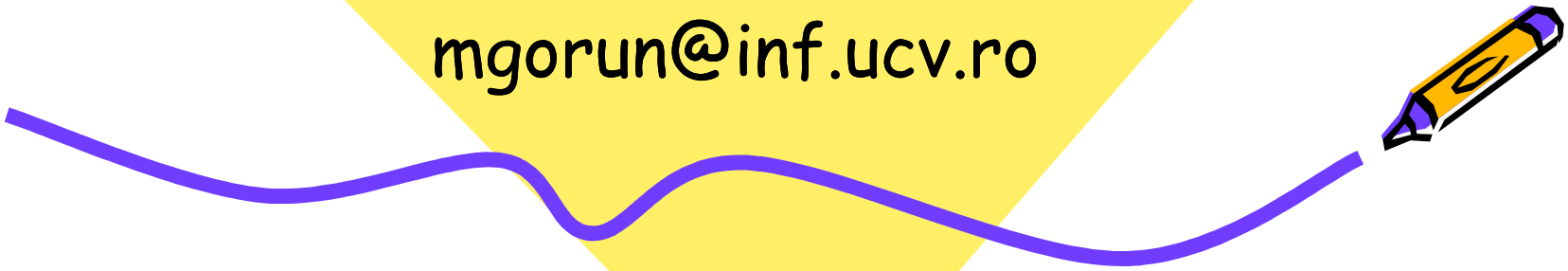
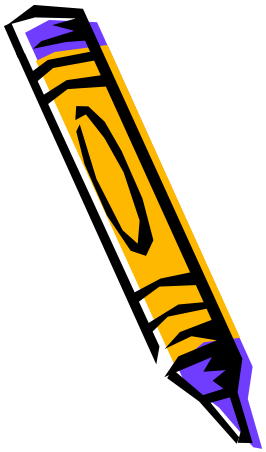
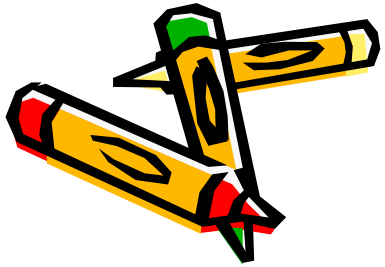
K nearest neighbor

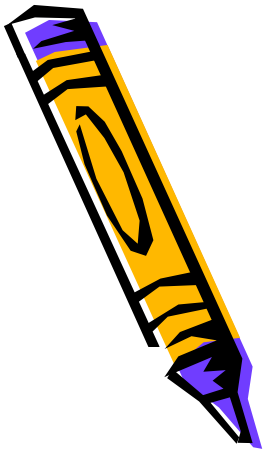
Marina Gorunescu
mgorun@inf.ucv.ro





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.

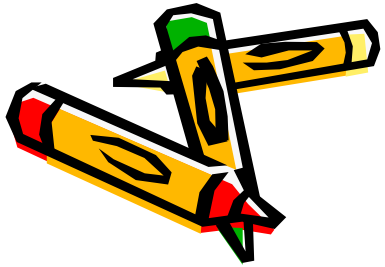




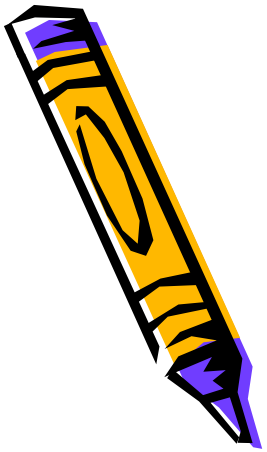
Algoritmul are ca parametru numărul k de vecini.

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



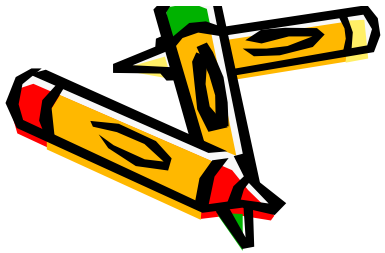
- în cazul valorilor continue, știm că:

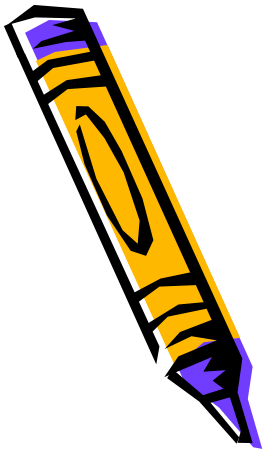
  $d(x,y) = |x - y|$;

  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.



