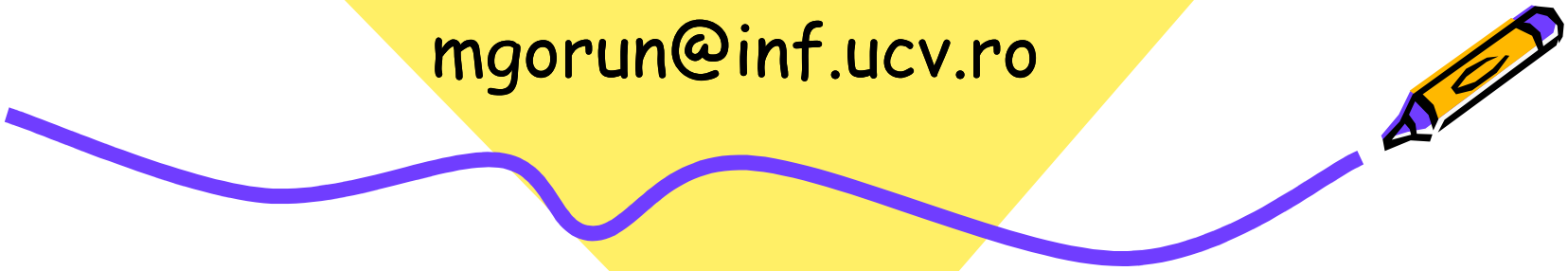
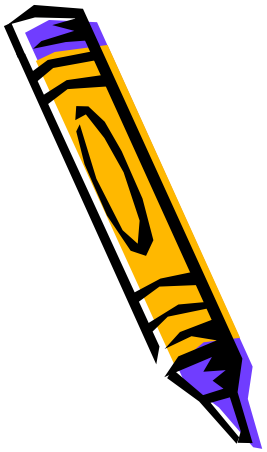
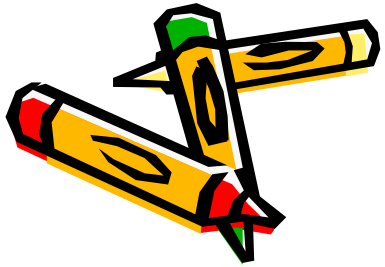
Clasificarea bazata pe reguli de asociere

Marina Gorunescu
mgorun@inf.ucv.ro

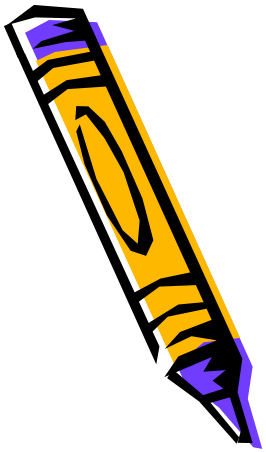




Metoda regulii de asociere (sau analiza asocierii) este o tehnică **nesupervizată**, care caută legături între înregistrările dintr-un set de date.

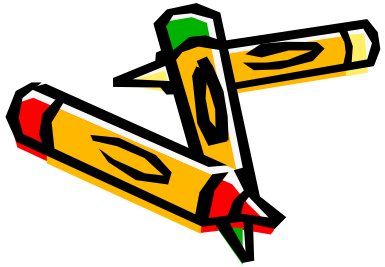


regula de asociere



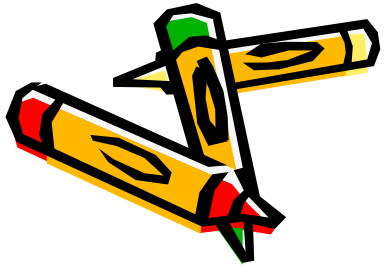
O *regulă de asociere* (*association rule*) este o expresie de implicare de felul „dacă (***IF***) X atunci (***THEN***) Y",
unde articolele (item-uri) X și Y sunt distincte,
$X \cap Y = \phi$.

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

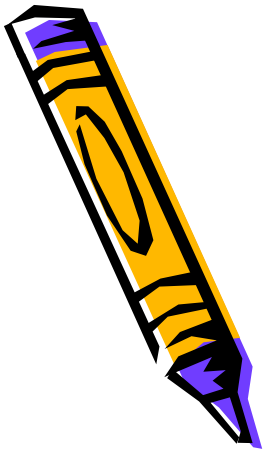


metode de construire a regulilor de asociere

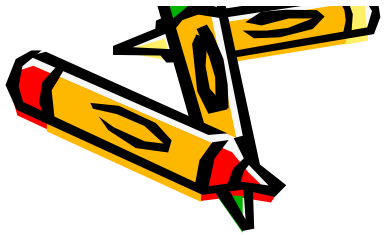
- metoda *indirectă*, care constă în extragerea regulilor folosind alți clasificatori, de exemplu arborii de clasificare și decizie;
- metoda *directă*, care constă în extragerea regulilor direct din date.



regulile de asociere pot fi:

