, cu venit mai mare de 1500 RON:

	concediu în țară	concediu în străinătate
căsătorit	2	2
necăsătorit	0	1

$$Gini_{split}\ (stare\ civila) = 0.4\ ;$$

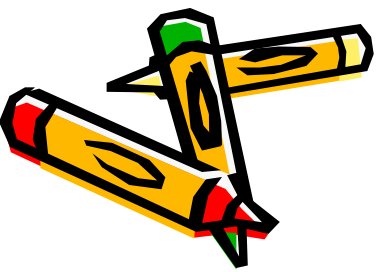
	concediu în țară	concediu în străinătate
superioare	2	3
medii	0	0

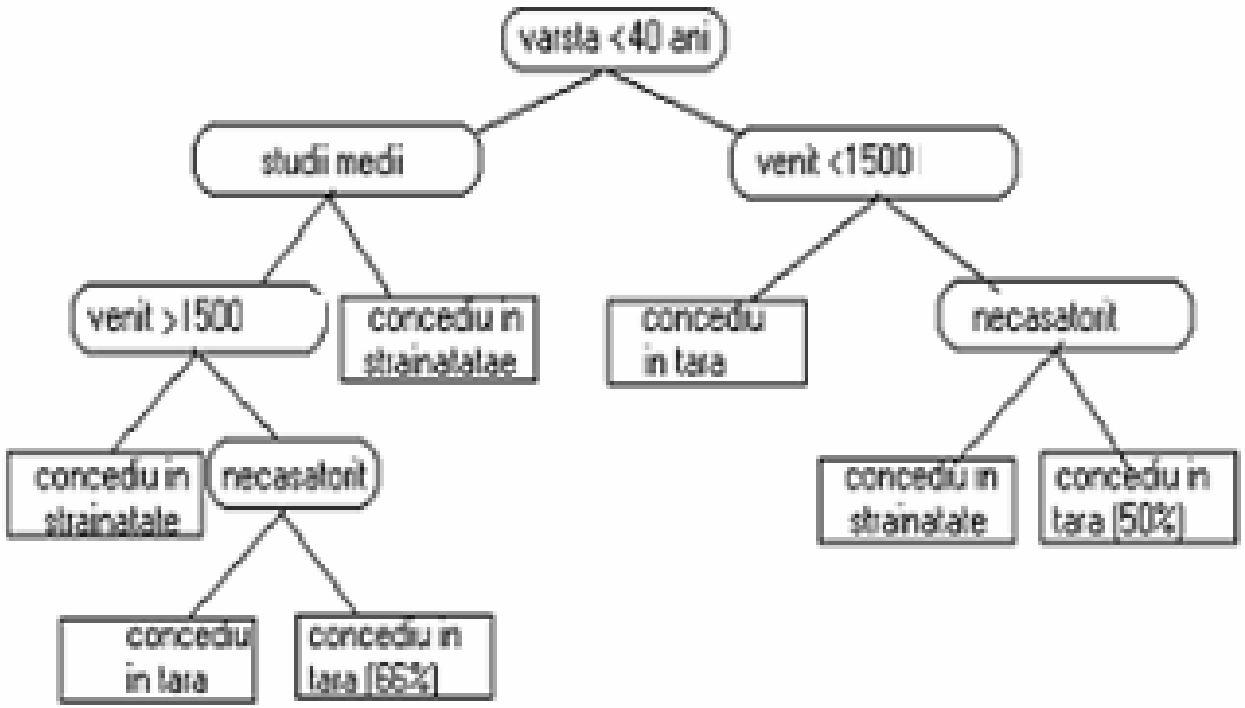
$$Gini_{split}\ (studii) = 0.4444\ .$$

Următorului subnod îi atribuim atributul *stare civilă* și anume necăsătorit.

Nu putem afirma nimic despre destinația de concediu a clienților agenției, în vârstă de peste 40 ani, cu venit mai mare de 1500 RON, căsătoriți, în raport cu nivelul studiilor acestora deoarece jumătate dintre ei aleg să-și petreacă concediul în țară.

În urmă acestui studiu, obținem următorul arbore de clasificare:

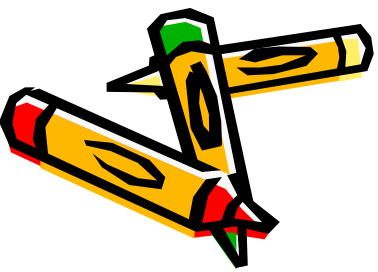



varsta <40 ani
studii medi
venit <1500
venit >1500
concediu in strainatatae
concediu in tara
necasatorit
concediu in strainatate
necasatorit
concediu in strainatate
concediu in tara (50%)
concediu in tara
concediu in tara (66%)

Cu ajutorul arborilor de clasificare şi decizie se pot formula reguli.

În exemplul prezentat, pe baza mulțimii de antrenament, construind arborele de clasificare, putem deduce din regulile obținute profilul clientului agenției, ce își petrece vacanța în țara sau străinătate.

Regulile deduse se bazează pe mulțimea de antrenament, în cazul nostru datele oferite de agenție.

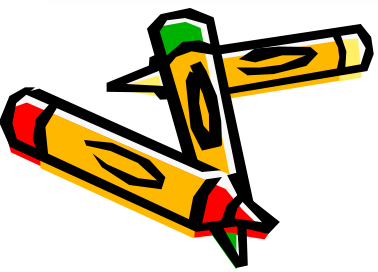


masura clasificarii gresite

măsura clasificării greşite (misclassification) se defineşte prin:

$$i_M(N) = 1 - \max_j P(\Omega_j)$$

Aceasta măsoară probabilitatea minimă ca un element din mulțimea de antrenament să fie greşit clasificat prin folosirea atributului A în nodul N.

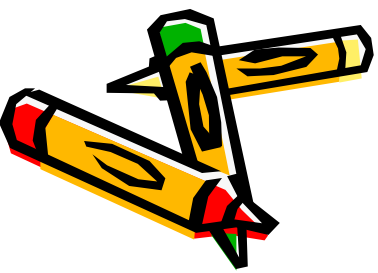


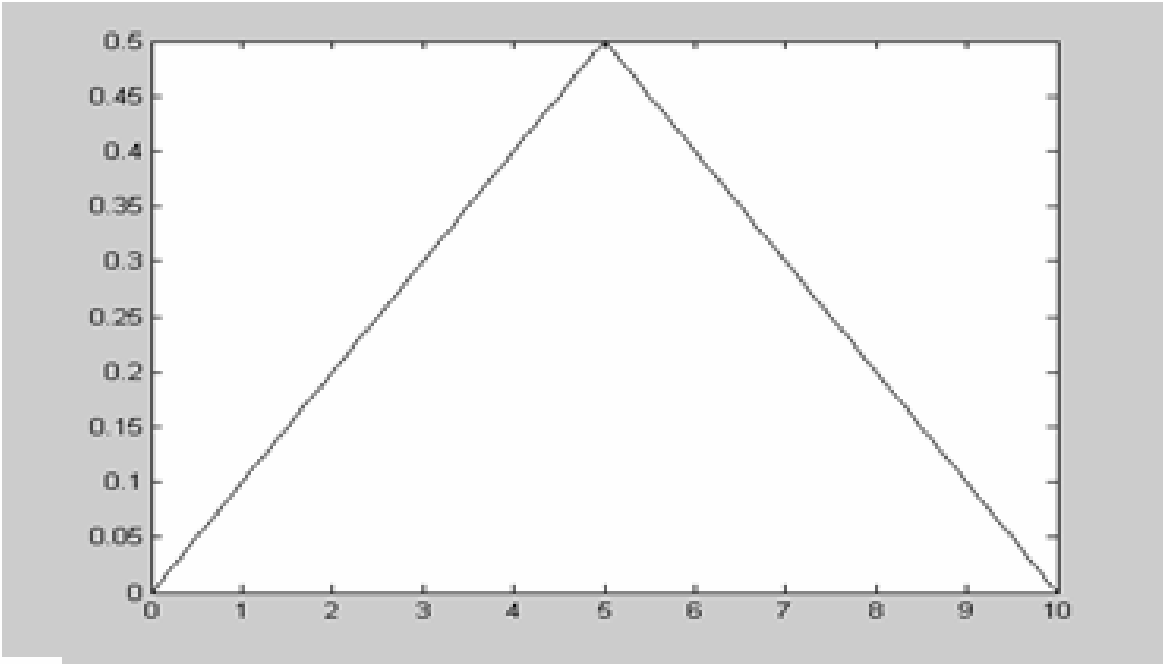
În cazul a două clase avem:

$$i_M(N) = 1 - \max\{P(\Omega_1), P(\Omega_2)\}$$

și anume dacă în nodul N avem n elemente aparținând claselor Ω_1 și Ω_2, dintre care x în clasa Ω_1, impuritatea entropiei în nod poate fi considerată a fi funcție de x:

$$f(x) = 1 - \max\{\frac{x}{n}, 1 - \frac{x}{n}\} = \begin{cases} \frac{x}{n}, & x > \frac{n}{2} \\ 1 - \frac{x}{n}, & x < \frac{n}{2} \end{cases}$$





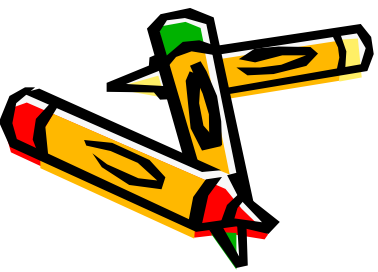
De reținut: cele trei funcții ale impurității, definite în cazul a două clase, au aceeași valoare a abscisei punctului de maxim.

exemplu

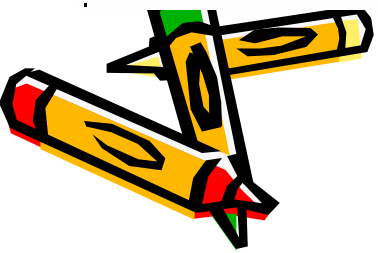
- Folosind impuritatea clasificării *greşite*, să conturăm profilul clientului supermarketului vs. clientul micului magazin din colțul străzii sau de la parterul blocului, utilizând următoarele atribute:

- *posesor sau nu automobil*,
- *venit* lunar (mai mic, respectiv mai mare de 1000 lei),
- *salariat/pensionar*

şi următoarea bază de date:



	client	venit	posesor auto	salariat/ pensionar
1	supermarket	peste 1000	da	salariat
2	supermarket	sub 1000	nu	salariat
3	mic magazin	sub 1000	nu	pensionar
4	mic magazin	sub 1000	nu	salariat
5	supermarket	sub 1000	da	salariat
6	supermarket	sub 1000	da	pensionar
7	mic magazin	peste 1000	da	pensionar
8	mic magazin	sub 1000	nu	pensionar
9	supermarket	peste 1000	nu	salariat
10	supermarket	sub 1000	nu	salariat
11	supermarket	sub 1000	nu	salariat



12	supermarket	peste 1000	da	salariat
13	supermarket	peste 1000	da	pensionar
14	supermarket	sub 1000	da	pensionar
15	mic magazin	sub 1000	nu	pensionar
16	mic magazin	peste 1000	nu	pensionar
17	supermarket	sub 1000	nu	salariat
18	supermarket	sub 1000	da	salariat
19	supermarket	peste 1000	da	salariat
20	supermarket	sub 1000	da	salariat
21	mic magazin	sub 1000	nu	pensionar
22	mic magazin	sub 1000	nu	salariat
23	mic magazin	sub 1000	da	pensionar
24	supermarket	peste 1000	da	salariat
25	supermarket	sub 1000	nu	salariat

Să decidem care va fi nodul rădăcină, folosind măsura de clasificare greşită:

	supermarket	mic magazin
venit lunar peste 1000 RON	6	2
venit lunar sub 1000 RON	10	7

$$i_M(peste\,1000) = 1 - \max\{\frac{6}{8}, \frac{2}{8}\} = 0.25\ ;$$

$$i_M(sub\,1000) = 1 - \max\{\frac{10}{17}, \frac{7}{17}\} = 0.4118\ ;$$

$$i_{split}(venit) = \frac{8}{25} \cdot 0.25 + \frac{17}{25} \cdot 0.4118 = 0.36$$

	supermarket	mic magazin
posesor auto	10	6
nu posedă automobil	7	2

$$i_{split}(auto) == 0.3198\ ,$$

	supermarket	mic magazin
salariat	13	2
pensionar	7	3

$$i_{split}(salariat\ /\ pensionar) == 0.200\ .$$

Conform celui mai mic indice de partiționare, în nodul rădăcină va fi atributul salariat.