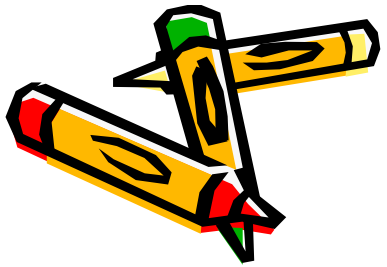


- în cazul a doi vectori cu *p* caracteristici $\mathbf{x} = (x_1,\ldots,x_p)$ și $\mathbf{y} = (y_1,\ldots,y_p)$, se calculează distanțele între caracteristici, $d_i(x_i, y_i)$ și apoi definim:

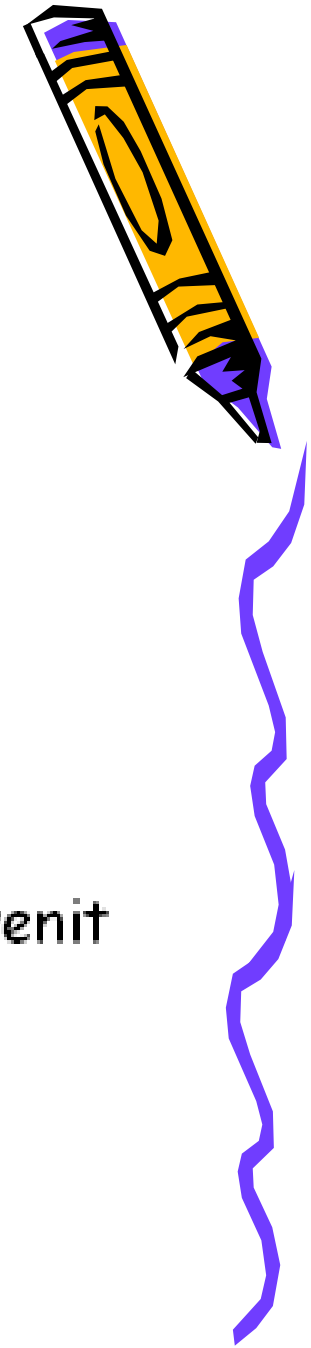
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{d_1^2(x_1, y_1) + \ldots + d_p^2(x_p, y_p)},$$

sau

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



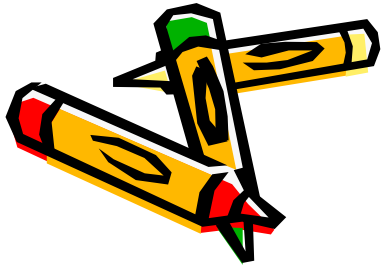
exemplu

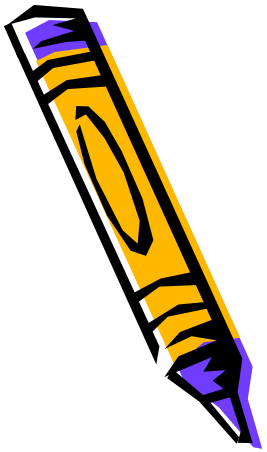


Să considerăm obiectele:

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

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.



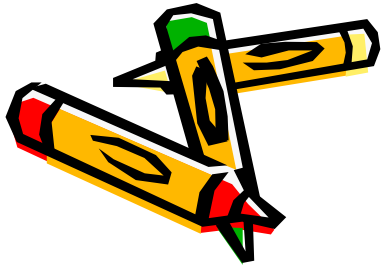


Vom calcula distanțele dintre cele trei obiecte, nu înainte de a face normalizarea acolo unde este cazul.

