Vezetői információs rendszerek

Adatbányászás I.

Adatbányászás

Növekvő adatmennyiség → egyre nehezebbé válik az adatokból a kívánt információ kinyerése.
Új technika szükséges, amely lehetővé teszi, hogy megismerhetővé, kinyerhetővé váljon a nagy adathalmazokban rejlő tudás. (adat ≠ tudás)
***Az új technika: Adatbányászás* (Data Mining)** →
az 1990-es években jelent meg az üzleti köztudatban.
Az adatbányászás olyan *adatelemzési folyamat,* amely nagy adatbázisokból *érvényes, hasznos, rejtett, előzőleg nem ismert* információkat tár fel.

Folyamat: nem automatikusan generálja a korábban nem látott adatokból a felhasználhat tudást, hanem egy rendkívül összetett folyamat révén.

Érvényes: az adatokból kinyert információnak pontosnak, teljesnek kell lennie.

Pl.: ha valamilyen szempont szerint kell kiválogatni egy adatbázisból a megfelelő ügyfeleket, fontos hogy minden, a kritériumnak megfelelő ügyfelet azonosítsunk.

Hasznos: nem elegendő a pontos ismeretek feltárása, a feltárt tudásnak az adott elemzés szempontjából hasznosnak kell lennie. (Hasznosság mérése nem mindig mérhető.)

Előzőleg nem ismert: az adatbányászás felfedező jellegű adatelemzés.
Adatelemzés célja: megerősítés és felfedezés.
Megerősítés: egy adott hipotézis megerősítése.
Felfedezés (tudásfeltárás): a rendszer autonóm módon generál hasznos mintákat az adatokból.

Adatok információvá történő átalakítása:

1. **Hagyományos mód** az adatok **manuális elemzése**, **interpretálása**.

   Az adott tématerület szakértői időszakosan elemzik az elmúlt időszakban összegyűlt adatok által mutatott trendeket.

   A szakértők jelentésekben foglalják össze tapasztalataikat → ezek a jövőbeli döntéshozatalban hasznosulnak.

   Ezekben az esetekben az adat alapú tudásfeltárás folyamata **lassú, drága** és **szubjektív.**

Adatok információvá történő átalakítása:

2. **Adatbányászási módszerekkel** :
   - így szakértői döntések **automatizálhatók**,
   - az automatizálás lehetővé teszi **emberi erőforrás megtakarítását,**
   - továbbá lehetővé válik a kérdésekre adott **válaszok reakció idejének lerövidítése.**

Adatok megőrzése: nem kétséges a folyamatosan gyűlő adatok megőrzésére szükség van.

Az egyre duzzadó adathalmaz értékes információkat rejthet.

Mekkora adathalmaz keletkezhet évente?

2002-es felmérés (Berkeley Egyetem) szerint abban az évben öt exabájt adat keletkezett: 92 %-a merevlemezen, 7 %-a filmen, 0,01 %-a papíron, 0,002 %-a optikai eszközön (DVD,CD) került tárolásra.

Mennyi az az öt exabájt?

Kb. annyi, mintha a föld minden lakosa (6.3 billió ember) évente egy CD-nyi (800 Mbájt) adatot termelne.

Kritikus jelenség: az adatok egyre nagyobb mennyiségben keletkeznek → a keletkezés sebessége az elmúlt 3 évben megháromszorozódott.

Adat ≠ tudás, információ ≠ tudás

Adatbányászás és statisztika:
Hasonlóság: mindkettő adatelemzési módszer.
Különbözőség:

1. A statisztika mintákból következtet az eredeti sokaság tulajdonságaira, az adatbányászási módszerek pedig a rendelkezésre álló adatbázist elemzik.
2. A statisztika az adatbázison kívüli információra támaszkodva állítja fel a modelljét, hipotézisét és teszteli azt. Az adatbányászás közvetlenül az adatbázisból képes előzőleg nem ismert összefüggések kinyerésére.