Pentru salariați avem:

	supermarket	mic magazin
venit lunar peste 1000 RON	5	0
venit lunar sub 1000 RON	9	1

$$i_{split}\,(venit) = 0.0667\ ;$$

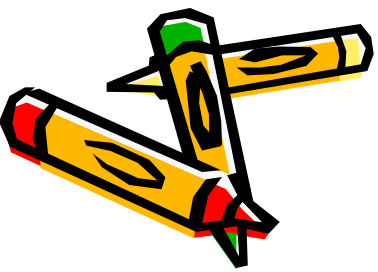
	supermarket	mic magazin
posesor auto	7	0
nu posedă automobil	6	2

$$i_{split}\,(auto) == 0.1778\ .$$

Următorului subnod îi atribuim atributul *venit* lunar sub 1000 RON.

Pentru salariați cu venit sub 1000 RON avem următoarea situație:

	supermarket	mic magazin
posesor auto	3	0
nu posedă automobil	3	2



Să studiem alegerea pe care o fac pensionarii:

	supermarket	mic magazin
venit lunar peste 1000 RON	1	2
venit lunar sub 1000 RON	2	5

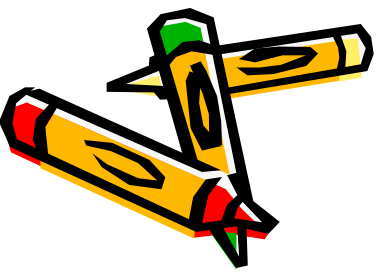
$$i_{split}\,(venit) = 0.300\,;$$

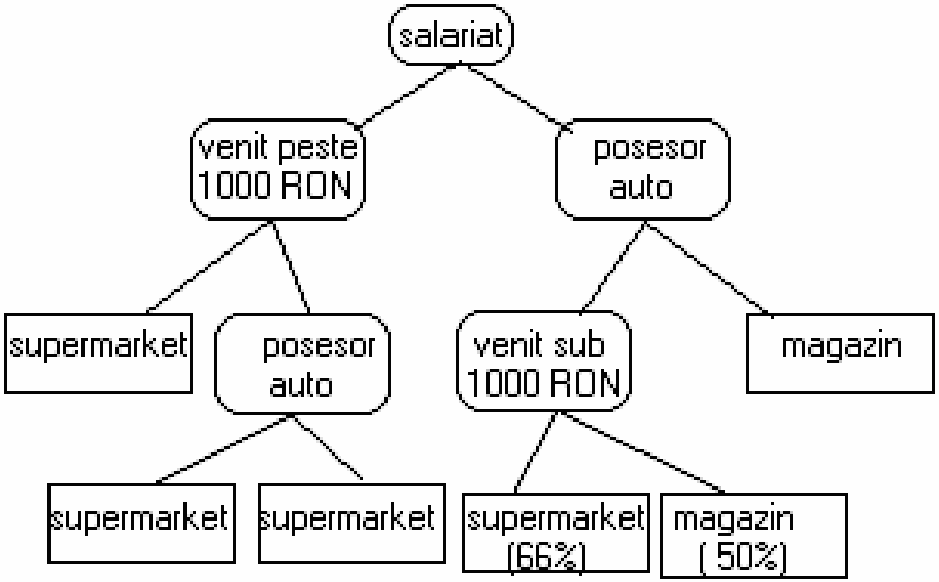
	supermarket	mic magazin
posesor auto	3	2
nu posedă automobil	0	5

$$i_{split}\,(auto) == 0.200\,;$$

Următorul subnod va avea atributul *posesor auto*.

	supermarket	mic magazin
venit lunar peste 1000 RON	1	1
venit lunar sub 1000 RON	2	1

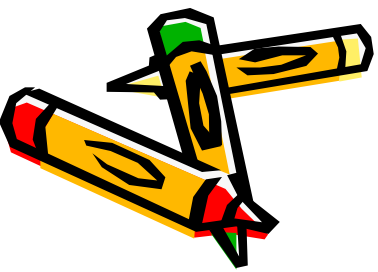




Regulile deduse din acesta, alcătuiesc profilul cumpărătorului din supermarket, respectiv din micul magazin.

Regulile deduse din acesta, alcătuiesc profilul cumpărătorului din supermarket, respectiv din micul magazin.
Teoretic, se afirmă că uneori indexul Gini descreşte, în timp ce măsura clasificării greşite nu.
Pe de altă parte, măsura Gini anticipează noile ramificaţii.

Intervine vreo modificare în exemplul nostru, folosind indicele Gini de partiţionare?



$$Gini_{split}\ (venit) = 0.3350;$$

$$Gini_{split}\ (posesor\ auto) = 0.2841;$$

$$Gini_{split}\ (salariat\ /\ pensionar) = 0.3067;$$

Nodul rădăcină va fi în acest caz posesor auto;

pentru cumpărătorii posesori auto avem:

	supermarket	mic magazin
salariat	7	0
pensionar	3	2

$$Gini_{split}(salariat\ /\ pensionar) = 0.200;$$

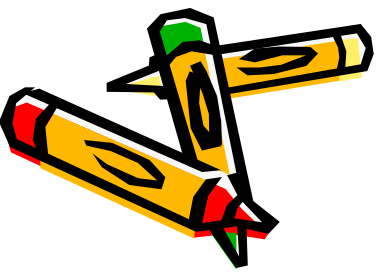
	supermarket	mic magazin
venit lunar peste 1000 RON	5	0
venit lunar sub 1000 RON	5	2

$$Gini_{split}(venit) = 0.2381\,.$$

- Subnodului ce urmează îi atribuim proprietatea salariat.

în cazul pensionarilor, posesori auto:

	supermarket	mic magazin
venit lunar peste 1000 RON	1	1
venit lunar sub 1000 RON	2	1



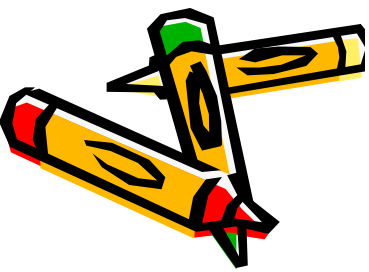
Ce se întâmplă în cazul celor ce nu au automobil:

	supermarket	mic magazin
salariat	6	2
pensionar	0	5

$$Gini_{split}(salariat / pensionar) = 0.200$$

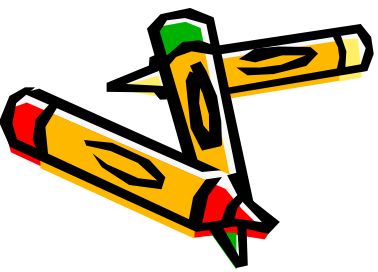
	supermarket	mic magazin
venit lunar peste 1000 RON	1	1
venit lunar sub 1000 RON	5	6

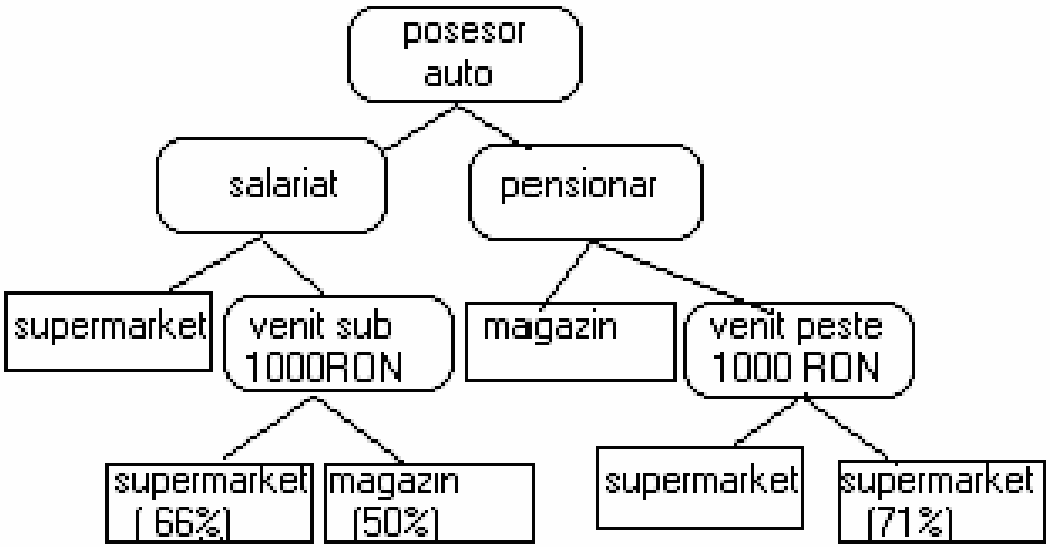
$$Gini_{split}(venit) = 0.4965$$



Situația cumpărătorilor salariați, ce nu au mașină este:

	supermarket	mic magazin
venit lunar peste 1000 RON	1	0
venit lunar sub 1000 RON	5	2



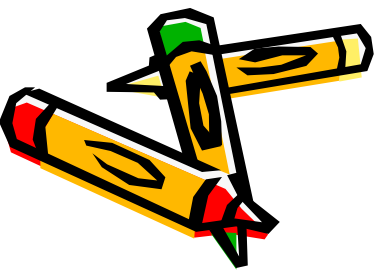


Puritatea frunzei „supermarket" este mai bună, cum arată un calcul simplu, dar regulile deduse, ce descriu profilul cumpărătorului, sunt aceleaşi.

Să folosim câştigul de informaţie pentru aceeaşi mulţime de antrenament.

Calculăm măsura entropiei tipului de magazin: supermarket/mic magazin:

$$i(\textit{magazinul ales}) = -\frac{16}{25}\cdot\log_2\frac{16}{25} - \frac{9}{25}\cdot\log_2\frac{9}{25} = 0.9427\ .$$



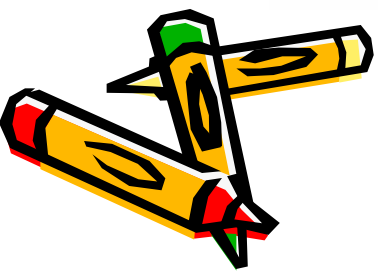
calculăm câştigul de informaţie folosind atributul ***venit***:

$$i(venit<1000) = -\frac{10}{17}\cdot\log_2\frac{10}{17} - \frac{7}{17}\cdot\log_2\frac{7}{17} = 0.9774;$$

$$i(venit>1000) = -\frac{6}{8}\cdot\log_2\frac{6}{8} - \frac{2}{8}\cdot\log_2\frac{2}{8} = 0.8113.$$

$$gain(venit) = i(magazinales) - \frac{17}{25}\cdot i(venit<1000) -$$

$$-\frac{8}{25}\cdot i(venit>1000) = 0.0184.$$



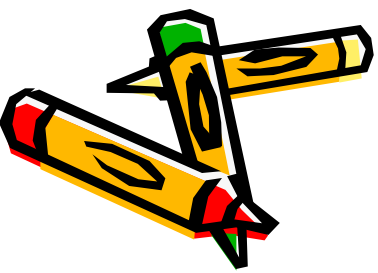
calculăm câștigul de informație folosind atributul *posesor auto*:

$$i(posesor\ \ auto - DA) = -\frac{10}{12}\cdot \log_2 \frac{10}{12} - \frac{2}{12}\cdot \log_2 \frac{2}{12} = 0.6500\ ;$$

$$i(posesor\ \ auto - NU) = -\frac{6}{13}\cdot \log_2 \frac{6}{13} - \frac{7}{13}\cdot \log_2 \frac{7}{13} = 0.9957\ \ ;$$

$$gain(posesor\ auto) = i(magazin\ ales) - \frac{12}{25}\cdot i(posesor\ auto - DA) -$$

$$- \frac{13}{25}\cdot i(posesor\ auto - NU) = 0.1129\ .$$



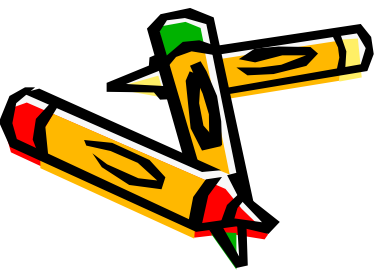
calculăm câştigul de informaţie folosind atributul salariat /pensionar:

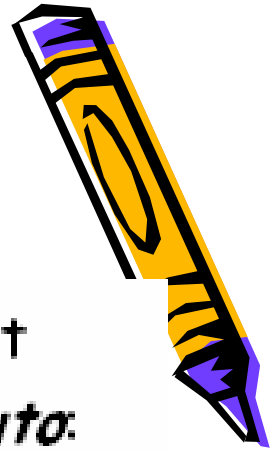
$$i(salariat) = 0.5665\ ;\ i(pensionar) = 0.8813\ ;$$

$$gain(salariat - pensionar) = 0.9427 - \frac{15}{25} \cdot 0.5665 - \frac{10}{25} \cdot 0.8813 =$$

$$= 0.2503\ .$$

Stabilim cel mai mare câştig de informaţie şi astfel în nodul rădăcină atributul va fi ***salariat***.





calculăm pentru salariat câştigul de informaţie obţinut prin folosirea atributelor ***venit***, respectiv ***posesor auto***:

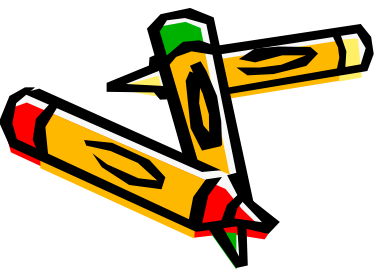
$$i(venit > 1000) = 0;\ i(venit < 1000) = 0.7219;$$

$$gain(venit) = 0.5665 - \frac{10}{15} \cdot 0.7219 = 0.0852;$$

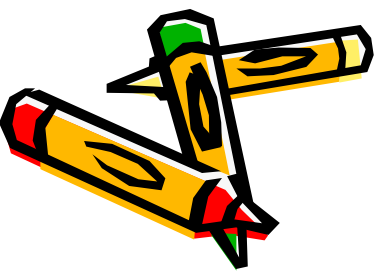
$$i(posesor\ \ auto - da) = 0;\ i(posesor\ \ auto - nu) = 0.8113;$$

$$gain(posesor\ \ auto) = 0.5665 - \frac{8}{15} \cdot 0.8113 = 0.1338.$$

În concluzie, în acest subnod atributul va fi ***posesor auto***;



dacă nu este posesor auto vom considera un subnod cu atributul venit >1000RON.



dacă este pensionar, calculăm câștigul de informație pentru ***venit***, respectiv pentru ***posesor auto***:

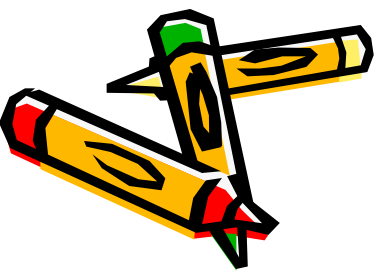
$$i(venit<1000)=0.8631;\ i(venit>1000)=0.9183;$$

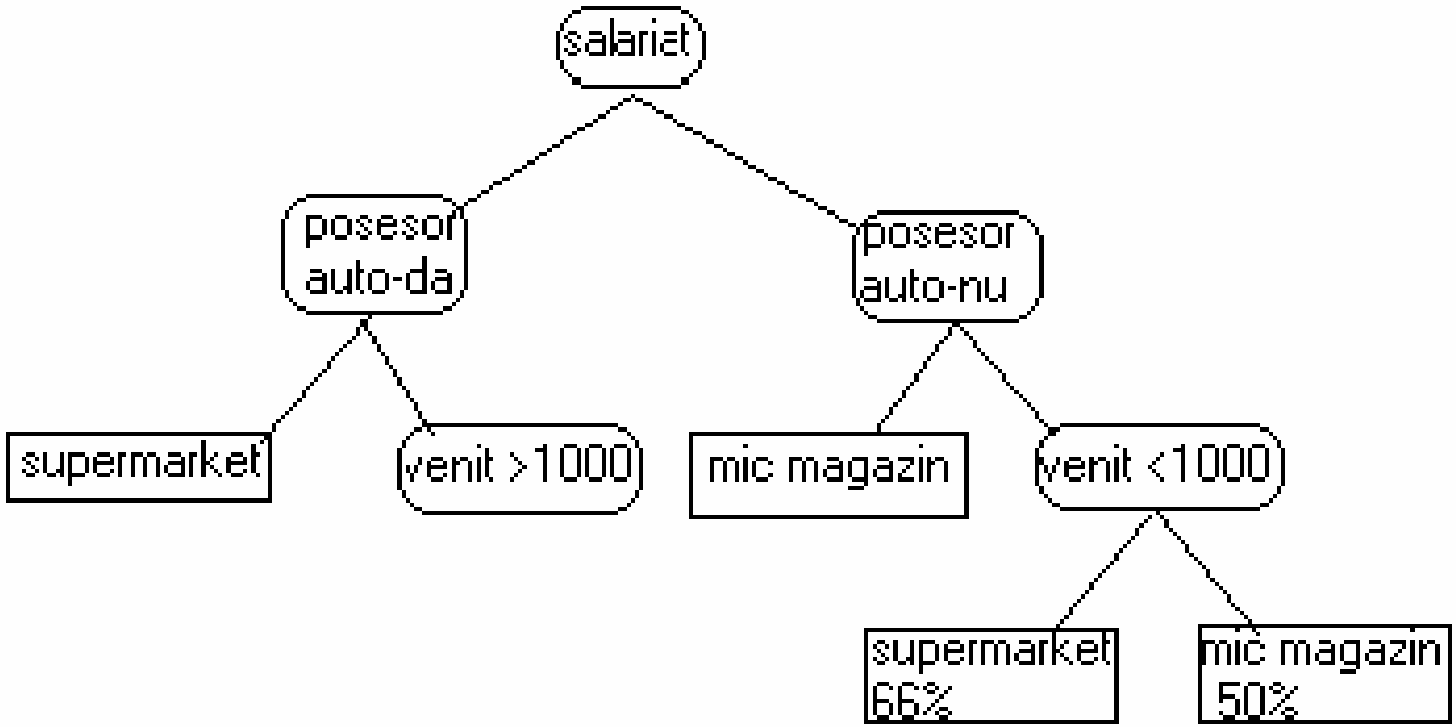
$$gain(venit)=0.8813-\frac{7}{10}\cdot 0.8631-\frac{3}{10}\cdot 0.9183=0.0016;$$

$$i(posesor\quad auto-da)=0.9710;\ i(posesor\quad auto-da)=0.9710;$$

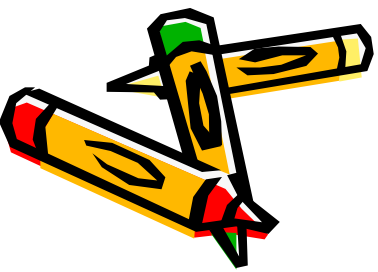
$$gain(posesor\quad auto)=0.8813-\frac{5}{10}\cdot 0.9710=0.3958$$

și astfel atributul va fi ***posesor auto***.




salariat
posesor auto-da
posesor auto-nu
supermarket
venit >1000
mic magazin
venit <1000
supermarket 66%
mic magazin 50%

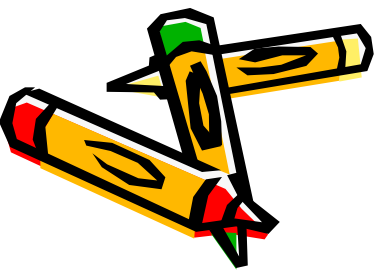
Scopul construcției arborilor de clasificare și decizie:
a obține o predicție cât mai precisă.
Costurile predicției sunt indicatori ai acurateței acesteia.
Consecințele unei clasificări eronate sunt deosebit de importante.



exemplu

în cazul în care un medic greşeşte diagnosticul benign/malign al unei tumori, ce este mai grav :

- tumoarea malignă să fie catalogată drept benignă?
- tumoarea benignă să fie considerată a fi malignă?



prior probabilities

Principalele costuri legate de procesul de clasificare sunt:

- *Probabilitățile prealabile* (*prior probabilities*) sunt acei parametri care specifică probabilitatea ca un obiect să aparțină unei anumite clase.

De obicei, se aleg acei parametri proporționali cu numărul de obiecte din fiecare clasă.

missclassification costs

- *Costuri de clasificare greşită* (*misclassification costs*) se referă la faptul că, în procesul de clasificare, unele categorii au nevoie de o clasificare mai precisă.

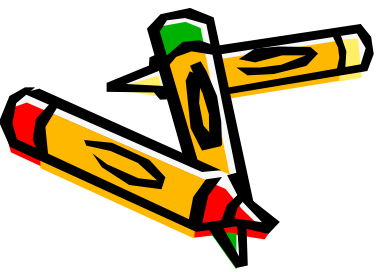
Reluând exemplul cu diagnosticul unei tumori, este mult mai importantă acuratețea clasificării uneia ca fiind malignă, decât ca fiind benignă.

Costurile de clasificare greşită sunt alese astfel încât să reflecte importanța fiecărei clase.

Problema alegerii criteriului de oprire a procesului de divizare a nodurilor.

Procesul de partiționare se derulează până când toate nodurile terminale -frunzele- sunt *pure*, adică vor conține numai elemente din aceeași categorie.

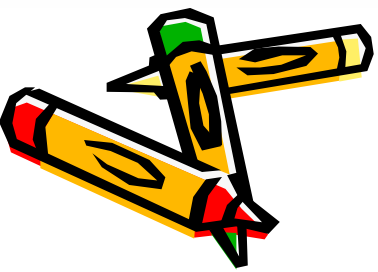
Este important ca pe mulțimea de testare/validare clasificatorul să aibă performanță maximă.



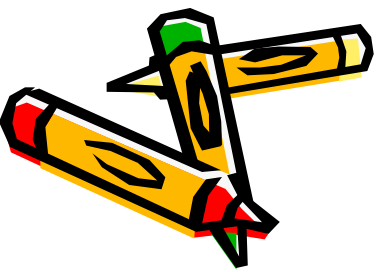
stop

Uzual, sunt utilizate două reguli de *Stop*:

- *Minimul n*: condiția de *Stop* specifică un număr minim de obiecte care să fie conținute în nodurile terminale. Divizarea unui nod ia sfârșit atunci când fie nodul este pur, fie conține numărul minim de obiecte.

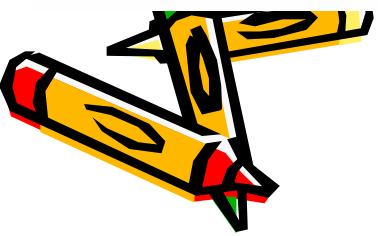


- *Proporția de obiecte*: condiția de *Stop* impune ca divizarea unui nod să ia sfârșit atunci când nodul este pur sau conține un procentaj minim de obiecte dintr-o anumită clasă.



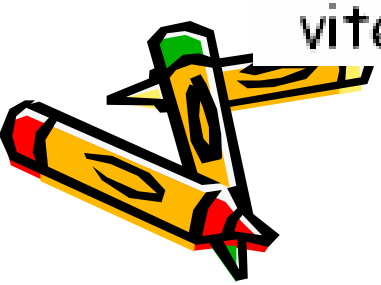
overfitting/underfitting

Un arbore fiind construit pentru a putea fi aplicat la diverse seturi de date, este necesară evitarea unei potriviri prea accentuate (*overfitting*) cu mulțimea pe care s-a făcut antrenamentul.
Când arborele este prea simplu față de datele utilizate la antrenament și, în consecință, atât eroarea de antrenament cât și cea de testare sunt mari, avem de-a face cu situația sub-potrivire (*underfitting*) a arborelui cu datele.



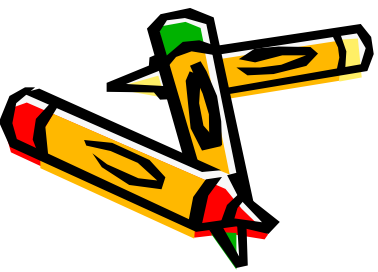
pruning

Overfitting-ul este cel mai adesea întâlnit.
În acest caz se utilizează metodă de fasonare (*pruning*) a arborelui.
Se utilizează metode statistice pentru îndepărtarea ramurilor nesemnificative, redundante, sau care nu urmează pattern-ul general al datelor, obținând astfel un arbore mai puțin „stufos", cu o mai mare viteză de clasificare.