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

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

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

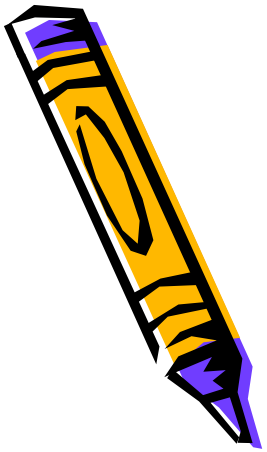


$$d_1(\mathbf{x}, \mathbf{y}) = \sqrt{\left(\frac{10}{15}\right)^2 + 1 + \left(\frac{7}{17}\right)^2} = 1.2704$$

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

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

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



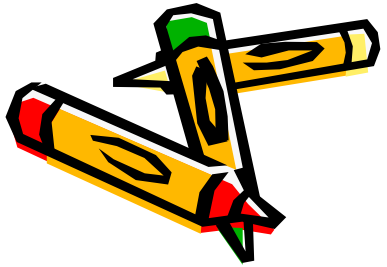
Rezultatul rămâne valabil şi dacă vom lua în considerare distanţa

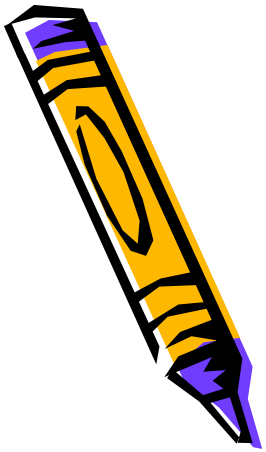
$d(\mathbf{x}, \mathbf{y}) = d_1(x_1, y_1) + d_2(x_2, y_2) + d_3(x_3, y_3)$, unde:

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

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

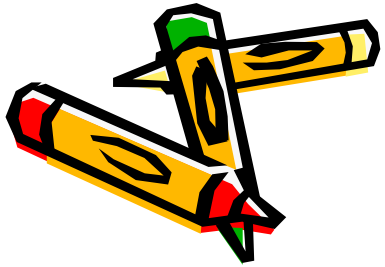
$d(\mathbf{x}, \mathbf{y}) = 2.0784$, $d(\mathbf{y}, \mathbf{z}) = 2.5002$, $d(\mathbf{x}, \mathbf{z}) = 1.4530$.

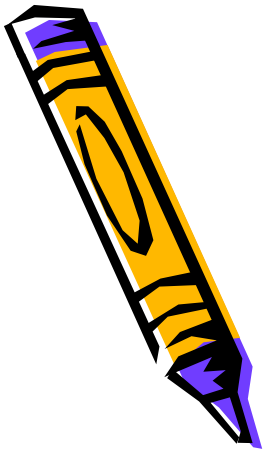




Metoda ***celui mai apropiat vecin*** ($k=1$) poate fi astfel descrisă:

„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."





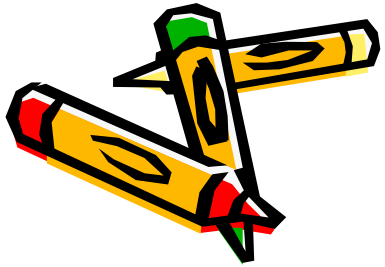
În general, se foloseşte metoda celor mai apropiaţi k vecini.

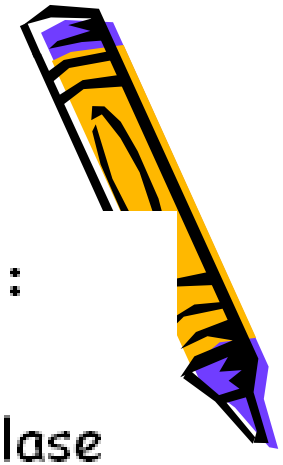
S-a dovedit experimental că o bună alegere a parametrului k este numărul cu 1 mai mare decât numărul atributelor.

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.

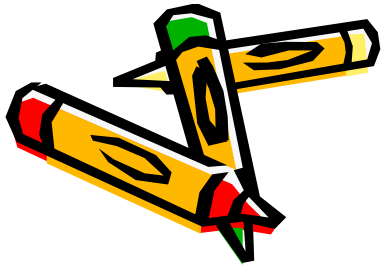




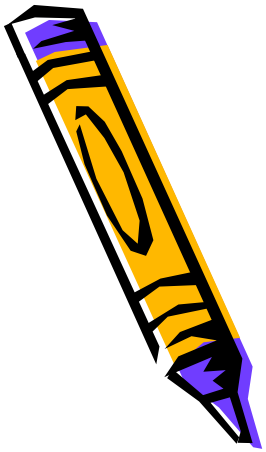
Pentru a găsi cărei clase îi aparține **y**, avem variantele:

- 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 **y** și obiectul $\mathbf{x}_i$.