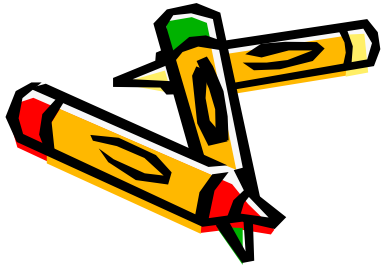


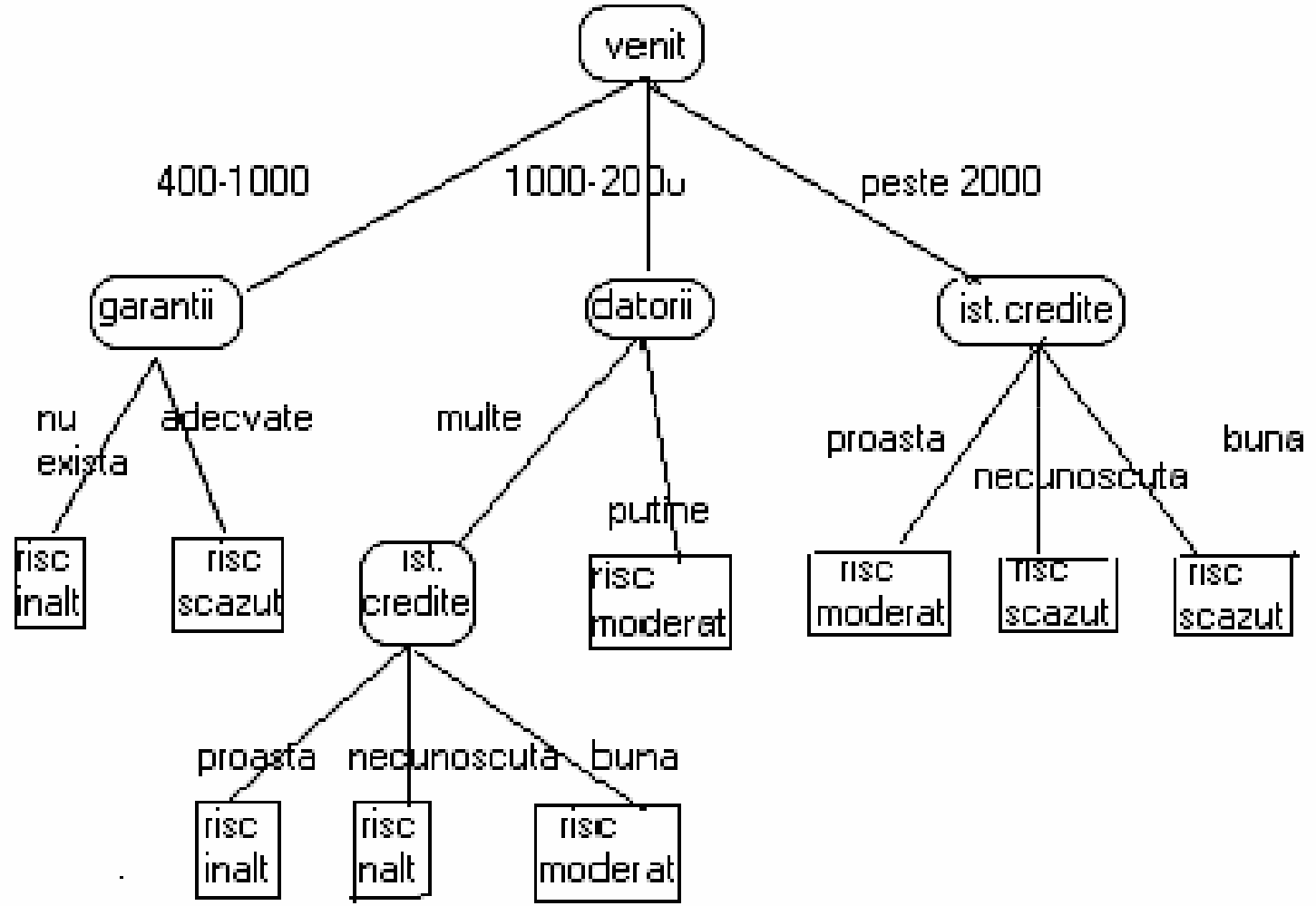
- reciproc *exclusive*, adică regulile sunt independente una de alta, fiecare articol fiind acoperit de cel puțin o regulă
- *exhaustive*, caz în care se consideră toate combinațiile posibile ale valorilor atributelor şi fiecare articol este acoperit de cel puțin o regulă.

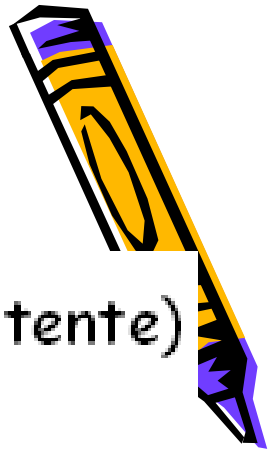


exemplu

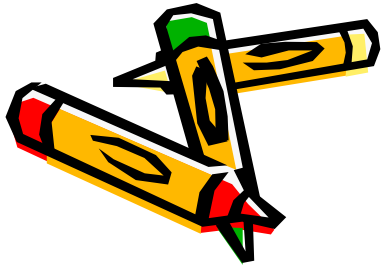
regulile de asociere pe baza arborelui de decizie, construit pentru exemplul referitor la estimarea riscului acordării unui credit




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 inalt
risc moderat

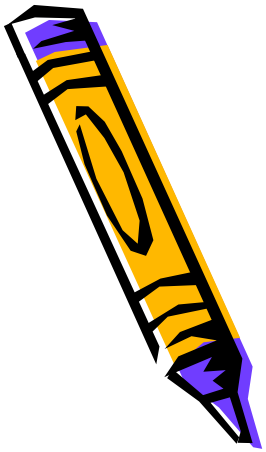


- **R1:** (venit lunar 400-1000 RON) și (garanții inexistente) ⇒(risc înalt).
- **R2:** (venit lunar 400-1000 RON) și (garanții adecvate)
- ⇒(risc scăzut).
- **R3:** (venit lunar 1000-2000 RON) și (datorii multe) și (istoria creditelor proastă) ⇒(risc înalt).
- **R4:** (venit lunar 1000-2000 RON) și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



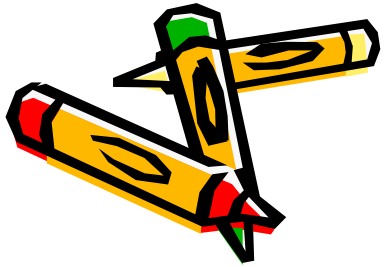
- **R6:** (venit lunar 1000-2000 RON) și (datorii puține) ⇒(risc moderat).
- **R7:** (venit lunar peste 2000 RON) și (istoria creditelor proastă) ⇒(risc moderat).
- **R7:** (venit lunar peste 2000 RON) și (istoria creditelor necunoscută) ⇒(risc scăzut).
- **R8:** (venit lunar peste 2000 RON) și (istoria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

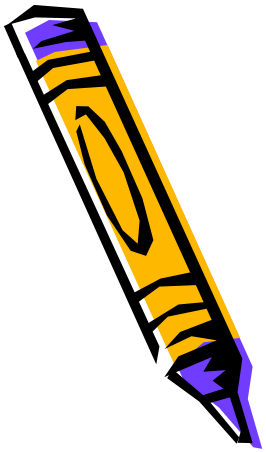
sunt reguli reciproc exclusive și exhaustive.