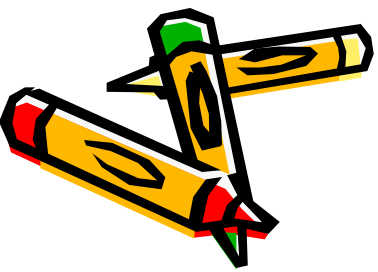


avantaje

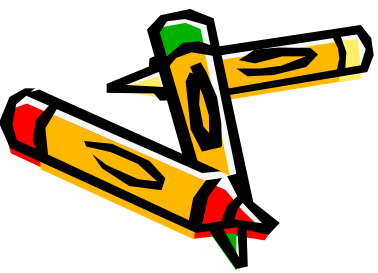
- Sunt uşor de înţeles şi interpretat, forma lor grafică reprezentând un atu puternic în acest sens;
- Necesită un volum mic de pregătire a datelor în raport cu alte tehnici;
- Permit utilizarea atât a datelor nominale cât şi a celor categoriale, fără nicio restricţie;
- Logica deciziei poate fi urmărită uşor, regulile de clasificare fiind la vedere;
- Permit utilizarea unor tehnici statistice clasice pentru validarea modelului;
- Lucrează bine cu mulţimi mari de date.



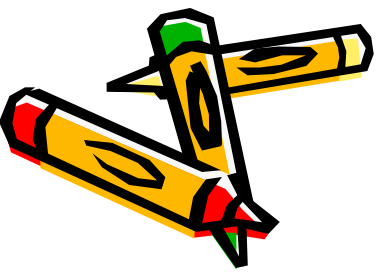
- în cazul unui număr prea mare de clase se poate deteriora rezultatul;
- algoritmul nu este incremental, în sensul că dacă apar date noi este necesară reluarea fazei de antrenament cu eşantionul complet format din vechile date şi cele noi.



Clasificarea bazată pe reguli de asociere



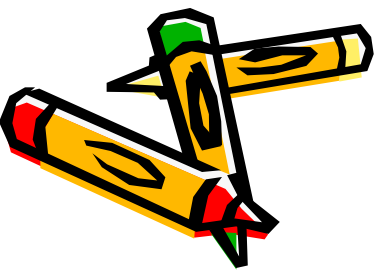
Metoda regulii de asociere (sau analiza asocierii) este o tehnică **nesupervizată**, care caută legături între înregistrările dintr-un set de date.



regula de asociere

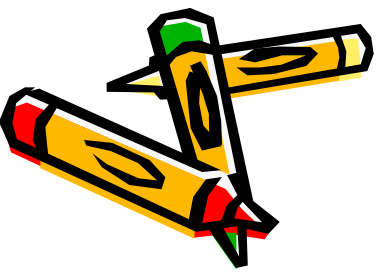
O *regulă de asociere* (*association rule*) este o expresie de implicare de felul „dacă (***IF***) *X* atunci (***THEN***) *Y*",
unde articolele (item-uri) *X* și *Y* sunt distincte,
$X \cap Y = \phi$.

- *X antecedentul* regulii
- *Y consecința* regulii.



metode de construire a regulilor de asociere

- metoda *indirectă*, care constă în extragerea regulilor folosind alți clasificatori, de exemplu arborii de clasificare și decizie;
- metoda *directă*, care constă în extragerea regulilor direct din date.

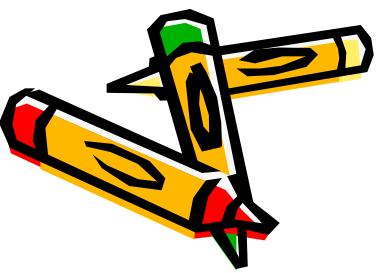


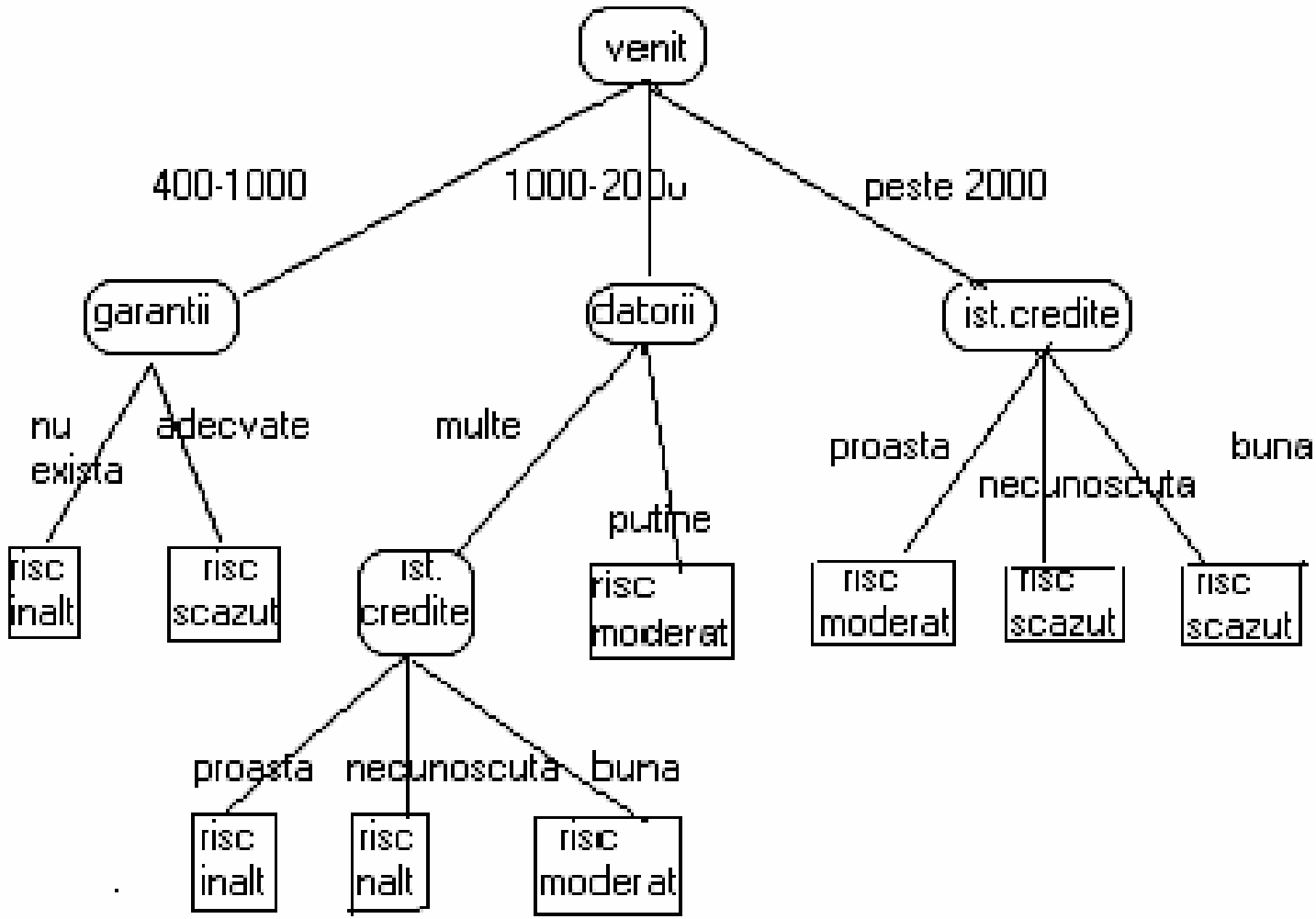
regulile de asociere pot fi:

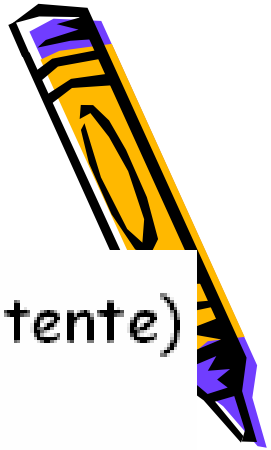
- reciproc *exclusive*, adică regulile sunt independente una de alta, fiecare articol fiind acoperit de cel puțin o regulă
- *exhaustive*, caz în care se consideră toate combinațiile posibile ale valorilor atributelor și fiecare articol este acoperit de cel puțin o regulă.

exemplu

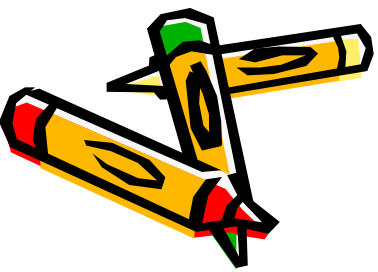
regulile de asociere pe baza arborelui de decizie, construit pentru exemplul referitor la estimarea riscului acordării unui credit




venit
400-1000
1000-2000
peste 2000
garantii
datorii
ist.credite
nu
exista
adecvate
multe
putine
proasta
necunoscuta
buna
risc
inalt
risc
scazut
ist.
credite
risc
moderat
risc
moderat
risc
scazut
risc
scazut
proasta
necunoscuta
buna
risc
inalt
risc
nalt
risc
moderat



- **R1**: (venit lunar 400-1000 RON) și (garanții inexistente) ⇒(risc înalt).
- **R2**: (venit lunar 400-1000 RON) și (garanții adecvate)
- ⇒(risc scăzut).
- **R3**: (venit lunar 1000-2000 RON) și (datorii multe) și (istoria creditelor proastă) ⇒(risc înalt).
- **R4**: (venit lunar 1000-2000 RON) și (datorii multe) și (istoria creditelor necunoscută) ⇒(risc înalt)
- **R5**: (venit lunar 1000-2000 RON) și (datorii multe) și (istoria creditelor bună) ⇒(risc moderat).

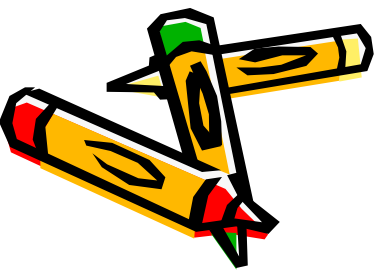


- **R6**: (venit lunar 1000-2000 RON) și (datorii puține) $\Rightarrow$(risc moderat).
- **R7**: (venit lunar peste 2000 RON) și (istoria creditelor proastă) $\Rightarrow$(risc moderat).
- **R7**: (venit lunar peste 2000 RON) și (istoria creditelor necunoscută) $\Rightarrow$(risc scăzut).
- **R8**: (venit lunar peste 2000 RON) și (istoria creditelor bună) $\Rightarrow$(risc scăzut)

sunt reguli reciproc exclusive și exhaustive.

Regulile de asociere pot fi caracterizate prin:

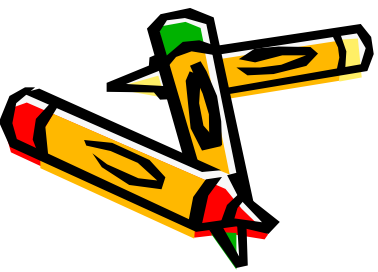
- *puterea de acoperire* a regulii, definită ca procentajul de articole care satisfac antecedentul regulii;
- *acurateţea* regulii, definită ca procentajul de articole care satisfac atât antecedentul cât şi consecinţa regulii.



În exemplul referitor la estimarea riscului acordării unui credit, în cazul regulii:

„ datorii multe $\Rightarrow$ risc înalt"

puterea de acoperire este 45% (9 din 20 cazuri), iar acuratețea este 33% (6 din 9 cazuri).



analiza cosului de consum

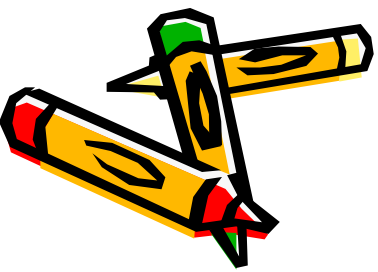
Metoda *directă* este câteodată definită ca ***analiza coşului de consum***, care este de altfel şi cea mai folosită aplicaţie a sa.

Analiza coşului de consum constă în găsirea de asocieri între produsele afişate pe bonurile de casă.

Se studiază astfel ce cumpărături fac clienţii, pentru a obţine informaţii asupra produselor ce tind a fi cumpărate în acelaşi timp.

Metoda poate fi aplicată în orice sector de activitate, pentru care este necesară găsirea de posibile grupări de produse sau servicii: servicii bancare, servicii de telecomunicații etc.

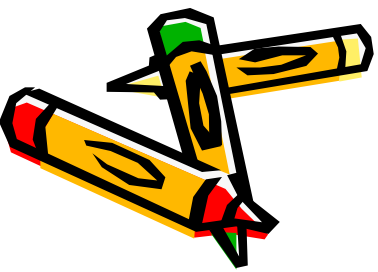
Poate fi folosită în domeniul medical pentru studiul complicațiilor apărute datorită asocierii unor medicamente sau în domeniul fraudelor, caz în care se caută asocierii neobișnuite.