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

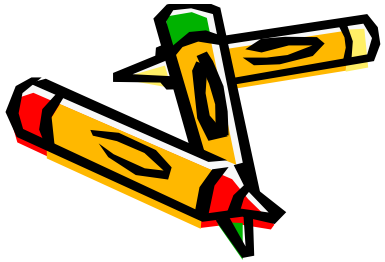


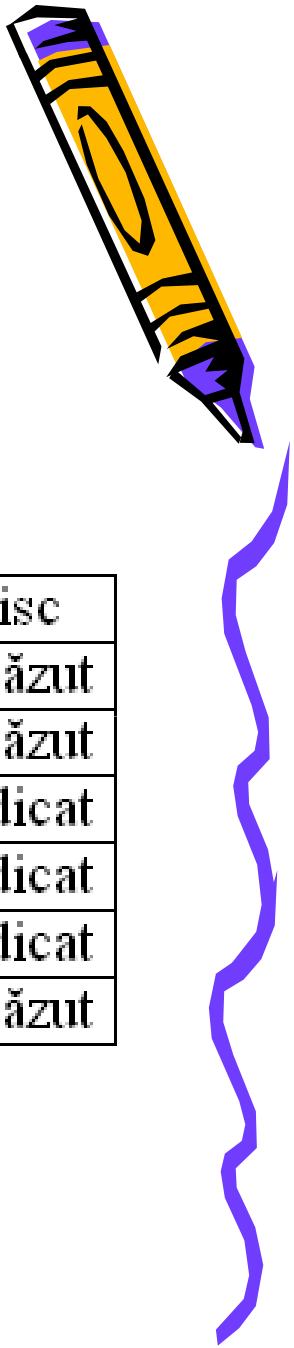
exemplul 1



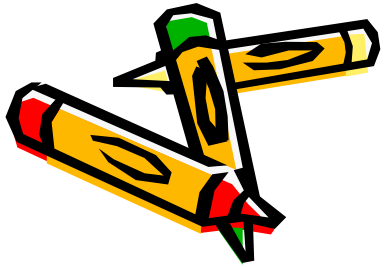
Prezentăm un exemplu referitor la evaluarea riscului bancar. Atributele iau valori continue, deci distanța va fi cea euclidiană. Mulțimea de antrenament are doar 6 obiecte, fiecare având două atribute: *venit lunar* (RON), *rata credit lunara* deja existentă (RON), ceea ce ne va permite să reprezentăm aceste obiecte în spațiul bidimensional.

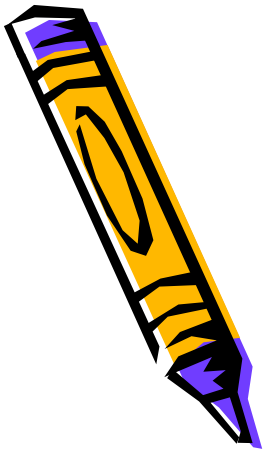
Există două clase: clienți ce prezintă risc scăzut şi respectiv clienți ce prezintă risc ridicat.





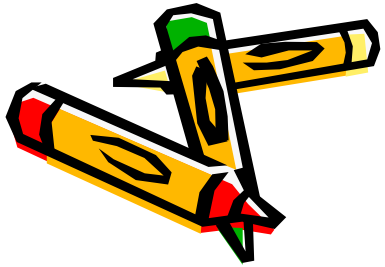
client	venit lunar (RON)	rata credit (RON)	risc
1	2500	500	scăzut
2	1500	200	scăzut
3	1200	400	ridicat
4	900	100	ridicat
5	2000	800	ridicat
6	1800	300	scăzut





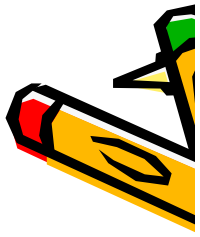
Pe baza acestei mulțimi de antrenament, folosind metoda *k-nearest neighbor*, vrem să evaluăm riscul prezentat de un nou client ce are un venit lunar de 1400 RON și are o rată la un credit contractat anterior de 300 RON.