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.

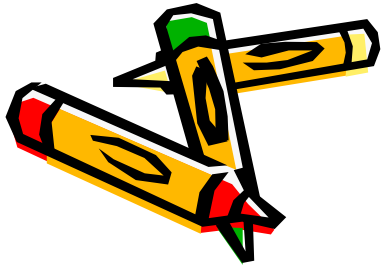




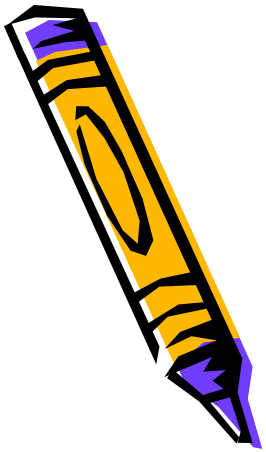
În exemplul referitor la estimarea riscului acordării unui credit, în cazul regulii:

„ datorii multe $\Rightarrow$ risc înalt"

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



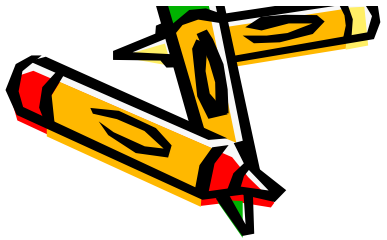
analiza cosului de consum



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

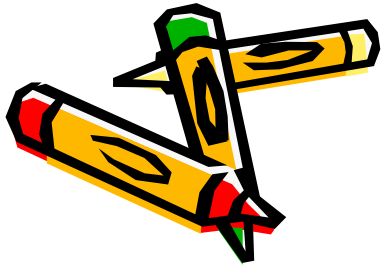
Analiza coşului de consum constă în găsirea de asocieri între produsele afişate pe bonurile de casă.

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.

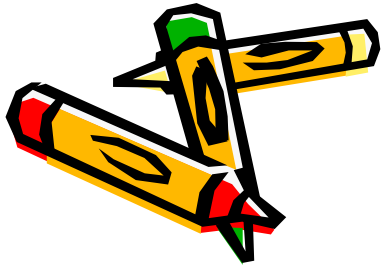
Poate fi folosită în domeniul medical pentru studiul complicațiilor apărute datorită asocierii unor medicamente sau în domeniul fraudelor, caz în care se caută asocierii neobișnuite.



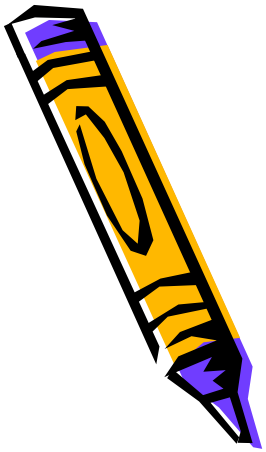
Rezultatele metodei sunt regulile de asociere, care sunt utile în marketing.

De cele mai multe ori metoda poate produce reguli interesante.

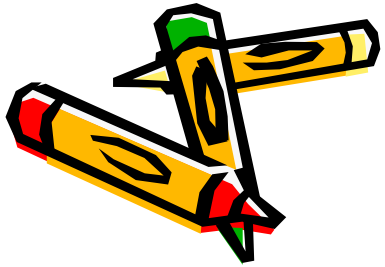
Uneori se obțin și reguli triviale.

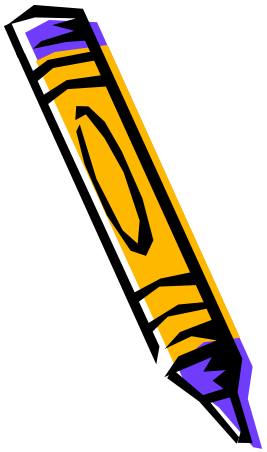


exemple



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

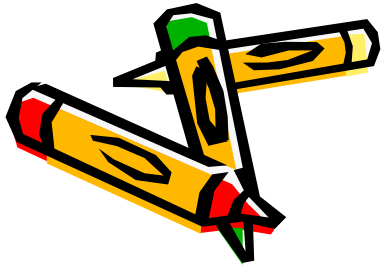




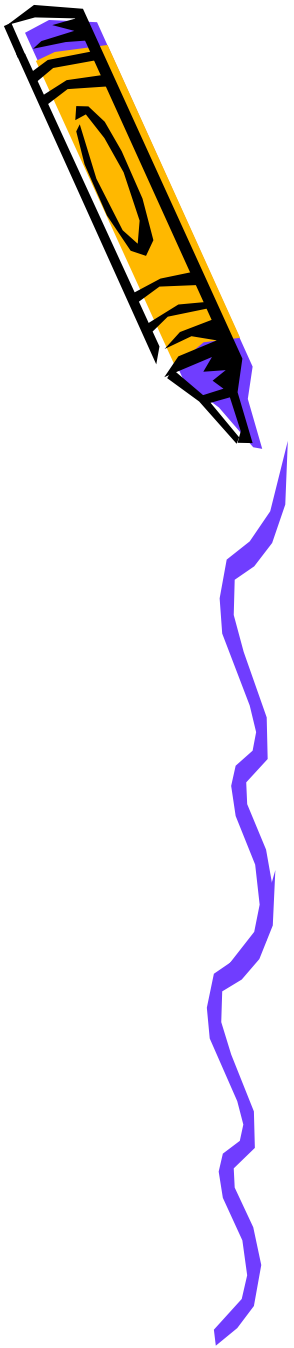
Considerăm următoarele date, provenite dintr-o listă de 10 clienți.