Rezultatele metodei sunt regulile de asociere, care sunt utile în marketing.

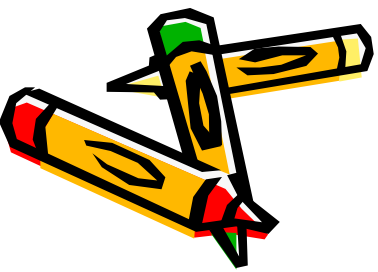
De cele mai multe ori metoda poate produce reguli interesante.

Uneori se obțin și reguli triviale.



exemple

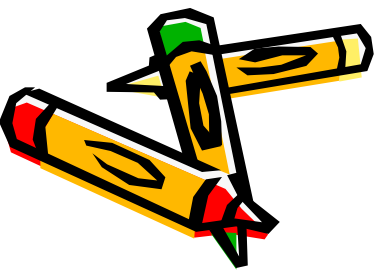
- dacă un client cumpără pește și lămâie, atunci va cumpăra vin alb (Franța); regulă trivială
- dacă un client (bărbat) cumpără scutece, va cumpăra și bere (SUA, exemplu clasic); regulă interesantă

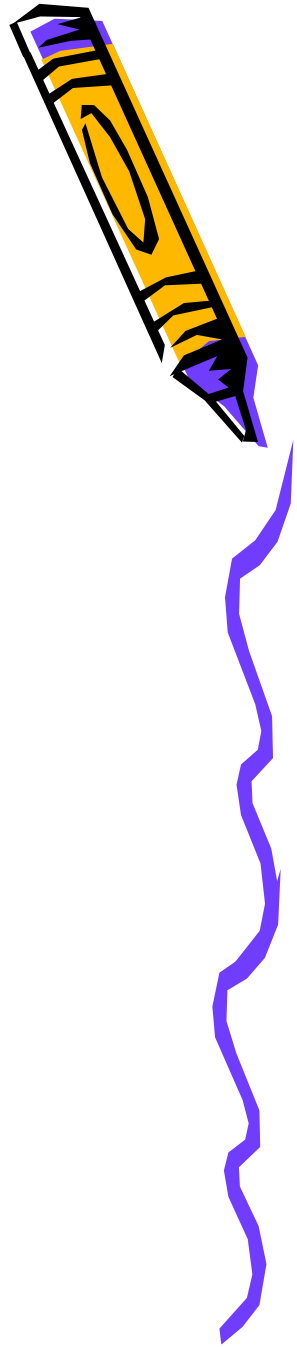


Considerăm următoarele date, provenite dintr-o listă de 10 clienți.

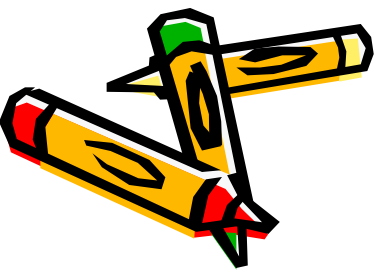
Un client a cumpărat o listă de articole, listă de lungime variabilă.

Articolele sunt notate A, B,..., G şi reținem că pe linie se află lista de articole a fiecărui client.





	A	B	C	D	E	F	G
1	×		×			×	
2	×	×		×			
3		×	×			×	
4		×		×		×	
5	×						×
6					×	×	
7	×		×				
8			×		×		
9	×	×	×	×		×	
10	×			×		×	



Creăm un tabel care să ilustreze de câte ori două produse sunt cumpărate simultan de un client:

	A	B	C	D	E	F	G
A	6	2	3	2	0	3	1
B	2	4	2	3	0	3	0
C	3	2	5	1	1	2	0
D	2	3	1	4	0	3	0
E	0	0	1	0	2	1	0
F	3	3	2	3	1	6	0
G	1	0	0	0	0	0	1

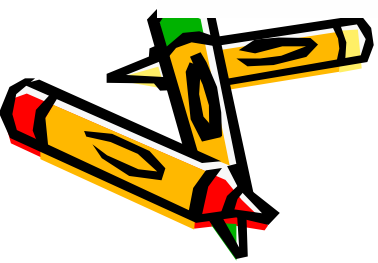
support

Una dintre caracteristicile ce măsoară robustețea unei reguli de asociere este *suportul* (*support*).

O regulă de asociere este de forma:

„dacă este îndeplinit *antecedentul* atunci avem *consecința*."

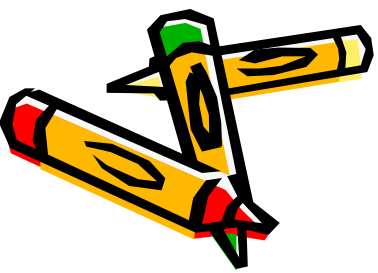
support (**R**) = frecvența (*antecedentul* și *consecință*)



notând cu x numărul de clienți care cumpără simultan articolele din antecedent și consecință și cu n numărul total de clienți, suportul regulii **R** este:

$$\text{support}(\mathbf{R}) = \frac{x}{n}$$

Suportul măsoară (procentual) cât de des se poate aplica regula la o mulțime de date, (cât de des apar anumite articole împreună în totalul tranzacțiilor.



exemplu

Considerăm regulile, obținute din tabelul prezentat anterior:

- **R1**: dacă *A* atunci B;
- **R2**: dacă *A* atunci *C*;
- **R3**: dacă *C* atunci *A*.

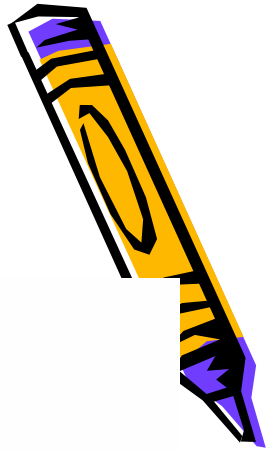
Articolele *A* și B sunt cumpărate simultan de 20% din clienți, adică suportul regulii 1 este de 20% .

Deoarece articolele *A* și *C* apar împreună în 30% din coșurile de cumpărături, regulile 2 și 3 au suportul de 30%.

confidence

O caracteristică ce măsoară robustețea unei reguli este *încrederea* (*confidence*) .

Încrederea este raportul dintre frecvența aparițiilor (antecedent și consecință) și frecvența aparițiilor (antecedent) adică raportul dintre numărul de clienți ce cumpără simultan articolele care apar în regulă și numărul de cumpărători ai articolelor ce apar în *antecedent*.



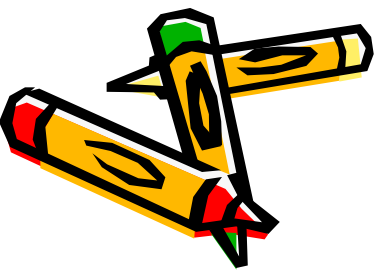
Astfel pentru regula 2 avem:

$$confidence(\mathbf{R2}) = \frac{3}{6},$$

în timp ce pentru regula 3 avem

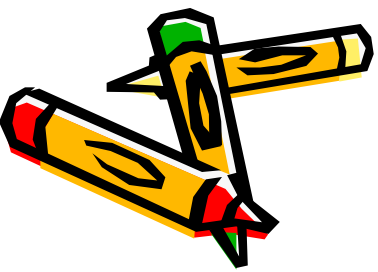
$$confidence(\mathbf{R3}) = \frac{3}{5}.$$

În concluzie, regula 2 prezintă o încredere de 50%, în timp ce regula 3 prezintă o încredere de 60% și astfel va fi preferată regula 3.



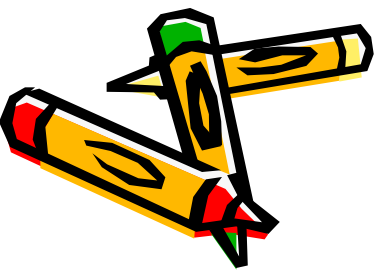
Putem spune că *încrederea* măsoară cât de mult depinde un articol de altul.

Să reținem: dintr-o mulțime de reguli ce au un suport suficient de mare, se alege aceea ce prezintă încrederea maximă.



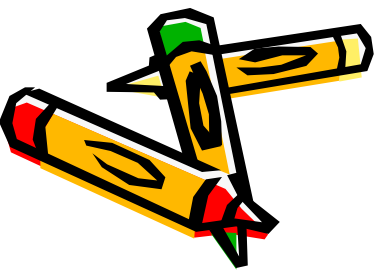
Cele două caracteristici, suportul şi încrederea, nu sunt suficiente pentru robustețea regulii de asociere.
Vom considera articolele A, B şi C şi frecvențele lor de apariție:

art.	A	B	C	A&B	A&C	B&C	A&B&C
frecv.	45%	42.%	40%	25%	20%	15%	5%



Dacă vom considera reguli cu trei articole, acestea vor avea același suport de 5%. Nivelul de încredere este:

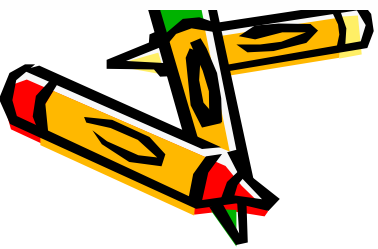
regula	*confidence*
dacă A şi B atunci C	0.20
dacă A şi C atunci B	0.25
dacă B şi C atunci A	0.33



Regula „dacă B şi C atunci A" prezintă cea mai mare încredere de 0.33, ceea ce înseamnă că:

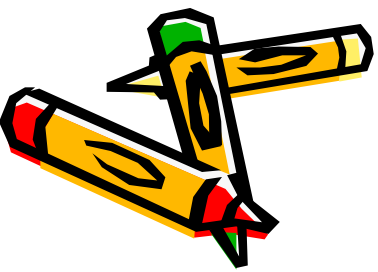
dacă articolele B şi C apar simultan pe un bon de casă, atunci articolul A va apărea pe acest bon cu o probabilitate de 33%.

Dacă studiem tabelul, observăm că A apare în 45% din coşurile de cumpărături, ceea ce înseamnă că este preferabil să prognozăm apariţia lui A, decât apariţia simultană a articolelor B şi C.



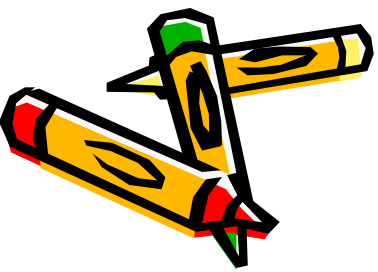
diferența de nivel (*lift*) permite compararea rezultatului predicției folosind regula, cu predicția obținută fără utilizarea regulii. *Diferența de nivel* este definită prin:

$$\text{Diferența de nivel} = \frac{confidence}{frecventa\ (\text{consecinta})}$$



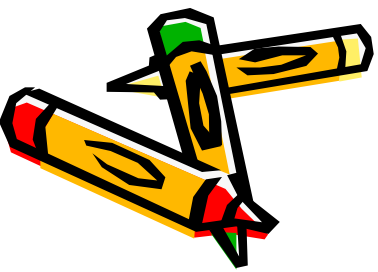
O regulă este interesantă dacă diferența de nivel este mai mare decât 1.

regula	confidence	frecvență	dif.de nivel
dacă A şi B atunci C	0.20	40%	0.50
dacă A şi C atunci B	0.25	42%	0.59
dacă B şi C atunci A	0.33	45%	0.74



Regula „dacă A atunci B" are un suport de 25%, o încredere de $\frac{25}{45} = 0.55$ şi o diferență de nivel de $\frac{55}{42} = 1.3$.

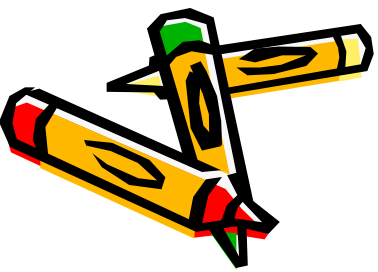
În general, regula cea mai bună conține mai puține articole.



exemplu

- comercializarea următoarelor două băuturi alcoolice: berea și whisky în 500.000 tranzacții.

- 5.000 tranzacții conțin whisky (1% din totalul tranzacțiilor);
- 30.000 tranzacții conțin bere (6% din totalul tranzacțiilor);
- 2.000 tranzacții conțin și bere și whisky (0.4% din totalul tranzacțiilor).