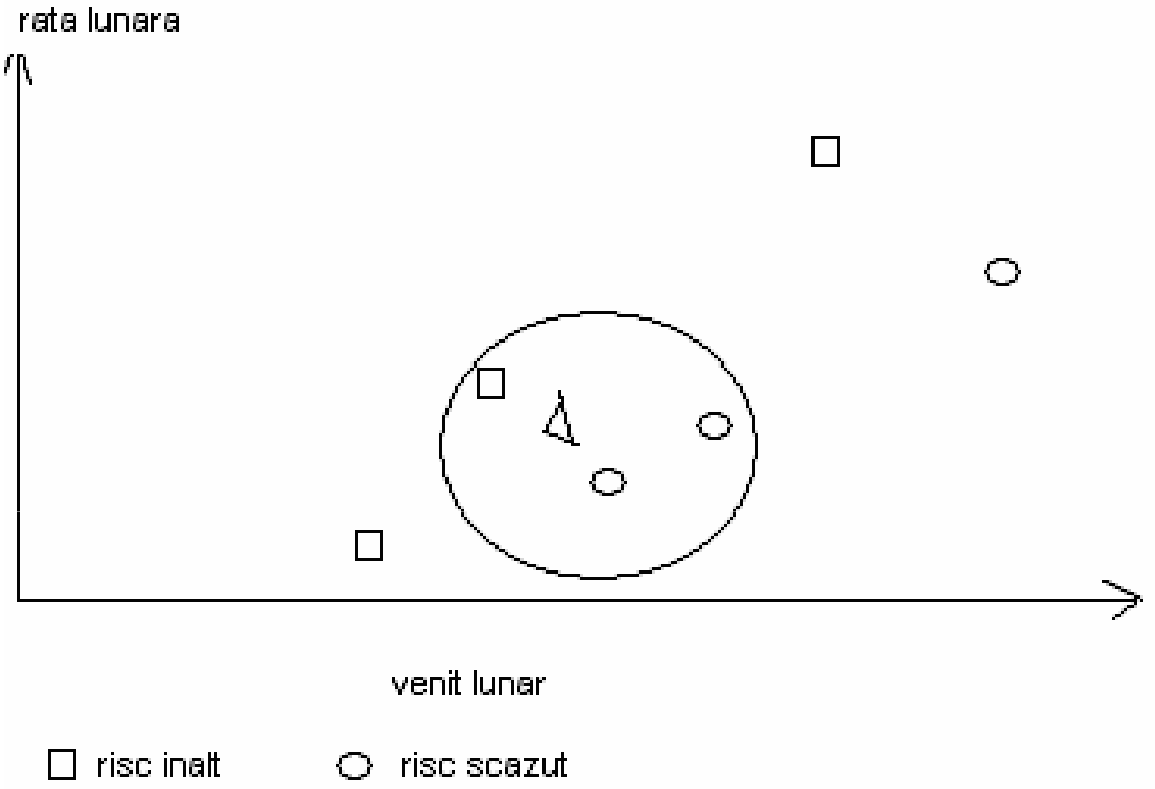
Deoarece obiectele au două atribute, vom considera $k = 3$.



