Adatbányászat

Bevezetés

Tikk Domonkos

Áttekintés

- Mi fán terem az adatbányászat
- Módszertan
- Tipikus feladatok
- Eszközök
- Esettanulmány

Forrás: http://www.adatmentes-adatvissza.hu/hu/adatmentes-adatbanyaszat-data-mining.html

Mottó

Megfulladunk az adatban és tudásra éhezünk

(We are drowning in information but starved for knowledge)

Adatbányászat – motiváció

- Gyorsuló ütemben növő adatmennyiség
- Üzleti igény az adatokban rejlő információk kinyerésére
- Definíció: döntéstámogatási **folyamat**, amely **érvényes**, **hasznos**, **rejtett** (korábban nem ismert) információt állít elő nagy mennyiségű – jellemzően adatbázisokban tárolt – adatból

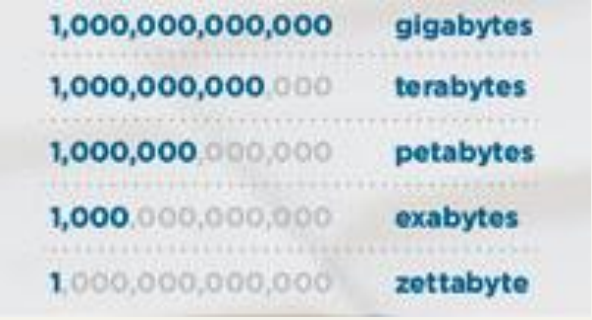
  (forrás: Abonyi (szerk): Adatbányászat a hatékonyság eszköze)
- Automatizálható folyamat
  - emberi erőforrás igénye alacsony
  - gyorsan generálhatóak az információk

A definíció elemei

- folyamat
  - nem dobozos termék, hanem átfogó tudást igényel az alkalmazása is
- érvényes
  - pontosság, statisztikai szignifikancia, teljesség
- hasznos
  - adjon új, értékes ismereteket
  - gyakran nehéz üzleti értéket meghatározni
- rejtett (előzőleg nem ismert)
  - hipotézis megerősítése vs. új felfedezése
  - prediktív vs. leíró adatbányászat

Honnan a sok adat?

- Életünk szinte minden rezdülése adatként le van tárolva valahol
  - személy: vásárlás, internet- és mobilhasználat, egészségi adatok, közlekedés (GPS)
  - vállalat: termelő rendszerek, kutatás, fejlesztés, szervezeti működés
- 2002: éves adattermelés 5 exabyte (800 MB/fő)
- 2012: 2,5 ZB adat csak vállalati területen: 360 GB/fő

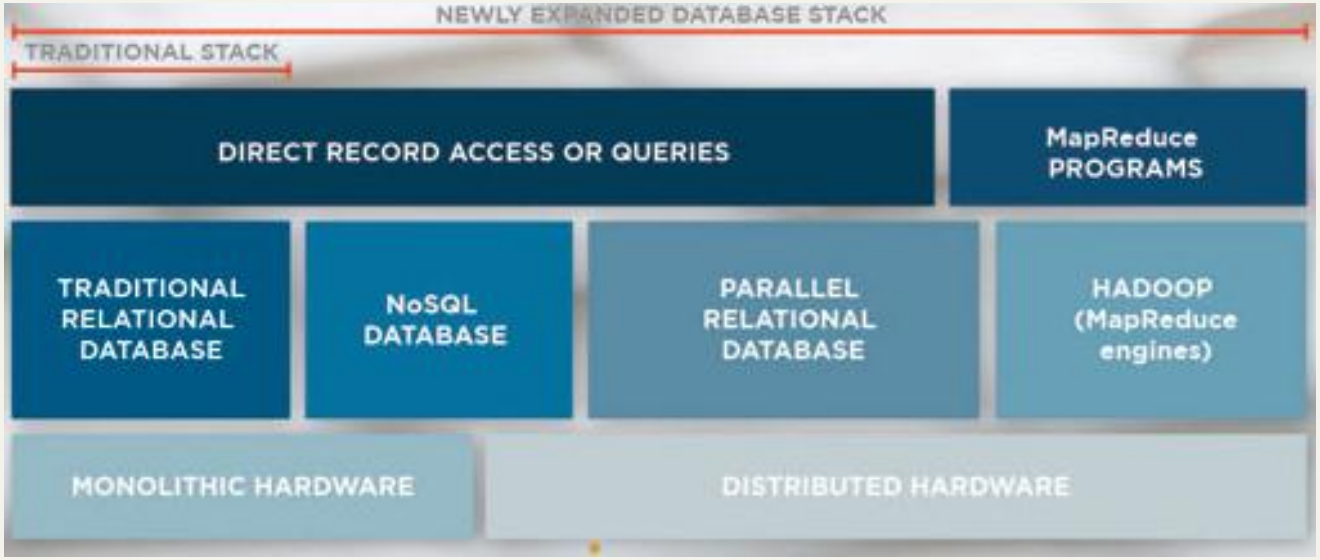


Trend


The production of data is expanding at an astonishing pace. Experts now point to a 4300% increase in annual data generation by 2020. Drivers include the switch from analog to digital technologies and the rapid increase in data generation by individuals and corporations alike.
Size of Total Data
Enterprise Created Data
Enterprise Managed Data
2020: MORE THAN 1/3 OF THE DATA PRODUCED WILL LIVE IN OR PASS THROUGH THE CLOUD.
2012: CUSTOMERS WILL START STORING 1 EB OF INFORMATION.
2009
2010
2015
2020
.79ZB
1.2ZB
.96ZB
.36ZB
2.37ZB
6.32ZB
7.9ZB
10.5ZB
28ZB
35ZB

Adatbázis dimenzió

- Tradicionális relációs adatbázisban már csak az adatok ~20%-át tárolják
- Új technológiák: No-SQL DB, párhuzamos RelDB, elosztott rendszerek (Hadoop)
- Hardver szerint: monolitikus vs. elosztott