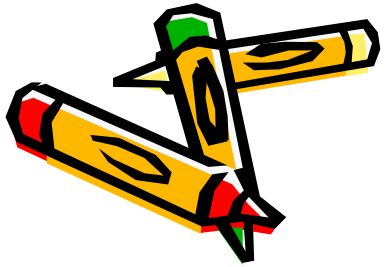
Un client a cumpărat o listă de articole, listă de lungime variabilă.

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



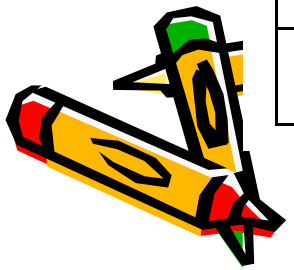
	A	B	C	D	E	F	G
1	×		×			×	
2	×	×		×			
3		×	×			×	
4		×		×		×	
5	×						×
6					×	×	
7	×		×				
8			×		×		
9	×	×	×	×		×	
10	×			×		×	



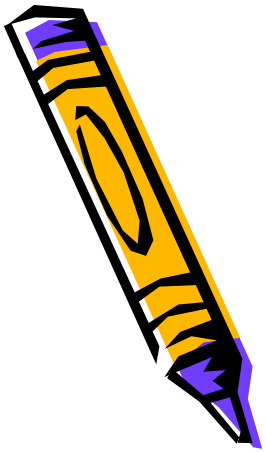


Creăm un tabel care să ilustreze de câte ori două produse sunt cumpărate simultan de un client:

	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:

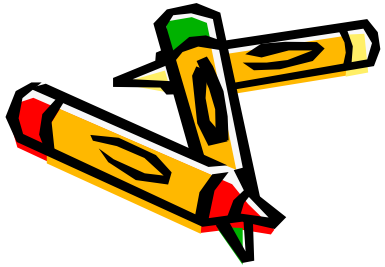
„dacă este îndeplinit *antecedentul* atunci avem *consecința*."

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

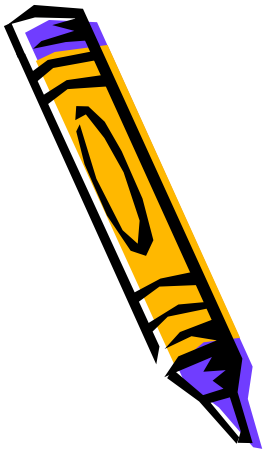
notând cu x numărul de clienți care cumpără simultan articolele din antecedent și consecință și cu n numărul total de clienți, suportul regulii **R** este:

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

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

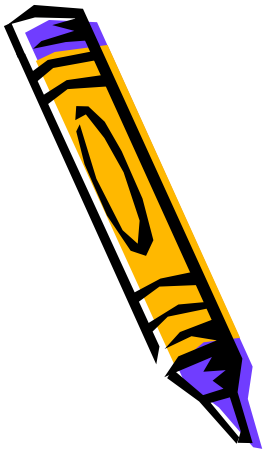


Considerăm regulile, obținute din tabelul prezentat anterior:

- **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*.

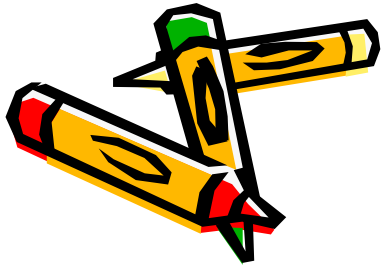
Astfel pentru regula 2 avem:

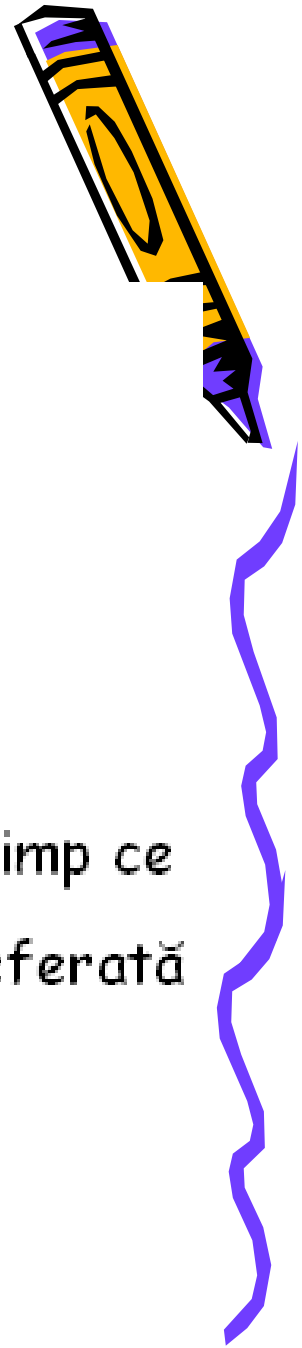
$confidence(\mathbf{R}2) = \frac{3}{5}$,

în timp ce pentru regula 3 avem

$confidence(\mathbf{R}3) = \frac{3}{6}$.

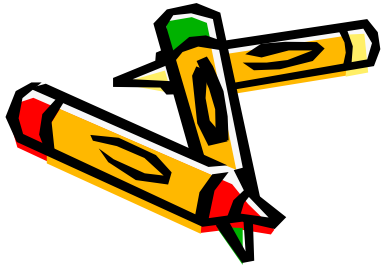
În concluzie, regula 3 prezintă o încredere de 50%, în timp ce regula 2 prezintă o încredere de 60% şi astfel va fi preferată regula 2.

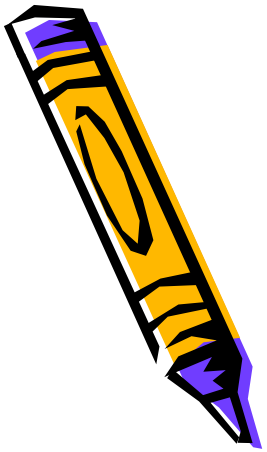




Putem spune că *încrederea* măsoară cât de mult depinde un articol de altul.