A tudásfeltárás folyamata és az adatbányászás

Adatbázisokban végzett tudásfeltárás :
(Knowledge Discovery in Databases, KDD)
1995-ben Montreál → itt tartották az első tudásfeltárásról szóló konferenciát .
Döntés:
1. KDD azt a *teljes folyamatot* jelentse, amelynek során az adatokból kinyerjük az információt.
2. Az adatbányászás a tudásfeltárás folyamatának az a *lépése*, amelyben az adatokban lévő összefüggések felfedezése történik.

A KDD folyamatának lépései

1. Adatkiválasztás: az elemzéshez szükséges adatok kiolvasása az adatbázisból.
2. Adatelőkészítés: adattisztítás, adatbővítés, adattranszformáció (kódolás) - a nem megfelelő adatok eltávolítása, esetleg a vizsgálatokhoz hiányzó adatok integrálása, az adatoknak a vizsgálatok számára használható alakra hozása.
3. Adatbányászás: olyan eljárás, amely során adatbányászási technikák (klaszterezés , osztályozás, stb.) alkalmazásával feltárjuk az ismeretlen, új trendeket, összefüggéseket, ill. mintázatokat.

A KDD folyamatának lépései

4. Jelentéskészítés: az előző lépés eredményeinek a végfelhasználó számára értelmezhető formában történő megadása.

A fenti folyamat egyes lépései általában különböző feladatköröket ellátó emberek **együttes munkája**.

Pl. az adatelőkészítési lépésben együtt kell dolgozniuk az adatgyűjtőknek az adatbányászokkal, a jelentés készítés pedig az adatbányászoknak a szakértőkkel való közös tevékenységét igényli.

Különböző feladatköröket ellátó szakemberek együttes munkája:

A KDD folyamatának jellemzői

A tudásfeltárás általában *nem lineáris* folyamat →

bármely lépésben előfordulhat hogy az előző lépésben kapott információ nem megfelelő →

ekkor vissza kell térni a megelőző lépés(ek)hez, és módosított adatokkal, paraméterekkel, esetleg más eszközzel kell folytatni a munkát, megismételve egyes lépés(eke)t →

a ***tudásfeltárás*** általában egy ***iteratív*** folyamat.

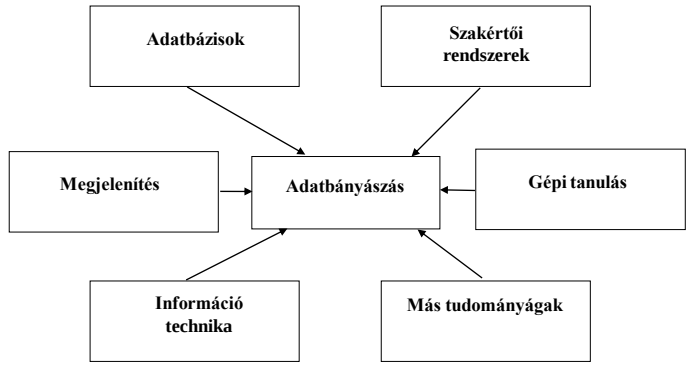
A KDD folyamatának jelemzői

A tudásfeltárás folyamatában az *új tudást* az *adatbányászási lépés* adja, ennek ellenére minden lépés egyforma fontossággal bír.
Megfelelő adatok nélkül nem lehetséges ugyanis az adatbázisban meglévő ismeretlen összefüggések megismerése.
A folyamat időigényét tekintve sem maga a 3. lépés tart a legtovább, a szükséges idő 80%-át az adatelőkészítés és a kapott eredmények értékelhető formába öntése teszi ki.

A KDD és az adatbányászat

Az adatbányászás – a fentiek szerint – egy lépés a tudásfeltárás folyamatában.
Ennek ellenére az iparban, a felhasználók körében az *adatbányászás* elnevezést a *KDD szinonimájaként* is használják.
Az adatbányászás terminológia a teljes folyamatra sokkal elterjedtebb, mint a jóval hosszabb „tudásfeltárás adatbázisokban” elnevezés.
Ezért szerepelt már a definícióban is az *adatbányászás* egy *adatelemző folyamatként*.