Méret és tartalom

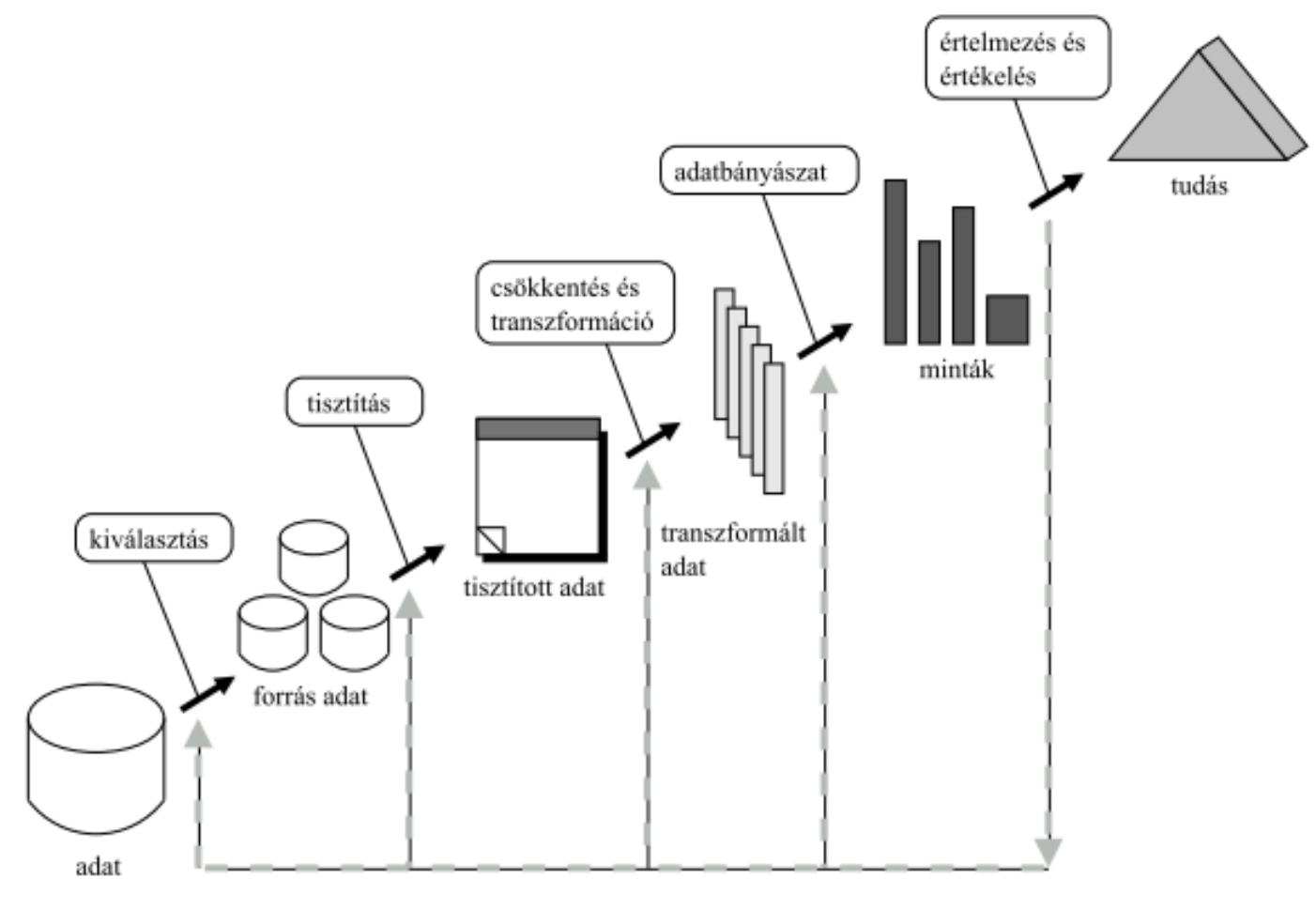
- Az emberi tudás összmennyiségét 300 EB-ra becsülték (2007)
- Avatar számítógépes grafikával készített effektjeinek adatmennyisége: 1 PB
- Strukturált adat:

  relációs DB
- Strukturálatlan adat:

  szöveg, audio, video

Forrás: http://www.csc.com/insights/flxwd/78931-big_data_growth_just_beginning_to_explode

A tudásfeltárás folyamata

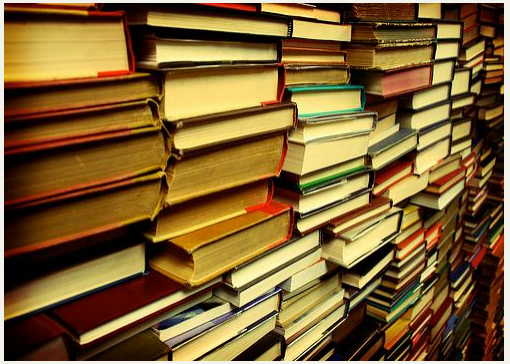


forrás: Bodon F.: Adatbányászat (elektronikus jegyzet)

Adattípusok

- Adatbázis-tartalom
- tranzakciós adatbázisok
- adattárházak
- térinformatikai adatok
- idősor és temporális adatok
- szöveges és multimédia adatok
- www
- heterogén adatbázisok

Company Name	CustomerTy	Order Date	Ship Date	PO Number
A. Datum Corporation	Commercial	3/23/2007	3/29/2007	10
A. Datum Corporation	Commercial	3/23/2007	3/29/2007	11
Contoso, Ltd	Government	10/16/2007	10/19/2007	26
Contoso, Ltd	Government	12/10/2007	12/11/2007	12A948
Trey Research	Government	6/21/2007	6/28/2007	30
Trey Research	Government	11/15/2007	11/18/2007	PO359194
Litware, Inc	Non-Profit	11/12/2007	11/14/2007	32
Litware, Inc	Non-Profit	1/12/2008	1/20/2008	PO1294204
New Customer	Non-Profit	12/20/2007	12/21/2007	PO128A438



Alkalmazási területek

- pénzügyi szektor
- tudomány
- gyártástechnológia
- közlekedés
- jog
- telekommunikáció
- energiaipar

Pénzügyi szektor

- Bankkártya bűncselekmények
- Hitelképesség-elemzés
- Ügyfélszegmentáció
- Ügyfélérték számítás
- Lojalitás vizsgálat
- Keresztértékesítés
- Kampánymenedzsment
- Vásárlói kosár elemzés

Tudomány/egészségügy

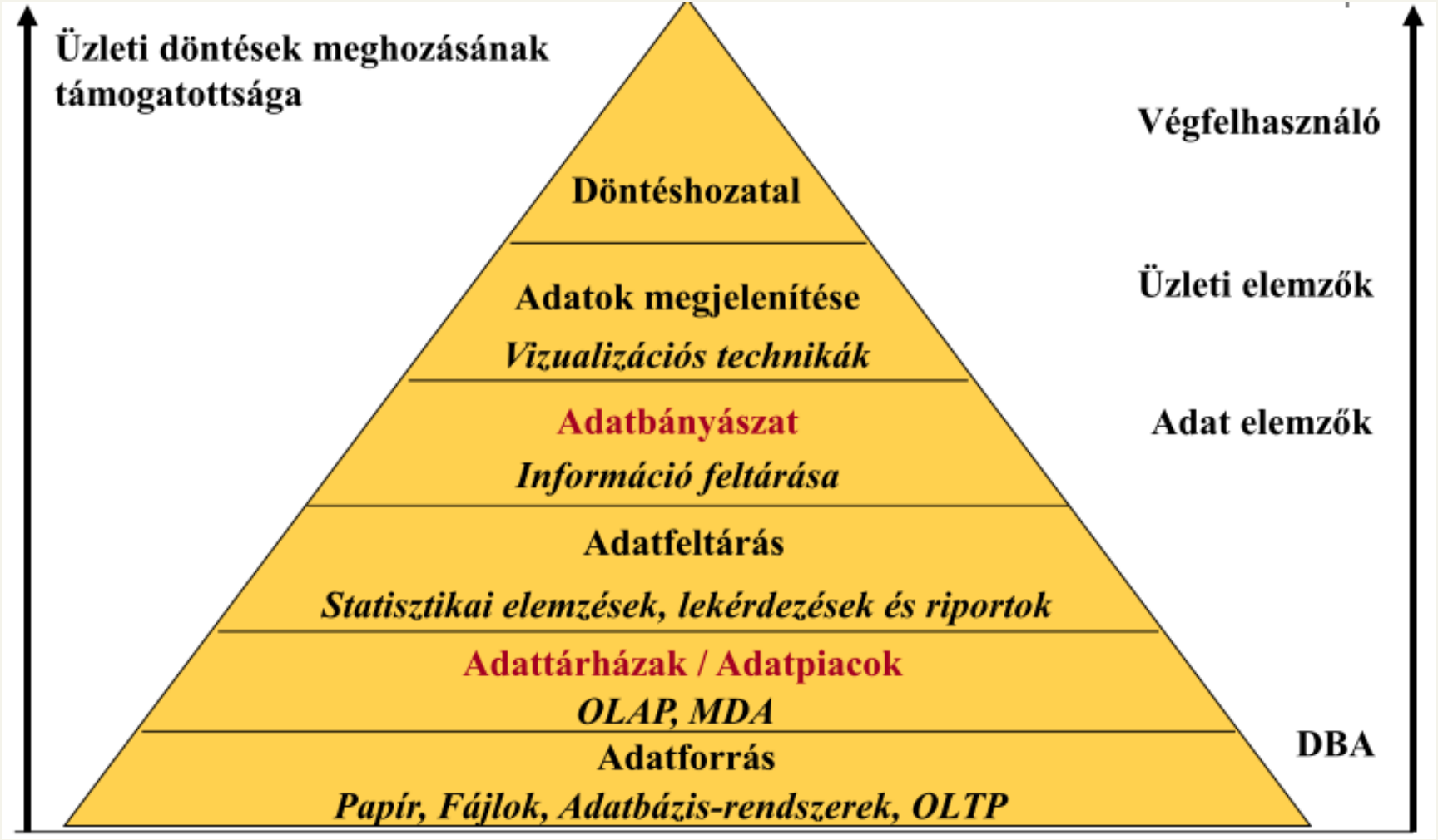
- Kutatási eredmények kiértékelése
- Képek osztályozása
- Új kapcsolatok keresése tényadatokból
- Korreláció elemzés (hipotézis és tényleges mérések között)
- Gyógyszerforgalmi előrejelzések
- Betegségek és fizikai megfigyelések korrelációvizsgálata
- Kórházi monitorozó rendszerek