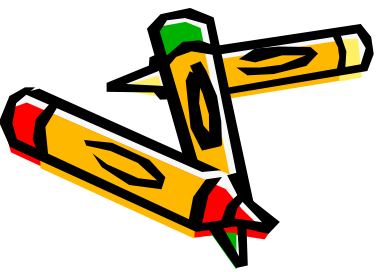
- *Suportul* este 0.4% din total (2.000/500.000);
  - Regula

    „Când oamenii cumpără whisky, cumpără de asemenea şi bere"

    are încrederea 40% (2.000/5.000);
  - Regula

    „Când oamenii cumpără bere, cumpără de asemenea şi whisky"

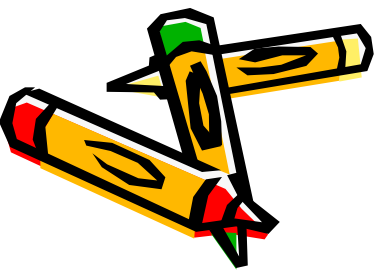
    are încrederea 6.66% (2.000/30.000).



Cele două reguli au același suport 0.4% și aceeași diferență de nivel 6.66.

Dacă nu mai există informații suplimentare despre alte tranzacții, putem face următoarele afirmații:

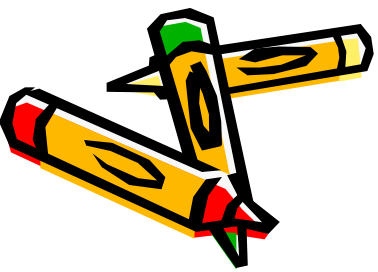
- 1% din clienții cumpără whisky;
- 6% din clienții cumpără bere.



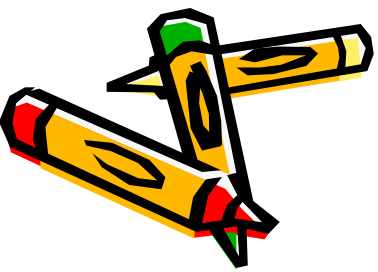
Cele două procentaje, 1% şi 6%, sunt numite *încrederea aşteptată* de a cumpăra whisky sau bere, indiferent de celelalte cumpărături.

Regula de cumpărare *whisky-bere* poate fi exprimată în termen de *diferenţă de nivel* astfel:

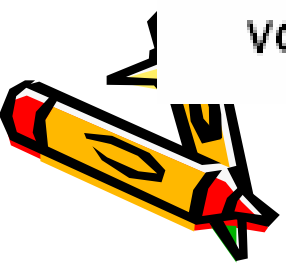
„Clienţii care cumpără whisky sunt de 6,66 *ori mai tentaţi să cumpere şi bere odată cu whisky".*



Fiind dată o listă de *n* articole, să considerăm o listă de *m* cumpărături (prin cumpărătură înțelegem coșul de cumpărături al unui client, ilustrat prin bonul de casă).
Pentru a descoperi regulile de asociere, procedăm astfel:

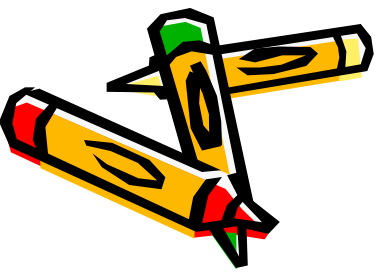


- calculăm numărul de apariții a fiecărui articol;
- construim tabelul de apariție simultană pentru perechile de articole;
- determinăm regulile ce conțin două articole, utilizând valorile de suport, încredere şi diferență de nivel;
- construim tabelul de apariție simultană pentru tripletele de articole;
- determinăm regulile ce conțin trei articole, utilizând valorile de suport, încredere şi diferență de nivel;



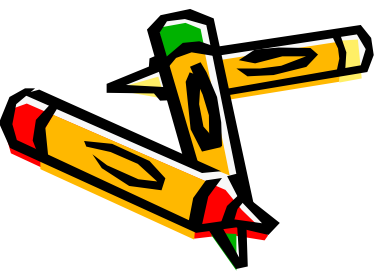
Valoarea m (numărul cumpărăturilor) este în general foarte mare.

Pentru a construi tabelul de apariție simultană, este necesară parcurgerea acestei liste de mai multe ori, așa că este necesară o arhitectură a acesteia care să permită acces rapid.



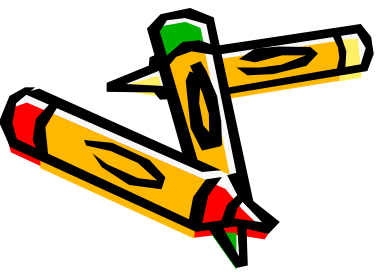
mărimea tabelelor, ca funcție de n și de număr de articole care apar în regulă.

n	C_n^2	C_n^3	C_n^4
100	4950	161700	3921225
10000	$\approx 5 \cdot 10^7$	$\approx 1.7 \cdot 10^{11}$	$\approx 4.2 \cdot 10^{14}$



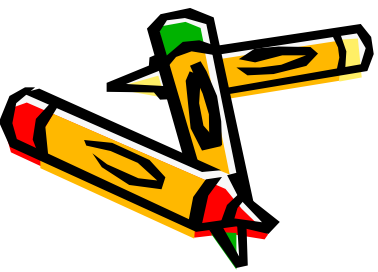
Definim procesul de descoperire a regulilor de asociere:

„Fiind dată o mulțime de tranzacții, să se descopere toate regulile posibile pentru care atât suportul cât și încrederea să fie mai mari sau egale decât anumite praguri prestabilite".



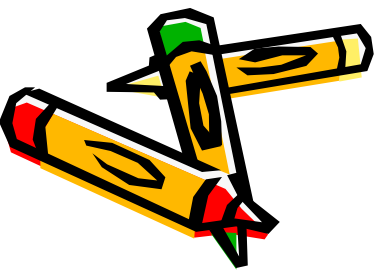
În acest sens, se poate arăta că numărul total *R* al regulilor care pot fi extrase dintr-un set de date, care conține un număr de *n* articole, este dat de formula:

$$R = 3^n - 2^{n+1} + 1,$$



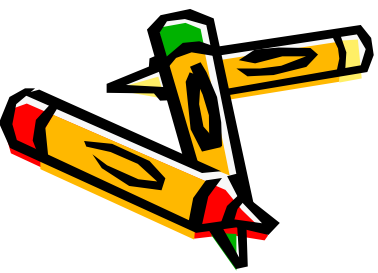
fasonare cu suport minim

O tehnică de reducere a numărului de articole şi a combinaţiilor lor luate în considerare, poartă numele de *fasonare cu suport minim*.

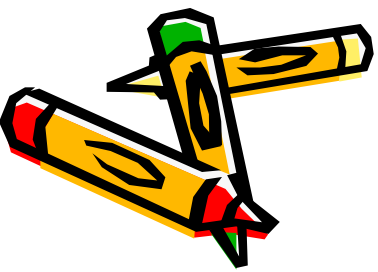


exemplu

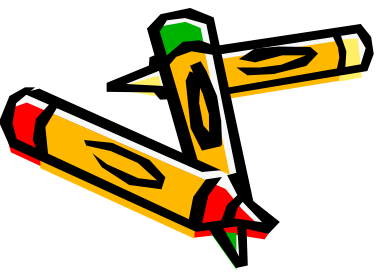
Căutăm reguli cu trei articole:
vom considera doar acele articole al căror suport este mai mare decât o valoare dată *apriori*.
De exemplu, pentru o listă de $m = 1000000$ cumpărături, dacă suportul minim este 2%, în etapa a doua, luăm în considerare doar regulile de forma „dacă X atunci Y", unde X și Y apar simultan în cel puțin 20000 de cumpărături.



Fasonarea cu suport minim permite eliminarea articolelor ce sunt mai puțin frecvente.
În funcție de etapă, putem varia suportul minim și astfel, diminuându-l, putem găsi combinații rare de articole frecvent cumpărate.
Dacă suportul minim va crește, vom obține combinații frecvente de articole rar cumpărate.



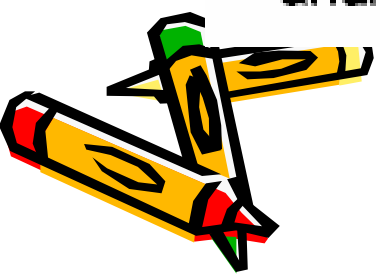
În anumite aplicații se poate limita numărul de combinații prin limitarea formei regulilor căutate.
De exemplu, concluzia regulii este restrânsă la o submulțime a mulțimii articolelor, cum ar fi ultimele modele primite.
Se pot limita calculele prin crearea de grupări de articole, ceea ce necesită sfatul specialiștilor în domeniu.



exemplu

în cazul unui supermarket, un anumit articol poate fi descris astfel:

- conservă;
- conservă de legume;
- conservă de legume de la un anumit producător;
- conservă de legume de un anumit gramaj;
- conservă de legume de la un anumit producător, de un anumit gramaj.

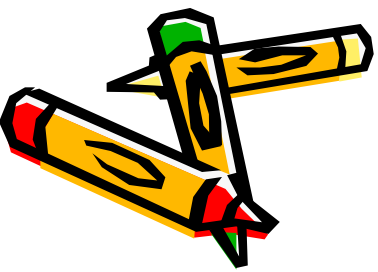


caracteristicile metodei

- regulile prezentate sunt uşor de folosit şi de interpretat în situaţii concrete;
- este o metodă de învăţare nesupervizată, pentru extragerea regulilor fiind necesară doar lista articolelor;
- cumpărăturile sunt de mărime variabilă;
- se poate introduce şi variabila timp, în sensul că se pot genera reguli de forma: "clientul care a cumpărat produsul A, va cumpăra probabil produsul B în doi ani";

- metoda şi calculele sunt elementare şi astfel se poate programa uşor în cazul unor baze de date de mărime rezonabilă;
- necesită mult timp de lucru şi, prin regruparea articolelor sau metoda suportului minim, se diminuează calculele, dar există riscul pierderii unor reguli importante;

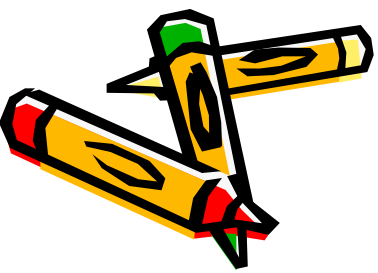
- metoda este mai puțin eficientă pentru articolele rar cumpărate, în aceste caz de obicei se variază suportul minim;
- se pot deduce reguli triviale sau reguli inutile



avantajele metodei

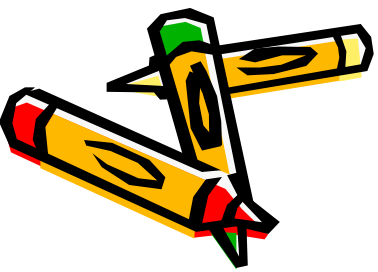
- expresivitate pronunțată;
- uşurință în interpretare;
- uşurință în generare;
- viteză ridicată de clasificare a unor noi instanțe;
- performanță generală comparabilă cu cea a arborilor de clasificare şi decizie.

K- nearest neighbor



Metoda clasifică un nou obiect pe baza cazurilor similare cele mai apropiate din mulțimea de antrenament.

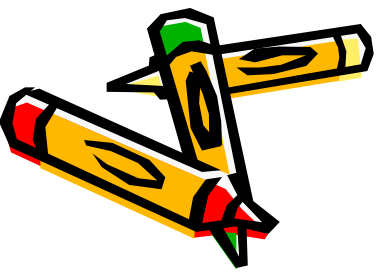
Se asociază mulțimii de antrenament o ***funcție distanță*** și o ***funcție de alegere*** a clasei de apartenență determinată de clasele de apartenență a vecinilor cei mai apropiați.



Algoritmul are ca parametru numărul k de vecini.

Se dă un eşantion de obiecte, a căror clasă de apartenență o cunoaştem $(\mathbf{x}, \Omega(\mathbf{x}))$, unde $\Omega(\mathbf{x})$, clasa căreia îi aparține obiectul $\mathbf{x}$.

Pentru un nou obiect $\mathbf{y}$, determinăm cele mai apropiate, în sensul distanței, k obiecte şi combinăm clasele cărora le aparțin într-o clasă Ω, care este clasa de apartenență a lui $\mathbf{y}$.



alegerea distantei

- în cazul valorilor continue, știm că:

  $d(x,y) = |x - y|$;

  în general se lucrează cu distanța normalizată:

  $$d(x,y) = \frac{|x - y|}{d},$$

  unde d este distanța maximă între două numere reale din domeniul considerat.