Adatbányászáshoz kapcsolódó tudományágak



Az adatbányászás alkalmazási lehetőségei

Az adatbányászási feladatok osztályozása (adatelemzési szempontból):

1. leíró adatbányászási feladat: az adatbázisban tárolt adatok alap (általános) jellemzőit határozza meg

2. előrejelző (következtetétes) ***adatbányászási feladat:*** a meglévő adatokból az alapvető összefüggések, mintázatok (adatok egymáshoz viszonyított elhelyezkedése) feltárásával prognosztizál.

Leggyakrabban használt adatbányászási technikák

1. A **társításelemzés** *társítási szabályok* (asszociációs szabályok) *feltárását* jelenti →

Azt vizsgájuk, hogy az adatbázis elemei között létezik-e összefüggés. Ha létezik, akkor ez adatbányászási eszközökkel feltárható és a kapcsolat erőssége is jellemezhető.

Ez az eljárás széles körben a kereskedelemben, a bevásárlókosár típusú elemzéseknél használható.

PL.: Egy adatbányászási elemzés során a következő társítási szabályt tárták fel:

életkor(XY, 30-40 év) és éves jövedelem (XY, 5-8 millió Ft) $\Rightarrow$ *autót vesz(XY, Audi), (gyakoriság= 5%, bizonyosság=30%),* ahol XY a vásárló azonosítója.

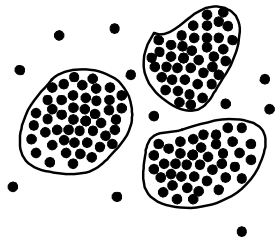
Ez a szabály azt fejezi ki, hogy a vizsgált vásárlóknak (az adatbázisban levő rekordoknak) az 5%-ára érvényes az életkorra és a jövedelemre vonatkozó feltétel és 30 % a valószínűsége, hogy az életkorra és a jövedelemre megfogalmazott feltételeknek eleget tevő vásárlók Audi gépkocsit vásárolnak.

Leggyakrabban használt adatbányászási technikák

2. Csoportosítás (klaszterezés) segítségével az adatoknak csoportokba (klaszterekbe) sorolása történik úgy, hogy az egyes csoportokba egymáshoz hasonló elemek kerüljenek, az egyes csoportok viszont jelentősen különbözzenek egymástól.

Az egy csoporthoz tartozó elemek esetén maximalizáljuk, a különböző csoportokhoz tartozó elemek esetében viszont minimalizáljuk a hasonlóságot.

A feltárt csoportok lehetnek egymást kizáróak, de akár egymással átfedők is.



Az adatok három csoportba (klaszterbe) sorolásának kétdimenziós szemléltetése

A klaszterezés tipikus példája: a piackutatás.

Vásárlási szokások alapján lehetőleg homogén vásárlói csoportokat határoznak meg.

Ezek a csoportok kereskedelmi célcsoportok → jól használhatók a marketing tevékenység optimalizálása érdekében → bizonyos reklám anyagot csak a megfelelő célcsoporthoz juttatnak el.

Leggyakrabban használt adatbányászási technikák

3. Az osztályozás jellemzője, hogy egy adott (kiválasztott) *ismérv* alapján akarjuk az adatbázis elemeit (rekordjait) megkülönböztetni, osztályokba sorolni.

Ezt a kiválasztott ismérvet szokás *osztálycímkének* nevezni.

Osztálycímke csak olyan ismérv lehet, ami véges számú különböző értéket vehet fel → ez azt jelenti, hogy ismert hány osztály létezik.

Az osztályba sorolás osztályozási szabályok alapján történik → különféle módszerek segítségével.

Cél minden esetben olyan szabály felállítása, amelynek segítségével a legpontosabban lehet szeparálni az adatokat a megfelelő osztályba.