Telekommunikáció/energiaszektor

- Lemorzsolódás-előrejelzés
- Ügyfél-szegmentáció és termék targetálás
- Véleményalkotók azonosítása – hívási gráf elemzések
- Hálózati hiba előrejelzése
- Túl- és alulszámlázások azonosítása
- Csalás-felderítés

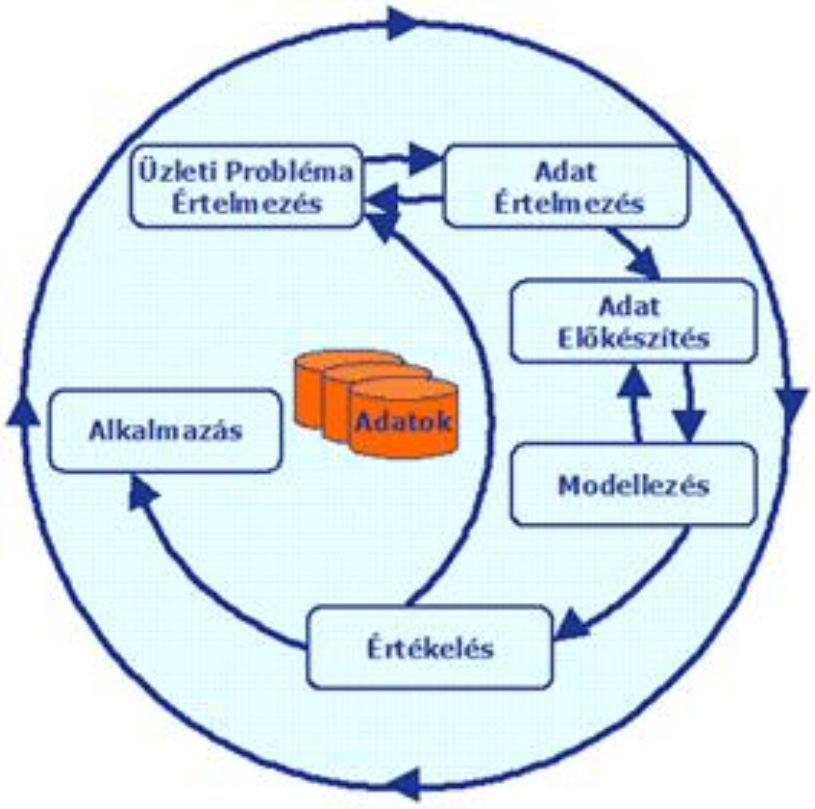
AZ ADATBÁNYÁSZAT MÓDSZERTANA

Az adatbányászat folyamata



CRISP-DM

- Cross-Industry Standard Process for Data Mining



Üzleti probléma értelmezése

- Üzleti célok megfogalmazása
  - üzleti háttér, üzleti cél és siker tényezők
- Helyzetfelmérés
  - erőforrások, követelmények, források, feltételezések
  - kockázatfelmérés, haszon és költségek
  - terminológia
- Adatbányászati célok definiálása
  - adatbányászati célok és sikerkritériumok
- Projektterv elkészítése
  - Eszközök és technikák értékelése

Adatértelmezés

- Kiindulási adatok gyűjtése
  - hozzáférés biztosítása, adatintegráció
- Rátekintés az adatokra
  - főbb jellemzők (típusok, értéktartományok)
- Alap statisztikai jellemzők feltárása
  - lekérdezés, vizualizálás, értelmezés
  - célparaméter eloszlása, főbb dimenziók mentén való szegmentálás
- Adatminőségi vizsgálat
  - feltöltöttség, lefedettség, adathelyesség
- Minderről beszámoló készül

Adatok előkészítése

- Adatkiválasztás
  - a célok eléréséhez mely adatok hasznosak
- Adattisztítás
  - adatkitöltés, inkonzisztencia megszüntetése
- Új paraméterek bevezetése
  - Származtatott adatok, generált rekordok
- Adatintegráció
  - több forrás esetén
- Adatformátum módosítása
  - adatbányászati modellhez igazítás

Modellezés

- Modellező technika kiválasztása
  - eszközt találni a célhoz, adatfeltáró elemzés
- Modell tesztelésének meghatározása
  - Kiértékelési módszer, vizualizálás
- Modellalkotás
  - Paraméterbeállítás, modellek, dokumentálás
- Modell kiértékelése és megjelenítése
  - Fontos a jól vizualizálható eredmény
  - beállítások felülvizsgálata

Üzleti értékelés

- A modell üzleti célú értékelése
  - üzleti elvárásoknak megfelel?
  - éles környezetben tesztelhető
- A teljes elemzési folyamat felülvizsgálata
  - pl. adatok hosszú távú elérhetősége
- Következő lépések
  - döntés a felhasználhatóságról, üzleti bevezetésről

Üzleti alkalmazás

- Alkalmazás megtervezése
  - beépítés az üzleti folyamatokba
- Alkalmazás fenntartás és monitoring
  - tesztesetek, ellenőrzések beállítása
- Projekttanulmány elkészítése
  - Beszámoló, prezentáció
- A projekt felülvizsgálata
  - éles eredmények kiértékelése (ROI!)
  - pozitívumok vs. negatívumok
  - elvárttól való eltérések elemzése

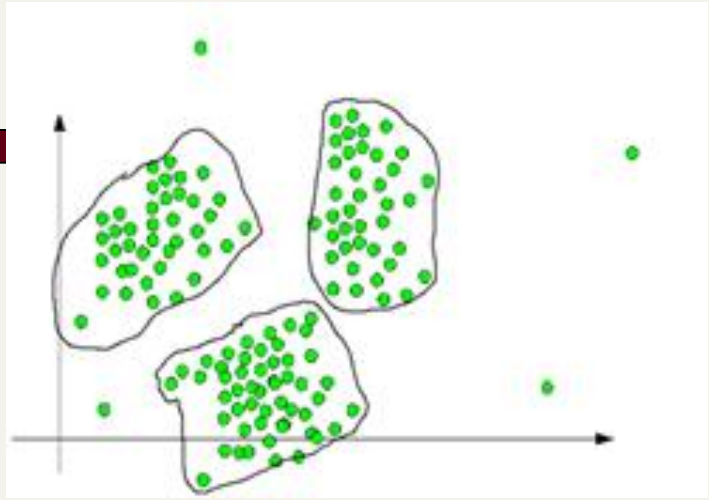
Mai óra

- Mi fán terem az adatbányászat
- Módszertan
- Tipikus feladatok
- Eszközök
- Esettanulmány

Forrás: http://www.adatmentes-adatvissza.hu/hu/adatmentes-adatbanyaszat-data-mining.html