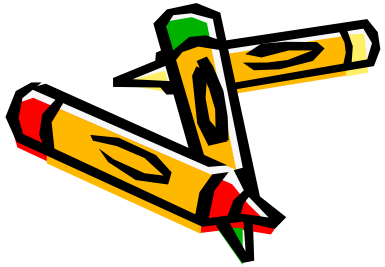
Să reținem: dintr-o mulțime de reguli ce au un suport suficient de mare, se alege aceea ce prezintă încrederea maximă.

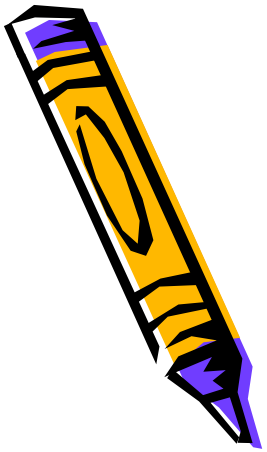




Cele două caracteristici, suportul şi încrederea, nu sunt suficiente pentru robustețea regulii de asociere.
Vom considera articolele *A*, *B* şi *C* şi frecvențele lor de apariție:

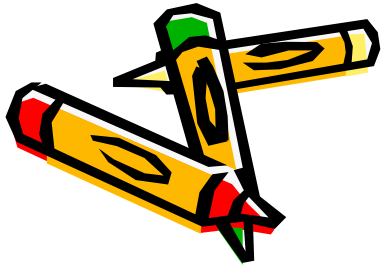
art.	A	B	C	A&B	A&C	B&C	A&B&C
frecv.	45%	42.%	40%	25%	20%	15%	5%

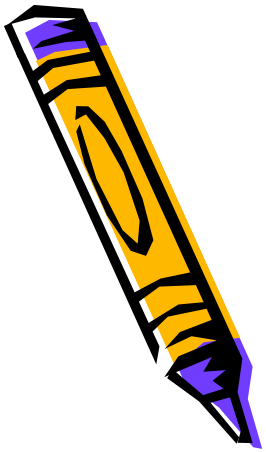




Dacă vom considera reguli cu trei articole, acestea vor avea același suport de 5%. Nivelul de încredere este:

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

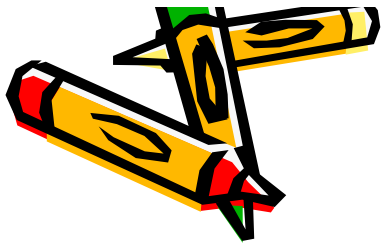


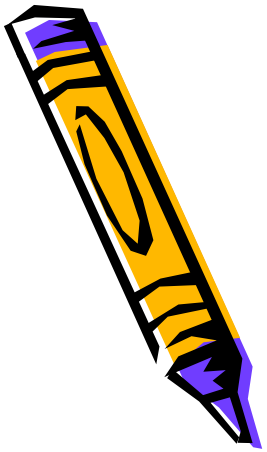


Regula „dacă B şi C atunci A" prezintă cea mai mare încredere de 0.33, ceea ce înseamnă că:

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%.

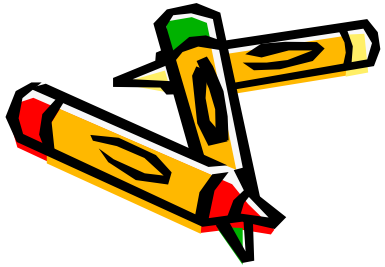
Dacă studiem tabelul, observăm că A apare în 45% din coşurile de cumpărături, ceea ce înseamnă că este preferabil să prognozăm apariţia lui A, decât apariţia simultană a articolelor B şi C.

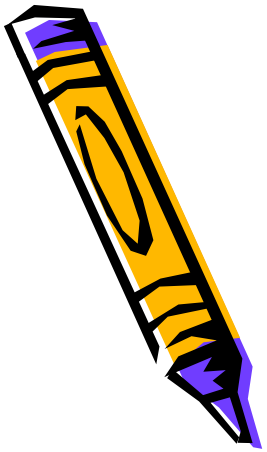




diferența de nivel (*lift*) permite compararea rezultatului predicției folosind regula, cu predicția obținută fără utilizarea regulii. *Diferența de nivel* este definită prin:

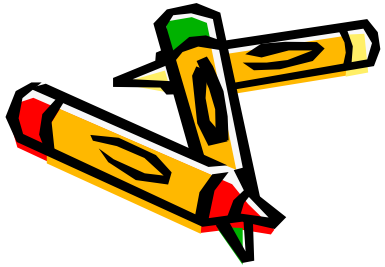
$$\text{Diferența de nivel} = \frac{confidence}{frecventa\ (\text{consecinta})}$$

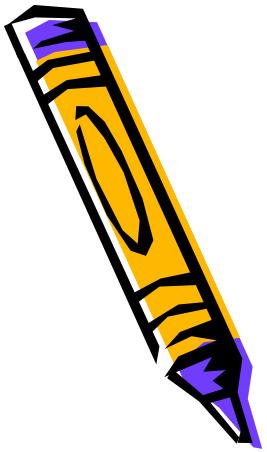




O regulă este interesantă dacă diferența de nivel este mai mare decât 1.

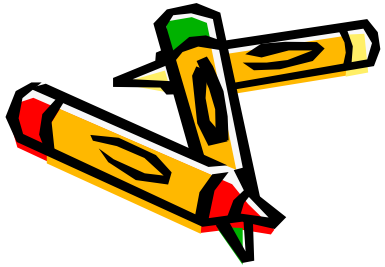
regula	confidence	frecvență	dif.de nivel
dacă A şi B atunci C	0.20	40%	0.50
dacă A şi C atunci B	0.25	42%	0.59
dacă B şi C atunci A	0.33	45%	0.74



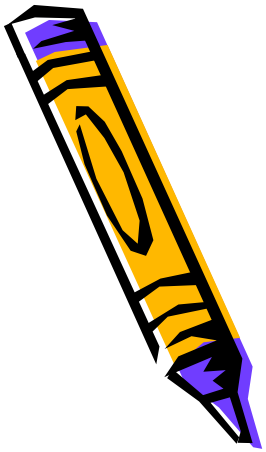


Regula „dacă A atunci B" are un suport de 25%, o încredere de $\frac{25}{45} = 0.55$ şi o diferență de nivel de $\frac{55}{42} = 1.3$.