Az osztályozást gyakran alkalmazzák pl. pénzintézeti vizsgálatoknál (ügyfelek hitelképességének megadása, biztosítási kockázatbecslés), orvosi alkalmazásoknál.

Például egy pénzintézet ügyfeleit hitelképessége szerint szeretné osztályozni jó, közepes és gyenge minősítésű osztályokba. Az osztályba sorolás alapján jellemezni lehet az egyes csoportokba tartozó ügyfeleket, és ezek alapján a pénzintézet egy új ügyfél hitelkérelméből azt is el tudja dönteni (előrejelzés), hogy hitel visszafizetés szempontjából jó ügyfél lesz-e.

A csoportosítás és az osztályozás hasonló technika.

Lényeges *eltérés* van a két módszer között:

Az osztályozásnál az osztálycímke megadásával ismertek az osztályok, annak számossága is, míg a csoportosításnál nem ismertek előre a csoportok, az adatok alapján kell létrehozni az adatokra jellemző csoportokat.

Ezek alapján a klaszterezés olyan osztályozásként is felfogható, ahol nem ismert az osztálycímke.

Ezek alapján az osztályozás a gépi tanulás területén a felügyelt tanulás (példák alapján történő tanulás) egy formája, míg a csoportképzés felügyelet nélküli tanulási forma.

Leggyakrabban használt adatbányászási technikák

4. A fejlődésanalízis az időben változó adatok időben változó viselkedési szabályosságait modellezi.

A **regresszió-vizsgálat** célja egy előrejelzésre alkalmas függvény megadása, amelynek segítségével ismert értékekből más numerikus érték(ek)re lehet következtetni. Pl. jó példa erre, amikor értékpapír befektetési döntésekhez az értékpapírárak alakulásának előrejelzéséhez az értékpapírt kibocsátó társaságok gazdasági fejlődésének jövőbeli szabályszerűségeit tárják fel adatbányászási módszerekkel.

Az **idősorok elemzése** akkor kerül be az adatbányászási feladatok közé, amikor a hagyományos statisztikai idősor elemzési eszközök már nem alkalmazhatók a feladat bonyolultsága (túl sok változó) miatt.

Az előállított minták használhatósága

Ezekkel a technikákkal elvben nagyon sok szabály, mintázat (minta) előállítható → Az előállítható minták jó része *nem „érdekes”* a felhasználók szempontjából.

Azt szokás mondani, hogy egy **minta érdekes**:

- ha egyszerűen érthető,
- adott megbízhatósággal érvényes új, vagy kísérleti adatokon,
- hasznos és újszerű.

A kapott eredmények érdekességi megítélésére **objektív** és **szubjektív érdekességi mértékek** adhatók meg.

Objektív érdekességi mérték lehet pl. társítási szabályok megadása esetén a kapott szabály megalapozottsága, gyakorisága, azaz az adatbázisban levő adatok (rekordok) százalékos aránya.

A **szubjektív érdekességi mérték** általában valamilyen felhasználói meggyőződést jelent:

- **nem az elvárásnak megfelelő** minták (felhasználói meggyőződésnek ellentmondó)
- **elvárásnak megfelelő** minták.

Ez utóbbiak akkor érdekesek, ha megerősítik a felhasználó feltételezését.

Az objektív és szubjektív mértékeket *mindig társítani célszerű* → pl. nem ugyanazok az eredmények lesznek érdekesek egy cég kereskedelmi igazgatójának és az alkalmazottak teljesítményét vizsgáló elemzőnek.

Az érdekességi mértékek figyelembevétele:

1. Ha az adatbányászás után történik → akkor a feltárt mintákat érdekességük szerint rangsorolják, elhagyva az érdekteleneket.
2. Sokkal előnyösebb azonban, ha a mértékek magát az adatbányászási folyamatot irányítják, korlátozzák: ekkor eleve kizárják az érdekességi mértéknek eleget nem tevő minták keresését.

Adatbányászási módszertanok

1990 évek elején kezdődik a speciális adatbányászási szoftverek kifejlesztése.