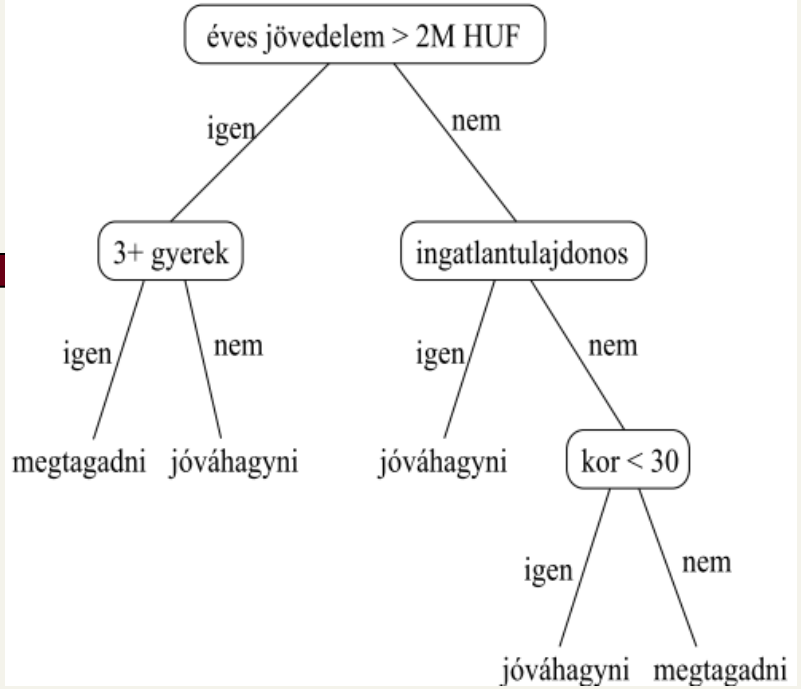
FELADATTÍPUSOK

Csoportosítás



- Szegmentálás, klaszterezés
- *Felügyelet nélküli tanulás*
- Adatok felosztása csoportokra
- Csoporton belül hasonlóak
- Csoportok viszont különbözőek
- Hány csoport legyen?
- Milyen legyen a felosztás struktúrája (egyszintű, hierarchikus)
- Példa: piacszegmentálás

Osztályozás

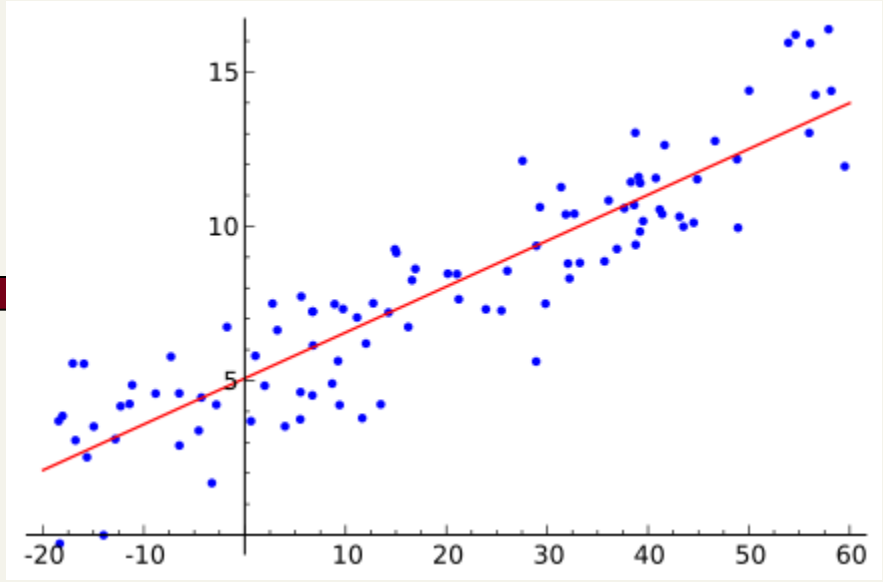


- Klasszifikáció, kategorizálás
- *Felügyelt tanulás*
- Tanuló és tesztadatok
- Modellépítés (-generálás)
- Alkalmazás (előrejelzés)
- Függvény: bemenetekből kimenetet állít elő (osztálycímke)
- Példa:
  - hitelelbírálás
  - égitestek besorolása (galaxis, közeli csillag, egyéb)

Asszociáció

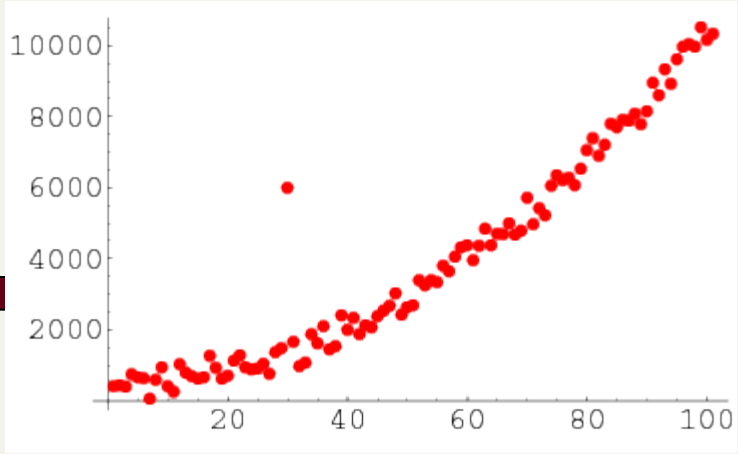
- Gyakori elemhalmazok
- Objektumok közti összefüggés
- Kapcsolat erőssége
- Példa: vásárlóikosár-elemzés
- Ha valaki vesz A és B terméket akkor C-t is vesz
- Konfidencia, támogatottság
- Más feladat: gyakori sorozatok (adatszekvenciák), gyakori epizódok (részben rendezett)

Regresszió



- Függvény illesztés
- osztályozáshoz hasonló
- az adatban rejlő sajátosságok modellezése,
- kimenet numerikus értéke, nem kategorikus adat (osztályozás)
- Adatbányász: modellkiválasztás (lineáris, polinom, logaritmikus, hiperfelület)
- Példa:
  - időbeni előrejelzés: BUX index alakulása
  - statikus: betegség valószínűségi orvosi adatok alapján