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

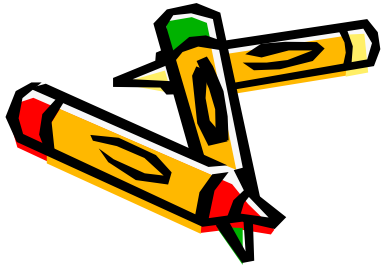


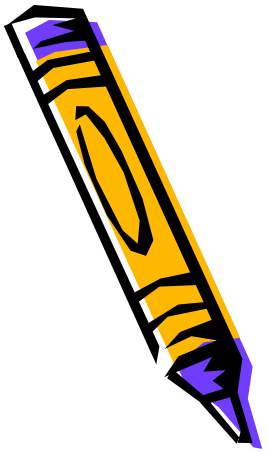
exemplu



- 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).





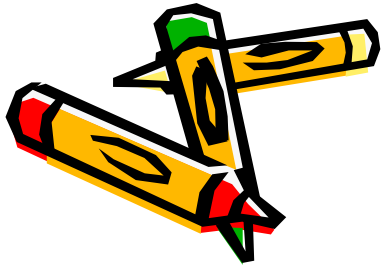
- *Suportul* este 0.4% din total (2.000/500.000);
  - Regula

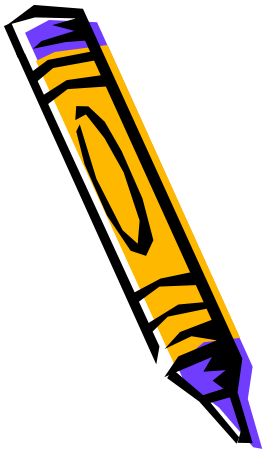
    „Când oamenii cumpără whisky, cumpără de asemenea şi bere"

    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).

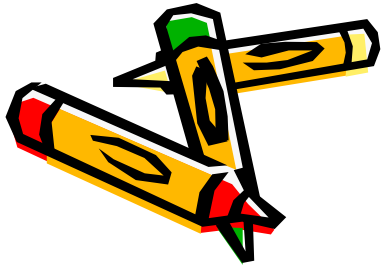




Cele două reguli au același suport 0.4% și aceeași diferență de nivel 6.66.

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

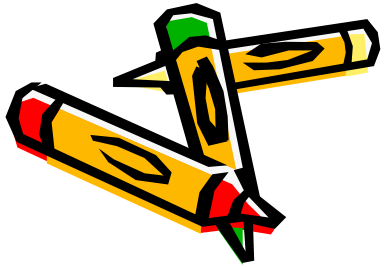
- 1% din clienții cumpără whisky;
- 6% din clienții cumpără bere.

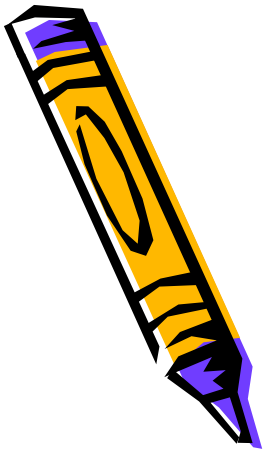


Cele două procentaje, 1% și 6%, sunt numite *încrederea așteptată* de a cumpăra whisky sau bere, indiferent de celelalte cumpărături.

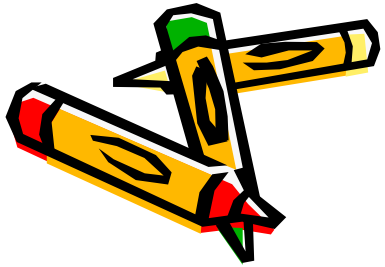
Regula de cumpărare *whisky-bere* poate fi exprimată în termen de *diferență de nivel* astfel:

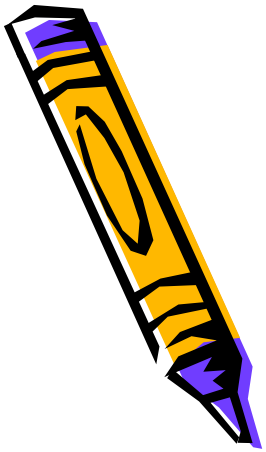
„Clienții care cumpără whisky sunt de 6,66 *ori mai tentați să cumpere și bere odată cu whisky*".



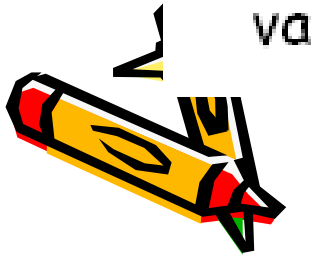


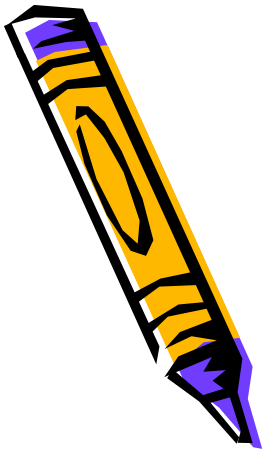
Fiind dată o listă de *n* articole, să considerăm o listă de *m* cumpărături (prin cumpărătură înțelegem coșul de cumpărături al unui client, ilustrat prin bonul de casă).
Pentru a descoperi regulile de asociere, procedăm astfel:





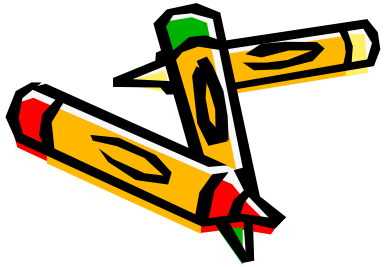
- calculăm numărul de apariții a fiecărui articol;
- construim tabelul de apariție simultană pentru perechile de articole;
- determinăm regulile ce conțin două articole, utilizând valorile de suport, încredere și diferență de nivel;
- construim tabelul de apariție simultană pentru tripletele de articole;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

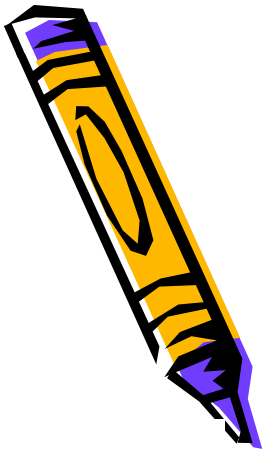




Valoarea m (numărul cumpărăturilor) este în general foarte mare.