Első üzleti célú adatbányászási program: **Clementine** (Integral Solutions Ltd. terméke)

Igény az adatbányászási feladatok végrehajtásához *módszertan kifejlesztésére*.

1996-ban készült el az iparágaktól független adatbányászási módszertani szabvány a : **CRISP_DM (Cross Industry Standard Process for Data Mining)**

A módszertan lényege:
az adatbányászási folyamat mindegyik szakaszában ellenőrzési listákkal, útmutatásokkal, célok és feladatok meghatározásával segíti az elemzési feladat végrehajtását.

A felmérések szerint a legelterjedtebb módszertan egyértelműen a CRISP_DM (http://www.crisp-dm.org/).

SAS cég is meghatározta a saját módszertanát:
SEMMA (Sample, Explore,Modify, Model, Assess)
(mintavételezés, vizsgálat, módosítás, modellezés és értékelés lépésekből áll a módszertan.)
(*www.sas.com/technologies/analytics/datamining/miner/**semma**.html*)
Nem tekinthető egy teljes adatbányászati projekt megvalósítását leíró módszertannak, hiszen annál szűkebb: csak az elemzés és a modell építés fázisaira szűkül.

Lépései:
Sample (mintavételezés): Amennyiben nagy mennyiségű adattal állunk szemben, vegyünk mintát, amely elég kicsi ahhoz, hogy kezelhető legyen, de elég nagy, hogy magában hordozzon minden lényeges információt, amit a teljes sokaság is! (Alkalmazása opcionális!)
Explore (vizsgálat) : Keressünk adatainkban összefüggéseket, trendeket, kiugró értékeket, ezzel segítve az adatok megértését, és hipotézisek felállítását!

Modify (módosítás) : Transzformáljunk változókat, hagyjunk el lényegteleneket, és hozzunk létre újakat a jobb modellalkotás elősegítéséhez!

Model (modellezés): Építsünk modelleket, melyek automatikusan megtalálják a szignifikáns változókat, amelyek jó becsléseket, előrejelzéseket képesek adni!

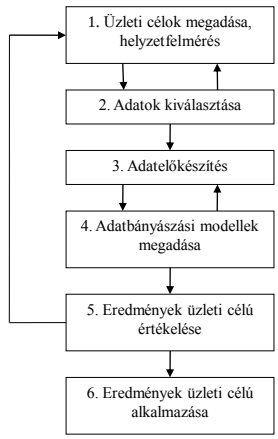
Assess (értékelés) : Értékeljük adatainkat az adatbányászati projekt eredményeinek hasznosulása és megbízhatósága alapján! Vizsgáljuk meg, mekkora pontossággal működnek modelljeink!

Adatbányászati módszertanok elterjedése:

(KDNuggets 2004-es felmérése 170 válasz)

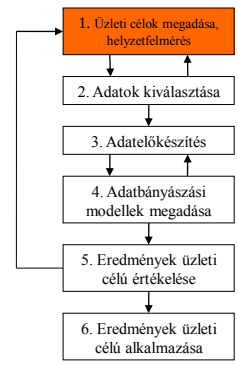
Módszertan	Százalékos arány
CRISP-DM (72)	42 %
SEMMA (17)	10 %
A cég saját módszertana (11)	6 %
Saját módszertan (48)	28 %
Egyéb (10)	6 %
Semmilyen (12)	7 %

Az adatbányászási folyamat lépései a CRISP-DM szerint



A nyilak mutatják, hogy az egyes fázisok között visszacsatolások lehetségesek, sőt bizonyos esetekben szükségesek.

1. Az üzleti célok definiálása, helyzetfelmérés

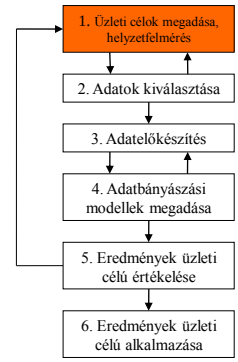


1.Üzleti cél pontos meghatározása, a korlátok és a lehetőségek felmérése.