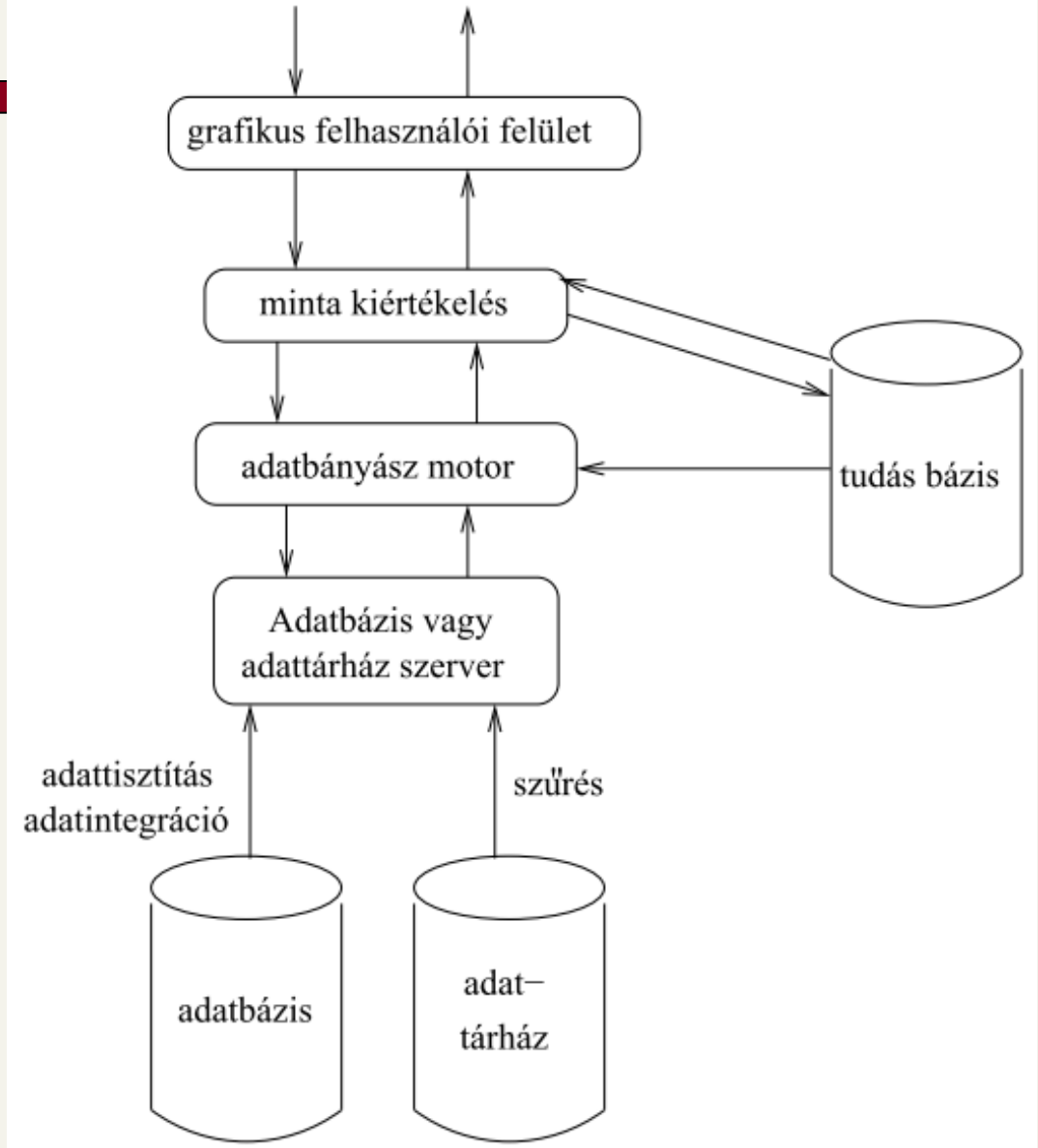
Eltéréselemzés



- outlier analízis
- olyan adatok azonosítása, amelyek eltérnek az elvárttól
- Lehet zaj, mérési hiba, kivétel (ekkor szűrni lehet)
- Alkalmas csalások kiderítésére
- Példa:
  - hitelkártya-visszaélések
  - áramlopás
  - biztonsági elemzés

ESZKÖZTÁRAK

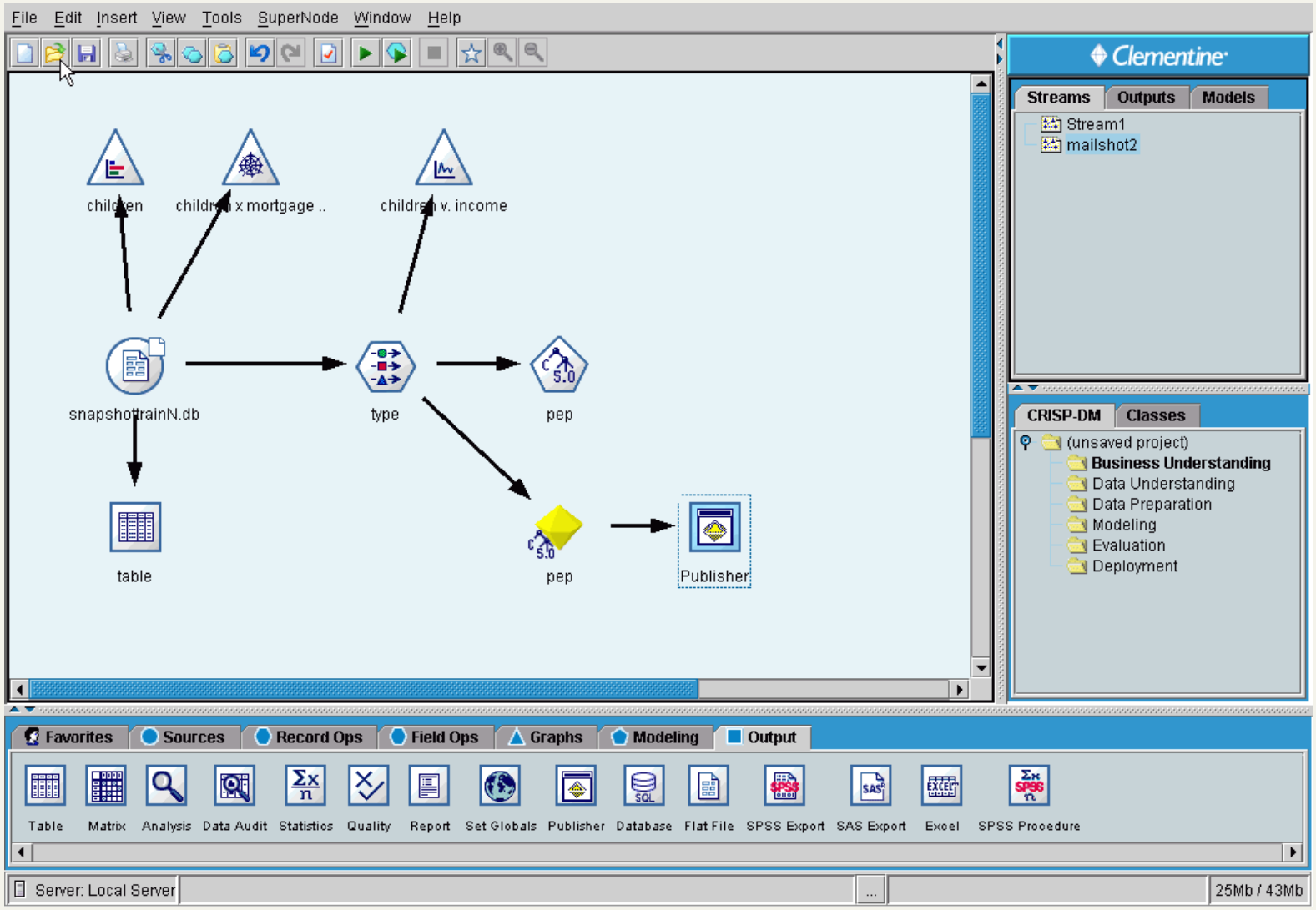
Alapelemek


grafikus felhasználói felület
minta kiértékelés
adatbányász motor
tudás bázis
Adatbázis vagy
adattárház szerver
adattisztítás
adatintegráció
szűrés
adatbázis
adat–
tárház