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
```

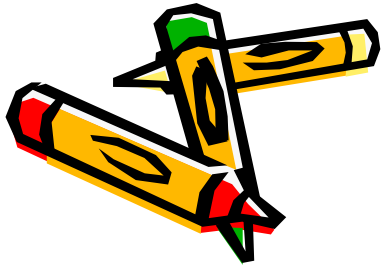


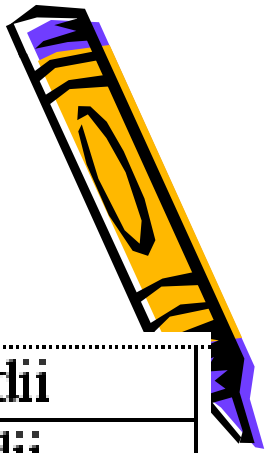


După cum se observă între cei trei cei mai apropiați vecini, doi prezintă risc scăzut, deci clientul nostru va prezenta un risc scăzut.

exemplul 2

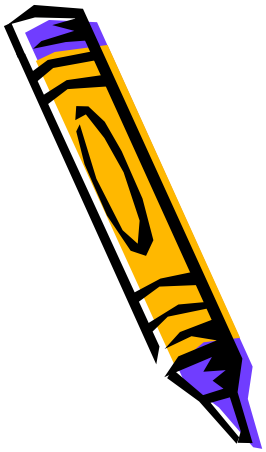
Reluăm exemplul referitor la profilul clientului ce alege să-şi petreacă concediul în ţară sau străinătate.





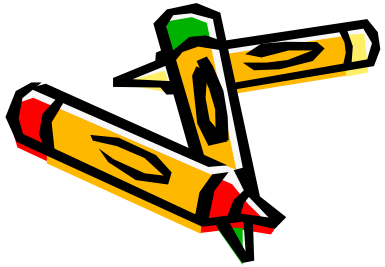
	Destinația	Vârsta	Stare civilă	Venit	Studii
1	țară	27	căsătorit	<1500	medii
2	străinătate	29	necăsătorit	>1500	superioare
3	țară	52	căsătorit	<1500	medii
4	străinătate	58	necăsătorit	>1500	superioare
5	țară	30	necăsătorit	<1500	medii
6	țară	39	căsătorit	<1500	medii
7	țară	60	căsătorit	<1500	medii
8	țară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

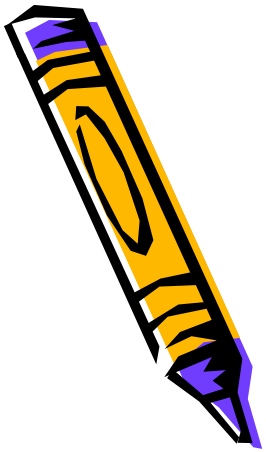
11	străinătate	64	căsătorit	>1500	superioare
12	străinătate	61	căsătorit	> 1500	superioare
13	îstrăinătate	29	căsătorit	> 1500	medii
14	țară	65	căsătorit	<1500	medii
15	țară	45	necăsătorit	< 1500	medii
16	străinătate	32	necăsătorit	>1500	medii
17	străinătate	34	căsătorit	< 1500	superioare
18	străinătate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



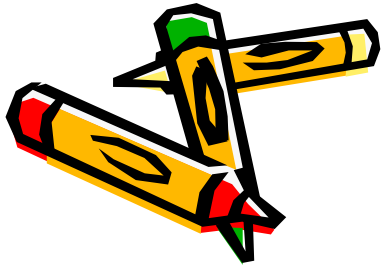
Fiecare client (obiect) având 4 atribute (vârsta, starea civilă, venit lunar (RON), studii).

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.

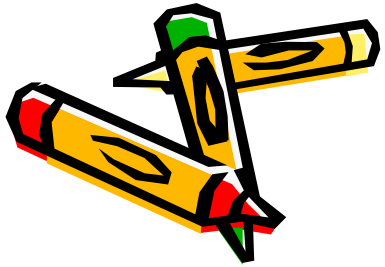




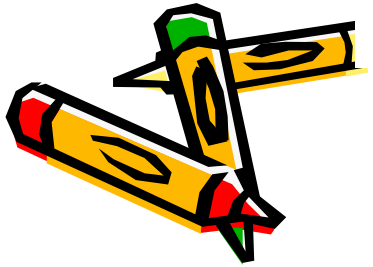
Conform celor menționate anterior luăm $k = 5$.
Atribuim valoarea 1 pentru căsătorit, pentru necăsătorit 0; analog pentru studii superioare 1, pentru studii medii 0. Pentru a calcula distanțele între obiectul nou $A(40,1,1,1500)$ și obiectele Ai din bază este necesară normalizarea valorilor continue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