- în cazul a doi vectori cu *p* caracteristici $\mathbf{x} = (x_1,\ldots,x_p)$ și $\mathbf{y} = (y_1,\ldots,y_p)$, se calculează distanțele între caracteristici, $d_i(x_i, y_i)$ și apoi definim:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{d_1^2(x_1, y_1) + \ldots + d_p^2(x_p, y_p)},$$

sau

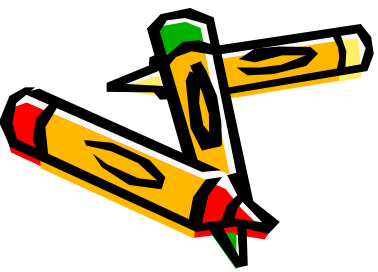
$$d(\mathbf{x}, \mathbf{y}) = d_1(x_1, y_1) + \ldots + d_p(x_p, y_p).$$

exemplu

Să considerăm obiectele:

$\mathbf{x} = (40,1,800)$, $\mathbf{y} = (30,0,1500)$, $\mathbf{z} = (45,1,2500)$,

unde prima componentă reprezintă atributul vârstă, a doua faptul că persoana este sau nu proprietara imobilului în care locuieşte şi a treia este atributul venit lunar.



Vom calcula distanțele dintre cele trei obiecte, nu înainte de a face normalizarea acolo unde este cazul.

$$d_1(x_1, y_1) = \frac{|40 - 30|}{d}, \quad d_1(x_1, z_1) = \frac{|40 - 45|}{d},$$

$$d_1(z_1, y_1) = \frac{|45 - 30|}{d},$$

unde $d = \max\{10, 5, 15\} = 15$

$$d_1(\mathbf{x}, \mathbf{y}) = \sqrt{\left(\frac{10}{15}\right)^2 + 1 + \left(\frac{7}{17}\right)^2} = 1.2704$$

$$d_1(\mathbf{y}, \mathbf{z}) = \sqrt{\left(\frac{15}{15}\right)^2 + 1 + \left(\frac{10}{17}\right)^2} = 1.5317$$

$$d_1(\mathbf{x}, \mathbf{z}) = \sqrt{\left(\frac{5}{15}\right)^2 + 1 + \left(\frac{17}{17}\right)^2} = 1.4530$$

Se observă că obiectele x şi y sunt cele mai apropiate între ele în sensul acestei distanțe

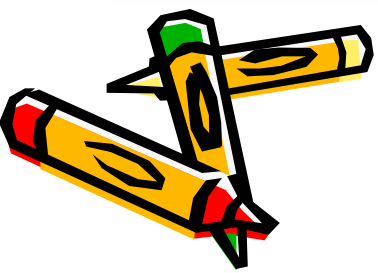
Rezultatul rămâne valabil şi dacă vom lua în considerare distanța

$d(\mathbf{x}, \mathbf{y}) = d_1(x_1, y_1) + d_2(x_2, y_2) + d_3(x_3, y_3)$, unde:

$d_2(x_2, y_2) = 1$, $d_2(x_2, z_2) = 0$, $d_2(y_2, z_2) = 1$

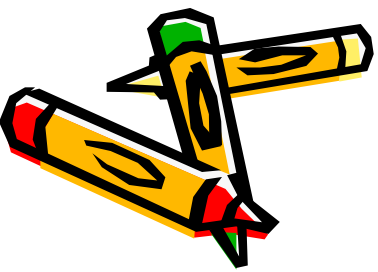
$$d_3(x_3, y_3) = \frac{700}{1700}, \quad d_3(x_3, z_3) = \frac{1700}{1700}, \quad d_3(y_3, z_3) = \frac{1000}{1700}$$

$d(\mathbf{x}, \mathbf{y}) = 2.0784$, $d(\mathbf{y}, \mathbf{z}) = 2.5002$, $d(\mathbf{x}, \mathbf{z}) = 1.4530$.



Metoda ***celui mai apropiat vecin*** ($k = 1$) poate fi astfel descrisă:

„având de clasificat un obiect $\mathbf{y}$, alegem cel mai apropiat obiect (în sensul distanței) din eșantionul dat, obiect a cărui apartenență o cunoaștem și atribuim lui $\mathbf{y}$ aceeași clasă."



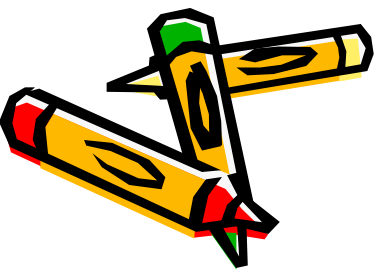
În general, se foloseşte metoda celor mai apropiaţi k vecini.

S-a dovedit experimental că o bună alegere a parametrului k este numărul cu 1 mai mare decât numărul atributelor.

Pentru clasificarea obiectului $\mathbf{y}$, determinăm

$(\mathbf{x}_1, \Omega(\mathbf{x}_1)),\ldots, (\mathbf{x}_k, \Omega(\mathbf{x}_k))$

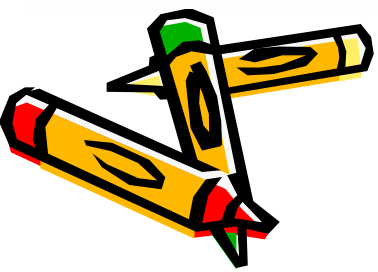
cele mai apropiate k obiecte şi clasele cărora le aparţin.



Pentru a găsi cărei clase îi aparține $\mathbf{y}$, avem variantele:

- alegem clasa ***majoritară***; în cazul unui număr par de clase se alege k impar;
- alegem clasa ***majoritară ponderată***. Fiecărei clase i se atribuie o anumită pondere: în general dacă $\mathbf{x}_i$ este vecinul considerat, ponderea atribuită clasei $\Omega(\mathbf{x}_i)$ este invers proporțională cu distanța dintre obiectul $\mathbf{y}$ și obiectul $\mathbf{x}_i$.

Este posibilă definirea ***încrederii*** în clasa atribuită lui $\mathbf{y}$, ca fiind raportul dintre numărul de apariții al clasei alese și k.

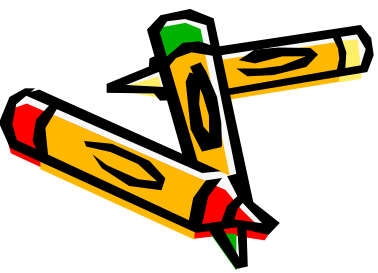


exemplul 1

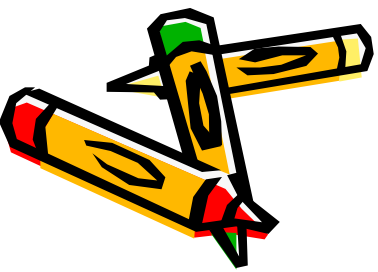
Prezentăm un exemplu referitor la evaluarea riscului bancar. Atributele iau valori continue, deci distanța va fi cea euclidiană.

Mulțimea de antrenament are doar 6 obiecte, fiecare având două atribute: *venit lunar* (lei), *rata credit lunara* deja existentă (lei), ceea ce ne va permite să reprezentăm aceste obiecte în spațiul bidimensional.

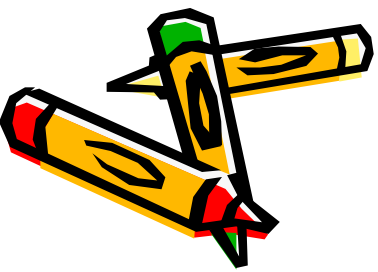
Există două clase: clienți ce prezintă risc scăzut şi respectiv clienți ce prezintă risc ridicat.



client	venit lunar (lei)	rata credit (lei)	risc
1	2500	500	scăzut
2	1500	200	scăzut
3	1200	400	ridicat
4	900	100	ridicat
5	2000	800	ridicat
6	1800	300	scăzut

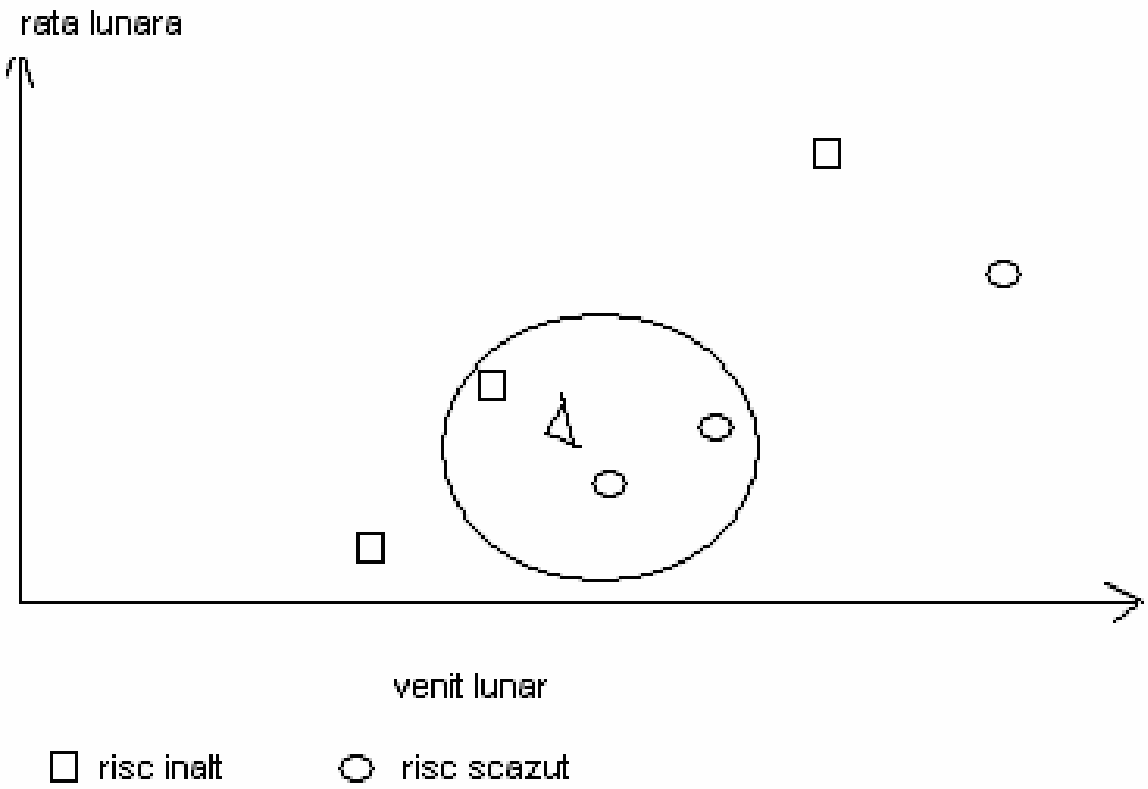


Pe baza acestei mulțimi de antrenament, folosind metoda *k-nearest neighbor*, vrem să evaluăm riscul prezentat de un nou client ce are un venit lunar de 1400 lei şi are o rată la un credit contractat anterior de 300 lei.