2.El kell végezni a helyzetfelmérést: rendelkezésre álló erőforrások feltárása (beleértve az idő- és anyagi erőforrást), korlátok, kockázatok feltárása és minden olyan egyéb tényező meghatározása, amely befolyással lehet a feladat sikeres teljesítésére. Nem hagyható el a költségek és a várható haszon vizsgálata sem.

3.Ezek után következik a meghatározott üzleti cél „lefordítása” adatbányászási feladattá.

4.Ennek a fázisnak az utolsó lépéseként el kell készíteni az adatbányászási feladat részletes megvalósítási tervét, amely konkrétan tartalmazza az elemzési lépéseket az alkalmazandó technikákkal és eszközökkel együtt.

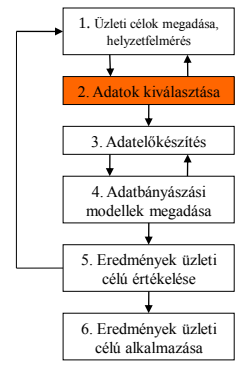


2. Az adatok kiválasztása

1.Meg kell adni a megfogalmazott feladat megoldásához szükséges adatokat.

2.Általában iterációs szakasz következik: összhangba kell hozni az elérendő célt és a rendelkezésre álló adatokat, ami az előző fázisba való többszöri visszalépéssel oldható meg.

3.A kiindulási adatok összegyűjtése saját adattárházból, információs rendszerekből, egyéni táblákból, ill. külső adatok beszerzése (pl. árfolyam adatok).



4.Érdemes ellenőrizni az adatokat feltöltöttség, lefedettség és adathelyesség szempontjából is.

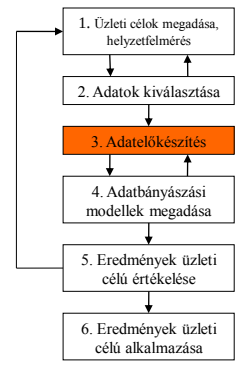
5.Célszerű az összegyűjtött adatokat elemezni: megvizsgálni a tulajdonságaikat, alapstatisztikai számításokat végezni.

3. Adatelőkészítés (előfeldolgozás)

Cél: javuljon az adatok minősége és ezzel az adatbányászás eredményessége.

Fontos, mert megalapozott döntések csak megbízható adatok ismeretében hozhatók.

Feladat: Az adatokban fellépő anomáliák megkeresése, kijavítása és szükség esetén az elemezendő adatmennyiség csökkentése → az ezekre fordított idő bőven megtérülhet a döntéshozatal során.



Az adatok előfeldolgozásának műveletei

Ügyfélkód	Név	Cím	Előfiz. dátum	Folyóirat
1001	Kiss János	Petőfi u.5	06.01.15.	HVG
1001	Kiss János	Petőfi u.5	03.02.01.	Autós élet
1001	Kiss János		05.05.01.	Blikk
1023	Nagy József	Viola u.10.	01.01.01.	Blikk
1045	Nagy János	Rózsa u. 3.	02.03.01.	Sport újság
1056	Kiss János	Petőfi u.5	01.01.01.	Lakáskultúra

Kiinduló adatok: Folyóirat előfizetéseket tartalmazó rekordok → ezek esetén nézzük meg az előfeldolgozás leggyakrabban előforduló feladatait.

Adattisztítás

Szükséges, mert a rendelkezésre álló adatok gyakran hiányosak, zajosak, duplán előfordulhatnak, inkonzisztensek (belső ellentmondást tartalmaznak).

Hiányzó adatok esetén használható eljárás:

- a sor törlése,
- a hiányzó érték manuális kitöltése,
- numerikus adatoknál az attribútum átlagértékkel történő pótlása (pl. ügyfél keresete adat esetén),
- legvalószínűbb érték használata a hiányzó érték pótlására.