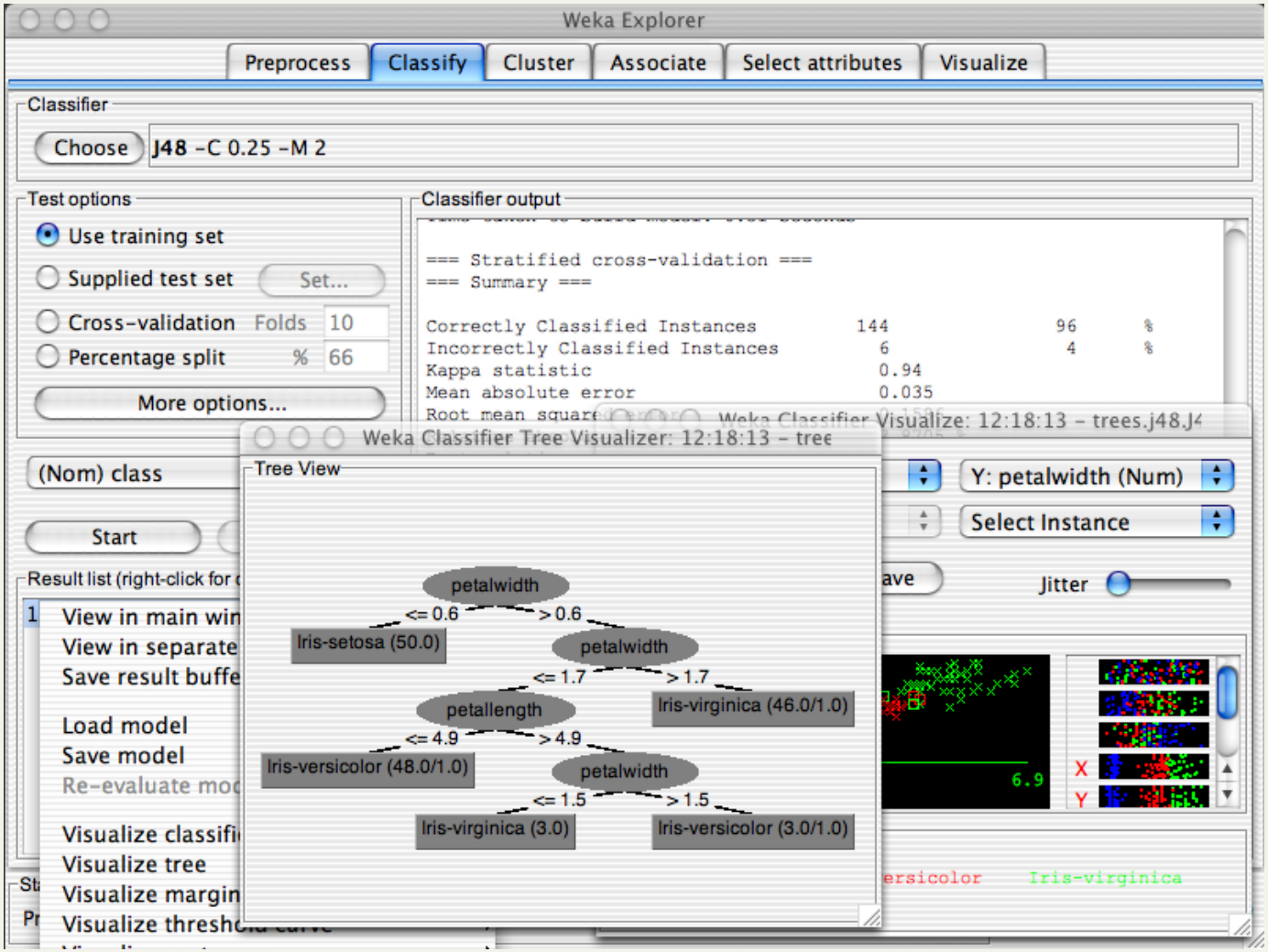
Eszközök

- általános adatkezelők
  - Excel
- programozási keretrendszer:
  - Matlab
- piaci szoftverek
  - IBM/SPSS Clementine, Statsoft Statistica, SAS Data Miner
- adatbázis-kezelők adatbányász kiegészítései
  - Oracle, MySQL, IBM
- Ingyenes rendszerek
  - WEKA, Rapidminer

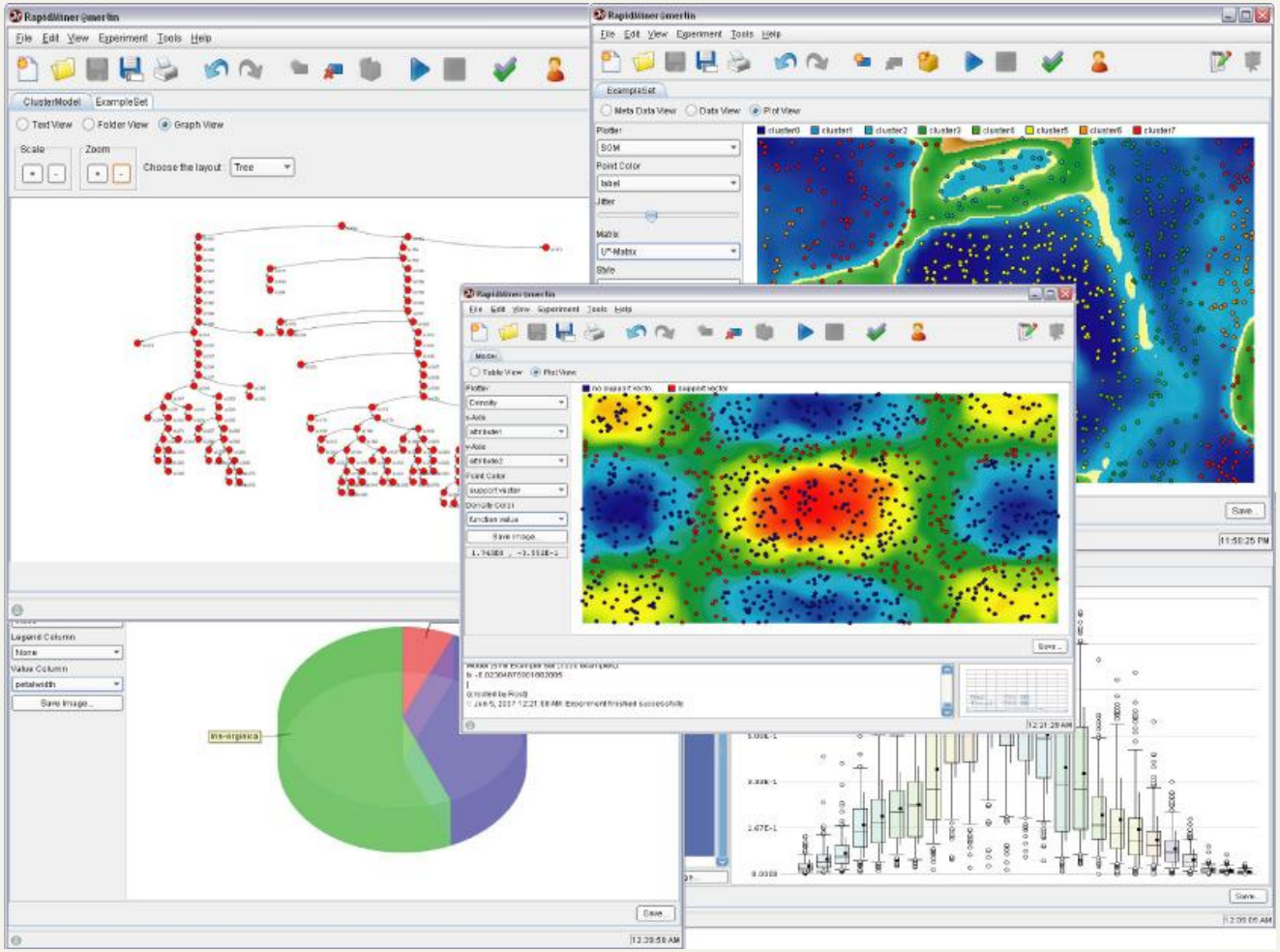
IBM/SPSS Clementine



WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 144 96 %
Incorrectly Classified Instances 6 4 %
Kappa statistic 0.94
Mean absolute error 0.035
Weka Classifier Tree Visualizer: 12:18:13 - tree
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
petallength
Iris-virginica (46.0/1.0)
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
petalwidth
<= 1.5
> 1.5
Iris-virginica (3.0)
Iris-versicolor (3.0/1.0)
(Nom) class
Y: petalwidth (Num)
Start
Select Instance
Result list (right-click for
Jitter
View in main win
View in separate
Save result buffe
Load model
Save model
Visualize classifi
Visualize tree
Visualize margin
6.9
X
Y
Iris-virginica

Rapidminer


File Edit View Experiment Tools Help
ClusterModel
ExampleSet
Text View
Folder View
Graph View
Scale
Zoom
Choose the layout
Tree
Meta Data View
Data View
Plot View
Plotter
SOM
Point Color
Label
Jitter
Matrix
U*-Matrix
Style
Legend Column
None
Value Column
petalwidth
Save Image...
Save
ins-erganica

ESETTANULMÁNY

Tüdőembólia detektálása

- Adatok:
  - numerikus gyanús régiókról (3D pixel – 116 jellemző); első 3: (x,y,z) az adatok szemantikája nem ismert; [-1; 1] be normált
  - beteg azonosító (egy beteghez több mérés)
  - pozitív és negatív minták címkével (beteg/nem)
- Feladatok:
  - új minták osztályozása
  - beteg emberek azonosítása
  - egészséges emberek azonosítása 100%-kkal