Calculăm distanțele între acest nou client şi cei existenți în baza de date:

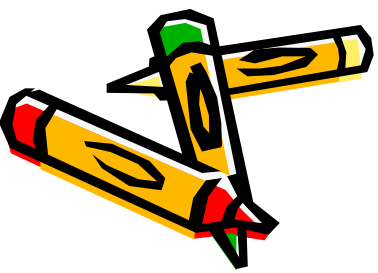
```
» a=[1400 300];b1=[2500 500]; n1=norm(a-b1)
      n1= 1.1180e+003
» a=[1400 300];b2=[1500 200];n2=norm(a-b2)
      n2=141.4214
» a=[1400 300];b3=[1200 400];n3=norm(a-b3)
      n3=223.6068
» a=[1400 300];b4=[900 100];n4=norm(a-b4)
      n4=538.5165
» a=[1400 300];b5=[2000 800];n5=norm(a-b5)
      n5=781.0250
» a=[1400 300];b6=[1800 300];n6=norm(a-b6)
      n6 =400
```

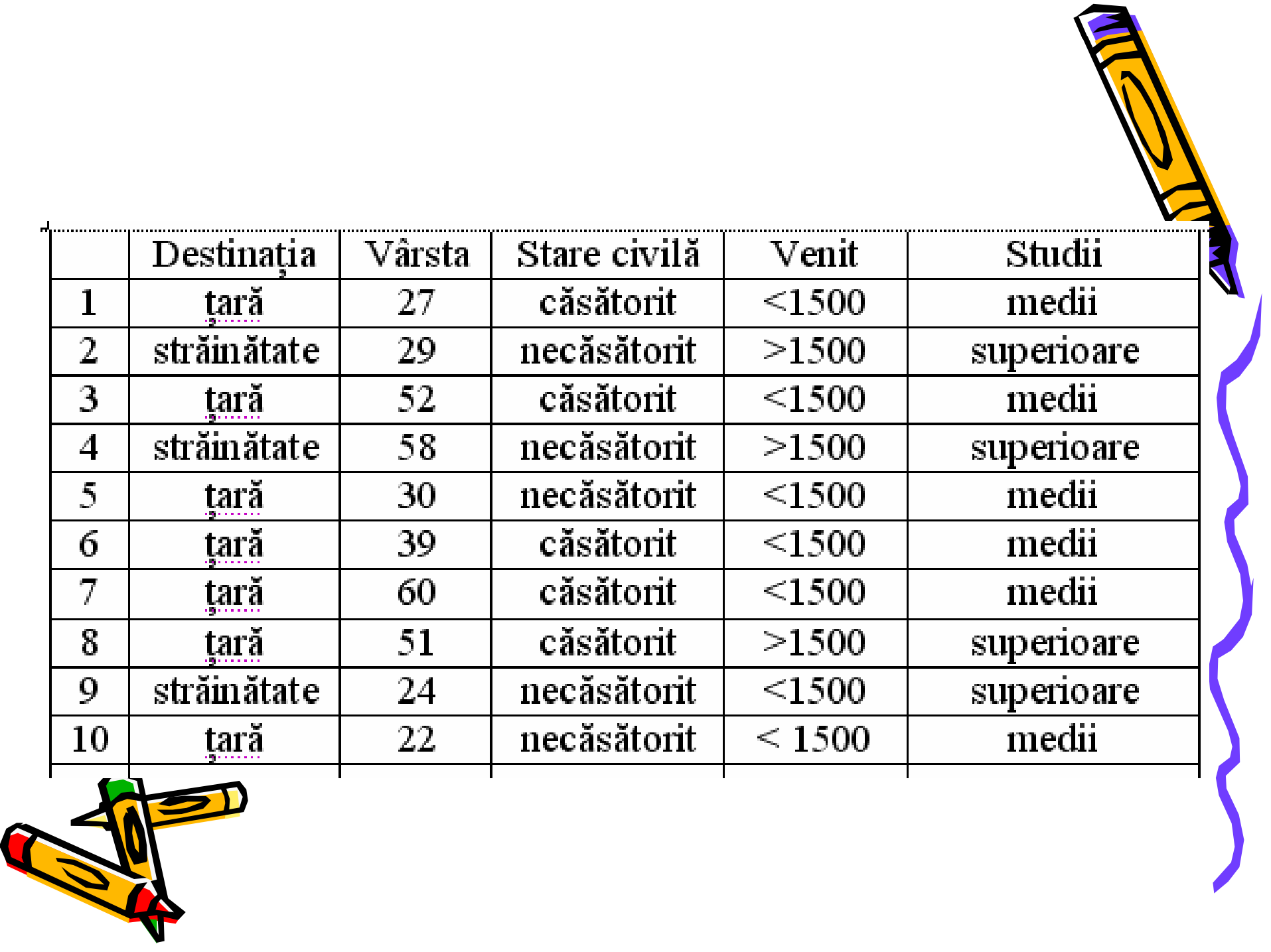


După cum se observă între cei trei cei mai apropiați vecini, doi prezintă risc scăzut, deci clientul nostru va prezenta un risc scăzut.

exemplul 2

Reluăm exemplul referitor la profilul clientului ce alege să-şi petreacă concediul în ţară sau străinătate.



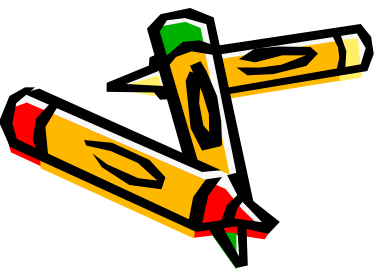


	Destinația	Vârsta	Stare civilă	Venit	Studii
1	țară	27	căsătorit	<1500	medii
2	străinătate	29	necăsătorit	>1500	superioare
3	țară	52	căsătorit	<1500	medii
4	străinătate	58	necăsătorit	>1500	superioare
5	țară	30	necăsătorit	<1500	medii
6	țară	39	căsătorit	<1500	medii
7	țară	60	căsătorit	<1500	medii
8	țară	51	căsătorit	>1500	superioare
9	străinătate	24	necăsătorit	<1500	superioare
10	țară	22	necăsătorit	< 1500	medii

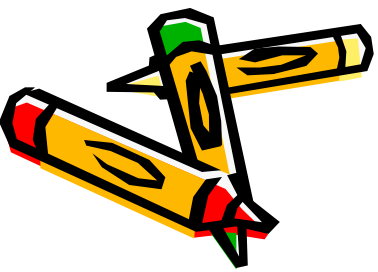
11	străinătate	64	căsătorit	>1500	superioare
12	străinătate	61	căsătorit	> 1500	superioare
13	înstrăinătate	29	căsătorit	> 1500	medii
14	țară	65	căsătorit	<1500	medii
15	țară	45	necăsătorit	< 1500	medii
16	străinătate	32	necăsătorit	>1500	medii
17	străinătate	34	căsătorit	< 1500	superioare
18	străinătate	38	necăsătorit	<1500	medii
19	țară	49	căsătorit	<1500	medii
20	țară	32	necăsătorit	< 1500	medii
21	țară	48	căsătorit	> 1500	superioare

Fiecare client (obiect) având 4 atribute (vârsta, starea civilă, venit lunar (lei), studii).

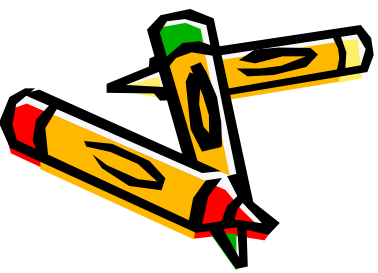
Să determinăm folosind metoda *k-nearest neighbor* unde își va petrece concediul un client în vârstă de 40 ani, căsătorit, cu studii superioare, cu un venit lunar de 1500 RON.



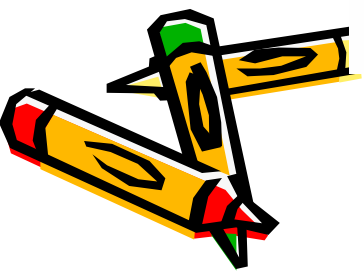
Conform celor menționate anterior luăm $k = 5$.
Atribuim valoarea 1 pentru căsătorit, pentru necăsătorit 0; analog pentru studii superioare 1, pentru studii medii 0. Pentru a calcula distanțele între obiectul nou $A(40,1,1,1500)$ şi obiectele Ai din bază este necesară normalizarea valorilor continue ale atributelor (deoarece atributele iau atât valori continue cât şi binare).



client	atribute	clasa	d(A,Ai)
A1	(27,1,0,1000)	în țară	1.0637
A2	(29,0 ,1,1000)	în străinătate	1.0392
A3	(52,1,0,1100)	în țară	1.0505
A4	(58,1,1,1600)	în străinătate	1.0848
A5	(30,0,0,800)	în țară	1.4603
A6	(31,1,0,1400)	în țară	1.0285
A7	(60,1,0,1000)	în țară	1.1209
A8	(51,1,1,1600)	**în țară**	**0.2589**
A9	(24,0,1,700)	în străinătate	1.1139
A10	(22,0,0,500)	în țară	1.5281



A11	(64,1,1,2500)	**în străinătate**	**0.6867**
A12	(61,1,1,2000)	**în străinătate**	**0.5277**
A13	(29,1,0,1800)	în străinătate	1.0392
A14	(65,1,0,800)	în țară	1.1901
A15	(45,0,0,900)	în țară	1.4391
A16	(32,0,0,2000)	în străinătate	1.4404
A17	(34,1,1,3000)	**în străinătate**	**0.6160**
A18	(38,0,0,1200)	în străinătate	1.4201
A19	(49,1,0,800)	în țară	1.0593
A20	(32,0,0,1000)	în țară	1.404
A21	(48,1,1,2200)	**în țară**	**0.3362**

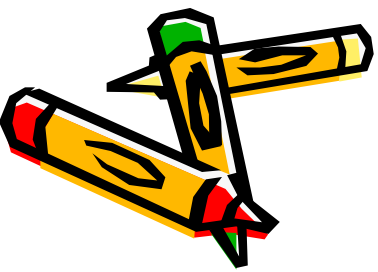


Distanțele au fost calculate după formula:

$$d(A,Ai)=\left(\left(\frac{A(1)-Ai(1)}{43}\right)^2+(A(2)-Ai(2))^2+\right.$$

$$\left.+(A(3)-Ai(3))^2+\left(\frac{A(4)-Ai(4)}{2500}\right)^2\right)^{\frac{1}{2}}$$

unde $A(j)$, $Ai(j)$ sunt notațiile pentru a j-a caracteristică, a obiectului nou A, respectiv a obiectului Ai, $j=1,\ldots,4$.

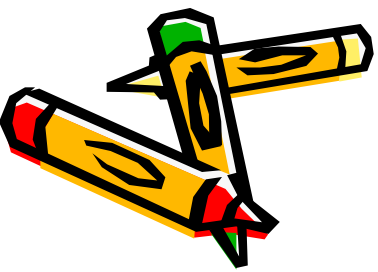


Calculând: $\max_{1 \le i \le 21} |A(1) - Ai(1)| = 43$,

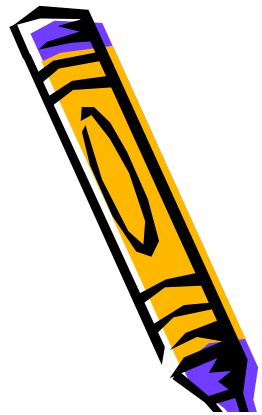
respectiv $\max_{1 \le i \le 21} |A(4) - Ai(4)| = 2500$,

am normalizat valorile atributelor 1 şi 4.

În concluzie din cei 5 vecini cei mai apropiați, votul majoritar spune că noul client își va petrece concediul în străinătate.



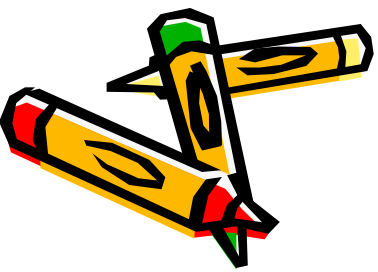
consideratii asupra metodei



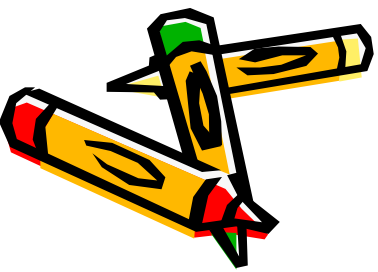
- nu necesită faza de antrenament;
- trebuie acordat maximum de atenție la alegerea atributelor, pentru a obține o bună clasificare;
- este necesar ca numărul de obiecte din eşantionul luat în considerare să fie suficient de mare în raport cu numărul atributelor; fiecare clasă trebuie să fie bine reprezentată.

avantaje

- se pot introduce în eşantionul considerat iniţial noi dat, ceea ce îmbunătăţeşte rezultatele şi nu necesită modificarea modelului;
- rezultatele sunt clare;

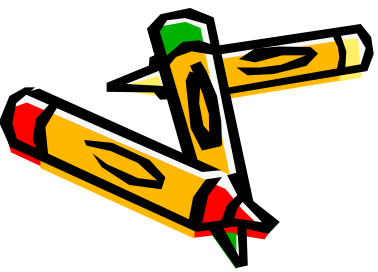


- metoda este aplicabilă la orice tip de date pentru care se pot defini distanțe, inclusiv pentru informații geografice, texte, imagini, sunete;
- obiectele din eșantion pot avea un număr mare de atribute, caz în care numărul obiectelor trebuie să fie mai mare. Pentru un număr mai mic de obiecte este necesară alegerea unor atribute definitorii.



dezavantaje

- se stochează în memorie întreg eșantionul;
- timpul de clasificare este mare, deoarece calculele se efectuează în timpul clasificării.



În general, distanțele simple funcționează bine.
Dacă nu, alegem alt parametru k: în caz de rezultat nesatisfăcător se alege altă distanță.
Dacă nici aceste modificări nu sunt suficiente este cazul să alegem altă metodă.