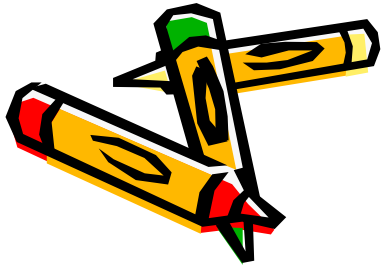
Pentru a construi tabelul de apariție simultană, este necesară parcurgerea acestei liste de mai multe ori, așa că este necesară o arhitectură a acesteia care să permită acces rapid.

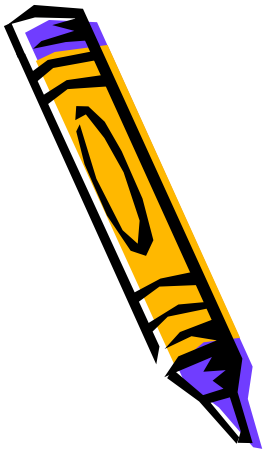




mărimea tabelelor, ca funcție de n și de număr de articole care apar în regulă.

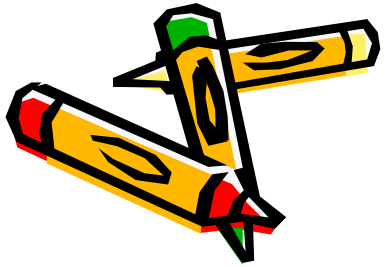
n	C_n^2	C_n^3	C_n^4
100	4950	161700	3921225
10000	$\approx 5 \cdot 10^7$	$\approx 1.7 \cdot 10^{11}$	$\approx 4.2 \cdot 10^{14}$

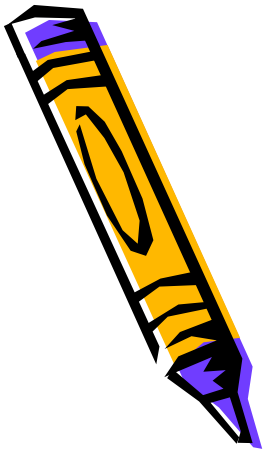




Definim procesul de descoperire a regulilor de asociere:

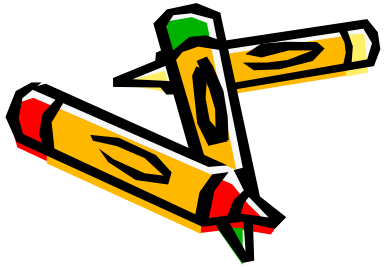
„Fiind dată o mulțime de tranzacții, să se descopere toate regulile posibile pentru care atât suportul cât și încrederea să fie mai mari sau egale decât anumite praguri prestabilite".



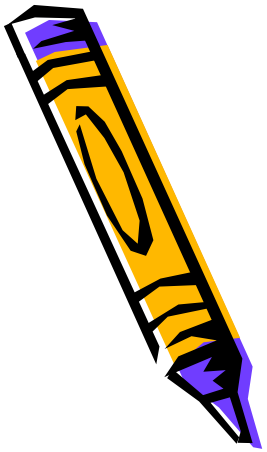


În acest sens, se poate arăta că numărul total *R* al regulilor care pot fi extrase dintr-un set de date, care conține un număr de *n* articole, este dat de formula:

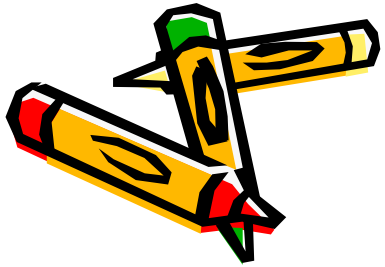
$$R = 3^n - 2^{n+1} + 1,$$



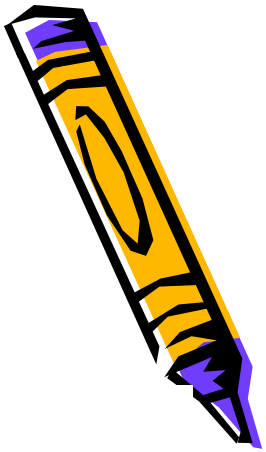
fasonare cu suport minim



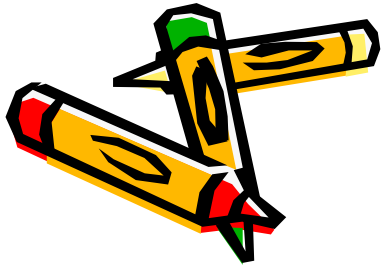
O tehnică de reducere a numărului de articole şi a combinaţiilor lor luate în considerare, poartă numele de *fasonare cu suport minim*.