Feladat felépítése

- tanítóadatok:
  - 3303 adatsor; 46 beteg és 20 egészséges eset
- tesztadatok:
  - 1391 adatsor, 33 eset
- Felügyelt tanulás (osztályozás)
  - olyan modell építése, amely egy adott mintáról el tudja dönteni, hogy beteg-e vagy sem

Kényszerfeltételek

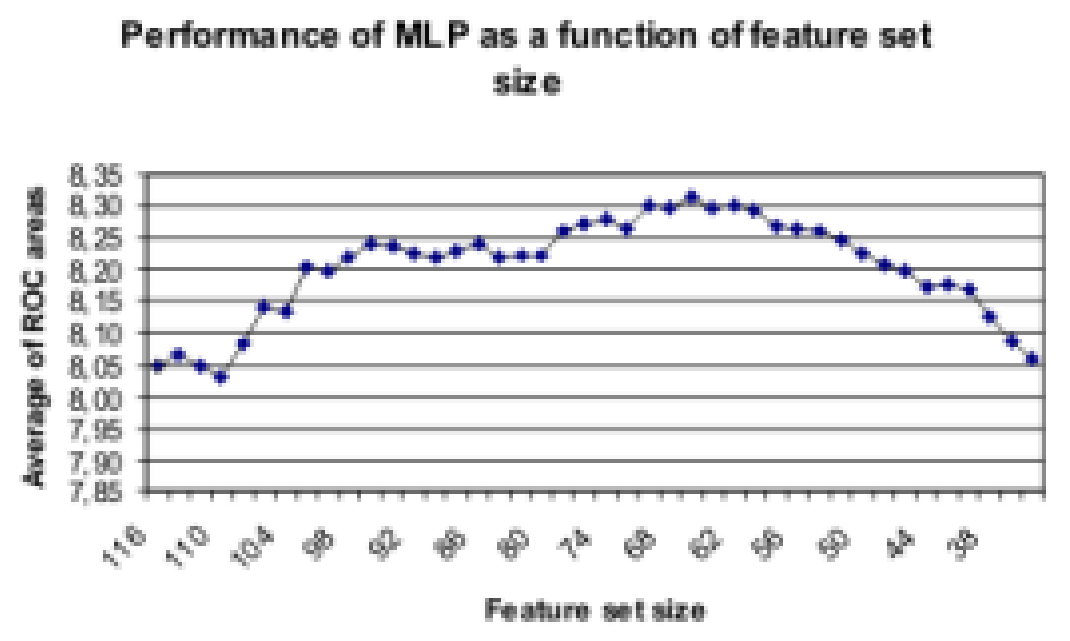
- hibás pozitív (FP) esetek minimalizálása
  - „farkast kiáltani” faktor csökkentése
- 3 küszöbérték adott, hogy páciensenként mennyi lehet a hibás esetek aránya (FP rates: 2; 4; 10)
- mérések:
  - helyes pozitív (TP) adatok azonosítása adott FP kényszerfeltételek mellett (#TP-PE)
  - hány beteg pácienst ismer fel adott FP kényszerfeltételek mellett a rendszer (#TP-P)
  - helyesen azonosított egészséges páciensek száma

Nehézségek

- Zajos és kevés adat
  - hogyan generálták az adatokat
  - különböző gépek, szakértői címkézés
- adatsorok szemantikája ismeretlen
- térbeli összefüggések az adatok között nem azonosíthatóak
- atipikus a célfüggvény és a kényszerfeltétel

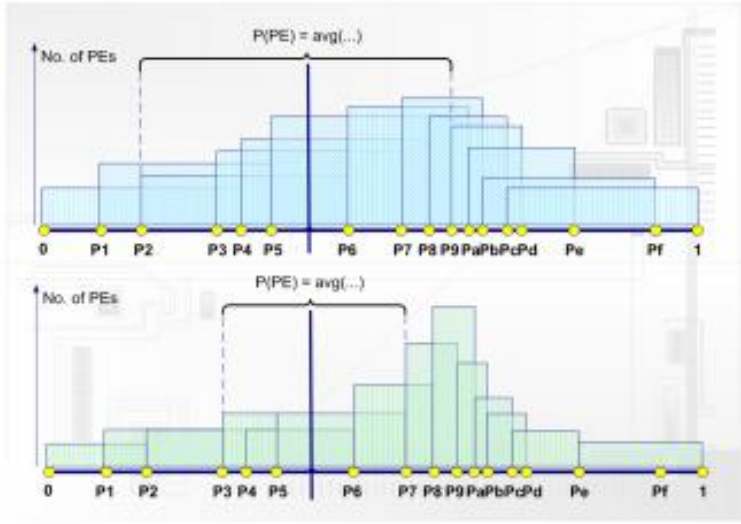
Adattisztítás

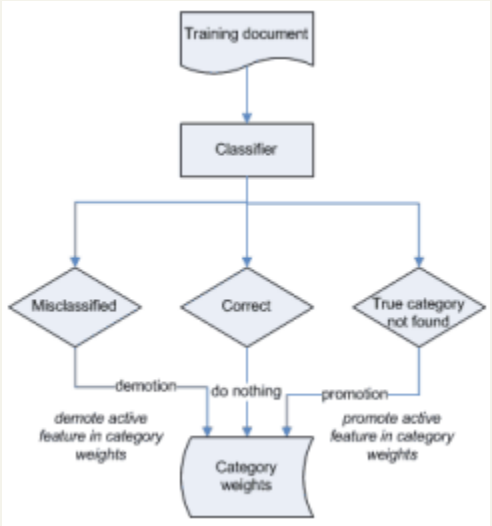
- A 116 adatelemből melyik használható
  - pontosít
  - gyorsít



Adatbányászati módszerek

- Osztályozók
  - statisztikai alapú
  - neurális háló A
  - neurális háló B
- Kombináció
  - osztályozó bizottság
  - feladat specifikusan
  - konfidencia értékek
    - osztályozó szinten
    - predikció szinten