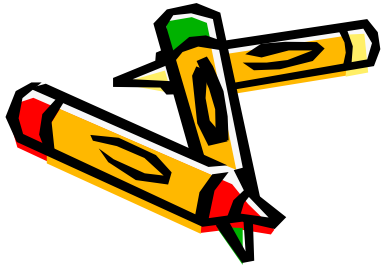


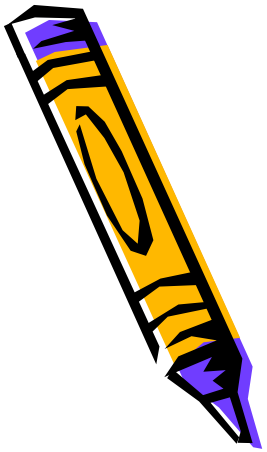
Distanțele au fost calculate după formula:

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

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

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



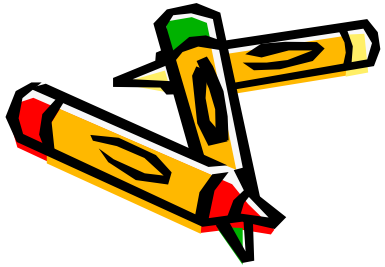


Calculând: $\max_{1 \le i \le 21} |A(1) - Ai(1)| = 43$,

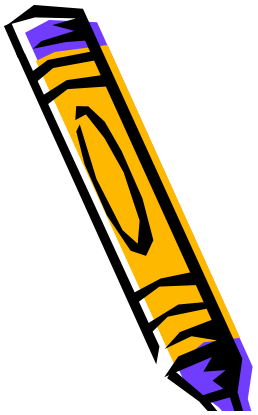
respectiv $\max_{1 \le i \le 21} |A(4) - Ai(4)| = 2500$,

am normalizat valorile atributelor 1 şi 4.

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



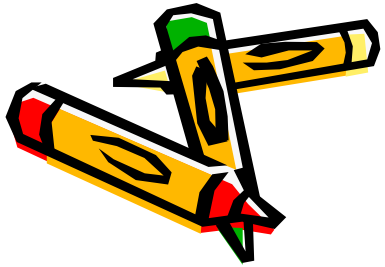
consideratii asupra metodei

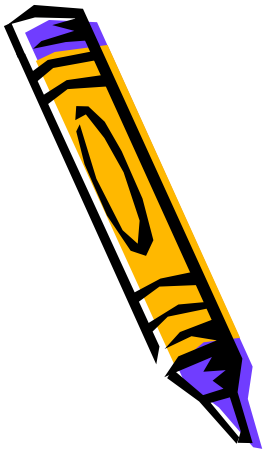


- nu necesită faza de antrenament;
- trebuie acordat maximum de atenție la alegerea atributelor, pentru a obține o bună clasificare;
- este necesar ca numărul de obiecte din eşantionul luat în considerare să fie suficient de mare în raport cu numărul atributelor; fiecare clasă trebuie să fie bine reprezentată.

avantaje

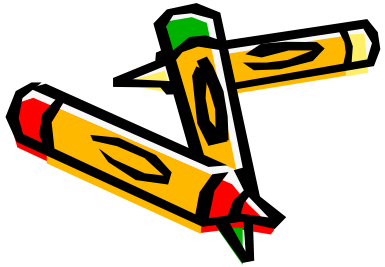
- se pot introduce în eşantionul considerat iniţial noi dat, ceea ce îmbunătăţeşte rezultatele şi nu necesită modificarea modelului;
- rezultatele sunt clare;





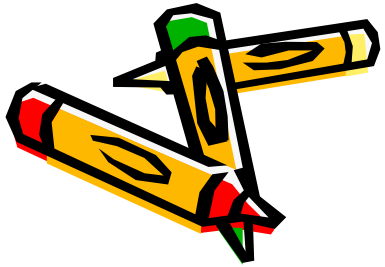
- metoda este aplicabilă la orice tip de date pentru care se pot defini distanțe, inclusiv pentru informații geografice, texte, imagini, sunete;
- obiectele din eșantion pot avea un număr mare de atribute, caz în care numărul obiectelor trebuie să fie mai mare.

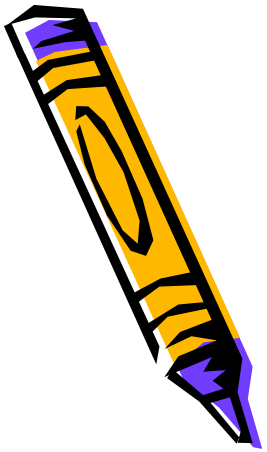
  Pentru un număr mai mic de obiecte este necesară alegerea unor atribute definitorii.



dezavantaje

- se stochează în memorie întreg eşantionul;
- 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.