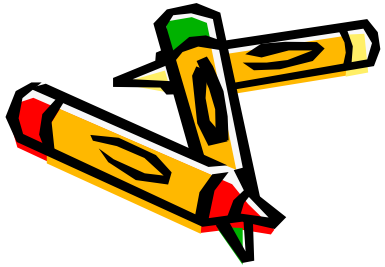
exemplu

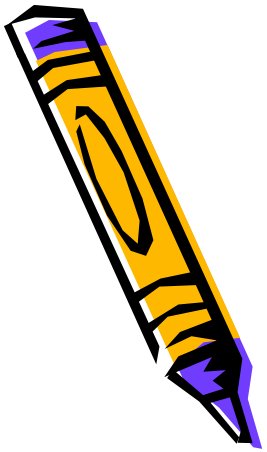


Căutăm reguli cu trei articole:
vom considera doar acele articole al căror suport este mai mare decât o valoare dată *apriori*.
De exemplu, pentru o listă de $m = 1000000$ cumpărături, dacă suportul minim este 2%, în etapa a doua, luăm în considerare doar regulile de forma „dacă X atunci Y", unde X și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

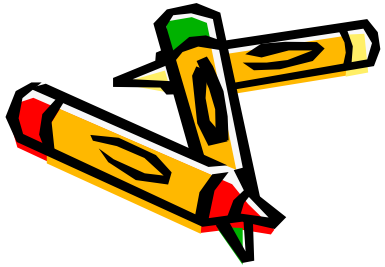


Fasonarea cu suport minim permite eliminarea articolelor ce sunt mai puțin frecvente.
În funcție de etapă, putem varia suportul minim și astfel, diminuându-l, putem găsi combinații rare de articole frecvent cumpărate.
Dacă suportul minim va crește, vom obține combinații frecvente de articole rar cumpărate.

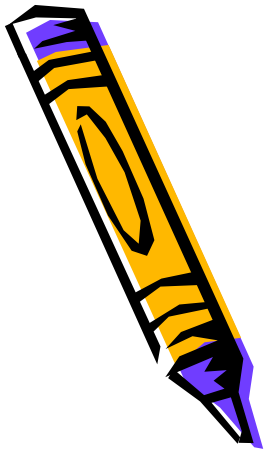




În anumite aplicații se poate limita numărul de combinații prin limitarea formei regulilor căutate.
De exemplu, concluzia regulii este restrânsă la o submulțime a mulțimii articolelor, cum ar fi ultimele modele primite.
Se pot limita calculele prin crearea de grupări de articole, ceea ce necesită sfatul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

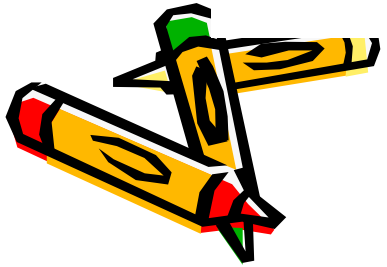


exemplu

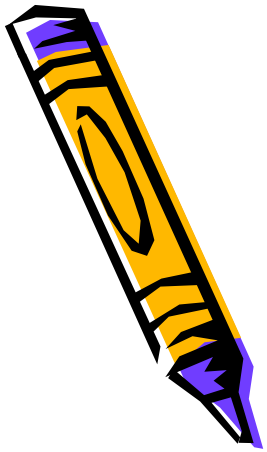


în cazul unui supermarket, un anumit articol poate fi descris astfel:

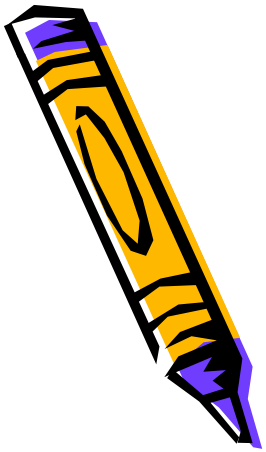
- conservă;
- conservă de legume;
- conservă de legume de la un anumit producător;
- conservă de legume de un anumit gramaj;
- conservă de legume de la un anumit producător, de un anumit gramaj.



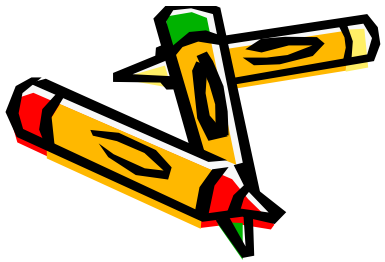
caracteristicile metodei

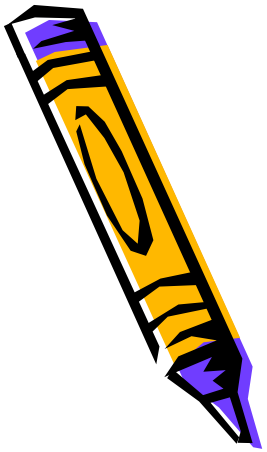


- regulile prezentate sunt uşor de folosit şi de interpretat în situaţii concrete;
- este o metodă de învăţare nesupervizată, pentru extragerea regulilor fiind necesară doar lista articolelor;
- cumpărăturile sunt de mărime variabilă;
- se poate introduce şi variabila timp, în sensul că se pot genera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

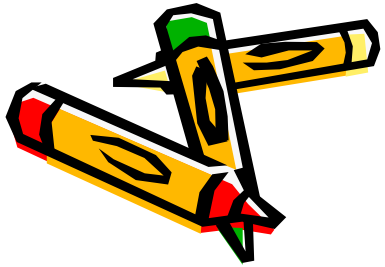


- metoda şi calculele sunt elementare şi astfel se poate programa uşor în cazul unor baze de date de mărime rezonabilă;
- necesită mult timp de lucru şi, prin regruparea articolelor sau metoda suportului minim, se diminuează calculele, dar există riscul pierderii unor reguli importante;

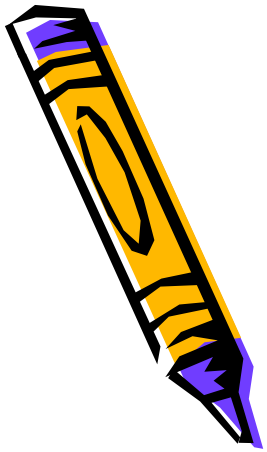




- metoda este mai puțin eficientă pentru articolele rar cumpărate, în aceste caz de obicei se variază suportul minim;
- se pot deduce reguli triviale sau reguli inutile



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.