Kombináció

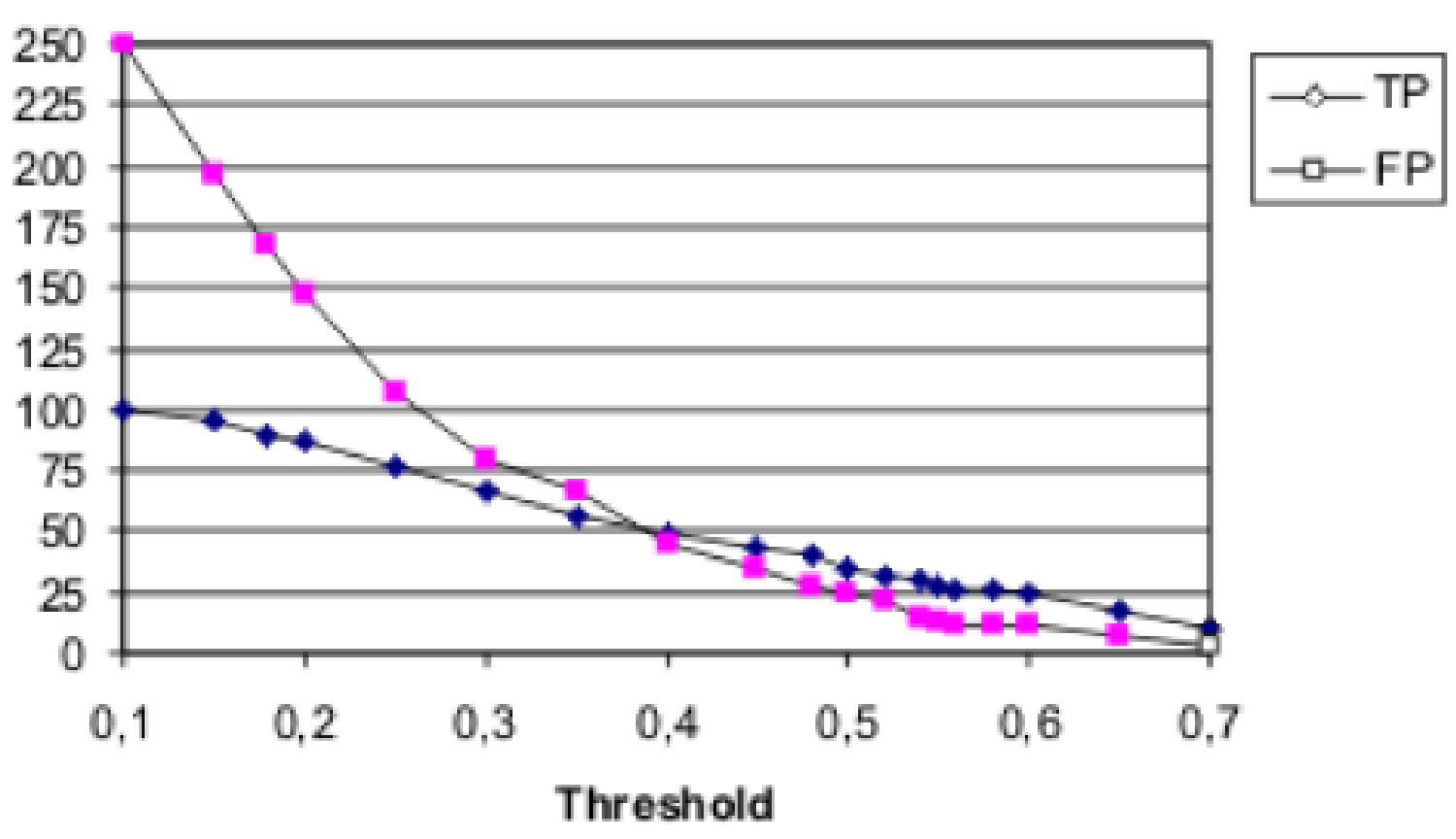
- Parametrizált vétó stratégia
- Biztossgái tagok:
  - 3 módszer, (1-1-2) beállítással: 4 adatsor
- Osztályozók kimenetének súlyozási szabályai
  - Egyöntetű pozitív döntés esetén: pozitív
  - 2-3 pozitív és nincs vétó: pozitív
  - 1 pozitív és nincs gyenge vétó: pozitív
  - különben negatív

Mitől függ a vétó értéke

- Megengedett hibás minták aránya (FP rates: 2; 4; 10)
- Osztályozó pontossága, amit keresztvalidációval mértünk
- esetenként az osztályozó által adott konfidencia-érték (nem mindenütt adott)
- osztályozók belső küszöbértéke

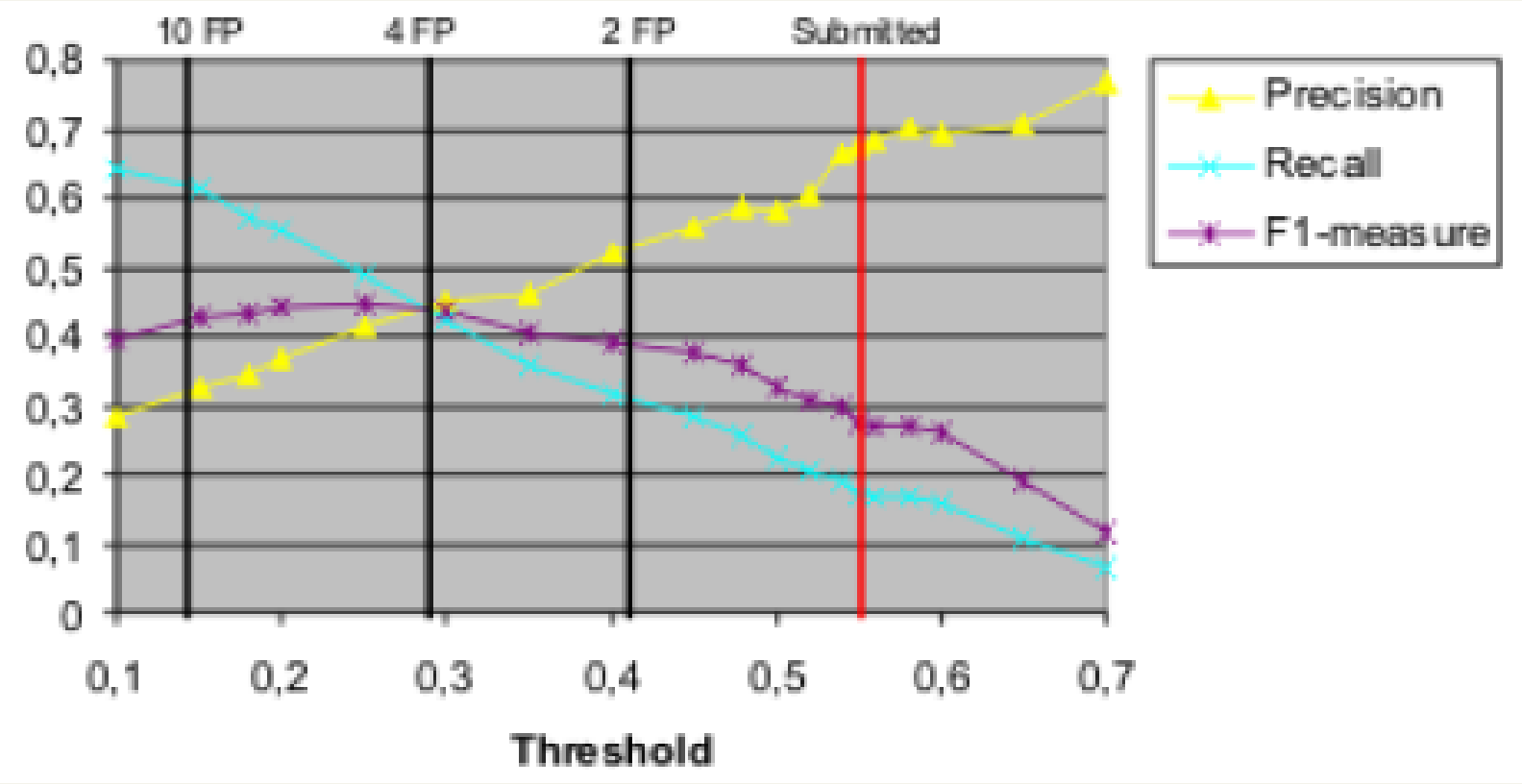
konzervatív vs. bátor

- a helyes és hibás találatok függvényében



Hagyományos IR mértékekkel

- pontosság (precision): a találatok közül mennyi helyes (PE)
- fedés (recall): hányat talál meg a tényleges PE-k közül



Eredmények

- Egyes osztályozókra

	Approaches			
	Statistical	ANN	Hitec-62	Hitec-116
Number of true PEs (23)	43	42	32	77
Number of false PEs (23)	62	12	18	91
Uniquely found PEs (23)	6	7	3	27

Kombináció

Combinations						Task1			Task2		
Stat & ANN	Stat & Hitec-62	Stat & Hitec-116	ANN & Hitec-62	ANN & Hitec-116	Hitec-62 & Hitec-116	Voting for 2FP	Voting for 4FP	Voting for 10FP	Voting for 2FP	Voting for 4FP	Voting for 10FP
22	22	36	24	33	27	51	80	90	49	64	98
8	5	19	3	5	8	21	57	109	18	58	144

Végeredmény

Method/submission	Precision (%)	Recall (%)	FP rate
Statistical	40.95	27.56	2.70
ANN	77.77	26.92	0.52
HITEC-116	45.83	49.35	3.96
HITEC-62	64.00	20.51	0.78
Task 1a	70.83	32.69	0.91
Task 1b	58.39	51.28	2.48
Task 1c	43.06	57.69	4.74
Task 2a	73.13	31.41	0.78
Task 2b	52.46	41.02	2.52
Task 2c	40.50	62.82	6.26

- pontosság: a találatok közül mennyi helyes (PE)
- fedés: hányat talál meg a tényleges PE-k közül