Adattisztítás

Zaj: az adatokra rakódott, véletlenszerű hibák.
Előfordulása: gyakori, ha az adatok mérési eredményekből adódnak.
Gond: A zaj miatt fellépő kiugró értékek általában távol vannak a helyes adatoktól.
Javítása:

- egyik módja az *adatsimítás*, amely a kiugró értéket a szomszédos adatok alapján módosítja,
- használható módszerek továbbá a kosarazás és a klaszterezés is.

Adattisztítás

Duplázódás : általában tévedés miatt.
Megszüntetése: a felhasználó döntése alapján, vagy mintaelemzési technikák segítségével történhet.
Ellentmondásos adatok kiszűrésére már az adatok megadásánál lehetőség van.
Az adatbáziskezelő rendszerek ugyanis lehetőséget nyújtanak arra, hogy a bennük tárolt információkra *megszorításokat lehessen előírni* (constraint).

Adatok integrálása és transzformálása

*Az **adatok integrálása*** az a folyamat, amikor a kitűzött feladat megoldásához szükséges, különböző forrásból származó adatokat egyetlen adattárban egyesítik.
Probléma:

- ***egyedazonosítás*** → El kell dönteni pl. a különböző helyről származó adatok esetében ugyanazt jelenti-e a vevő_kód és a vásárló_ID?
- ***redundáns adatok*** *kiszűrése,*
- ***ellentmondó adatértékek felderítésére*** és kijavítása.

Új adatok beszerzése és a felismert hibák javítása utáni állapot:

Ügyfél -kód	Név	*Szül. dátum*	*Jövedelem Ft-ban*	*Gépkocsi*	Cím	Előfiz. dátum	Folyóirat
1001	Kiss János	*75-01-23*	*220 000*	*van*	Petőfi u.5	06.01.15.	HVG
1001	Kiss János	*75-01-23*	*220 000*	*van*	Petőfi u.5	03.02.01.	Autós élet
1001	Kiss János	*75-01-23*	*220 000*	*van*	*Petőfi u.5*	05.05.01.	Blikk
1023	Nagy József	*85-05-03*	*90 000*	*van*	Viola u.10	01.01.01	Blikk
1045	Nagy János	*80-09-13*	*150 000*	*nincs*	Rózsa u.3	02.03.01.	Sportújság
1001	Kiss János	*75-01-23*	*220 000*	*van*	Petőfi u.5	01.01.01.	Lakáskultúra

Adatok transzformálása

A kiindulási adatokat az adatbányászás céljainak, az alkalmazni ***kívánt technikának megfelelő alakra*** kell hozni.
Pl. a neuron hálózatok alkalmazása *numerikus adatokat* igényel, míg a döntési fáknál *szimbolikus értékek* is használhatók.
A transzformációnál vigyázni kell arra, hogy *ne történjen információvesztés.*
Gyakran használt transzformáció a ***normalizálás,*** amikor az adatokat egy előre megadott tartományba vetítik le, pl. a [0,1] intervallumba.

Az adatok redukálása

Cél: az adathalmaz elemzése ne váljon lehetetlenné az elemzéshez szükséges idő miatt.
Csökkentés végrehajtása: adathalmaz méretében jóval kisebb legyen, de a redukált adathalmazon végzett bányászat várhatóan az eredetivel azonos, vagy majdnem azonos eredményt adjon.
Lehetséges módszerek:

- ***dimenziócsökkentés,***
- ***adattömörítés,***
- ***számosságcsökkentés.***

Az adatok redukálása

Dimenziócsökkentés:redukálja az adathalmaz méretét. Az adatbányászási feladat szempontjából érdektelen attribútumo(ka)t eltávolítja, vagy minden attribútumot figyelembe véve új, kevesebb számú attribútumot hoz létre.
Adatok tömörítése: az adatokat kódolással, vagy transz-formációs műveletekkel hozzuk „tömörebb" alakra.
Számosság csökkentés: a tárolandó rekordok számát csökkentjük
Mindhárom esetben számos módszer létezik.

A végső adatok

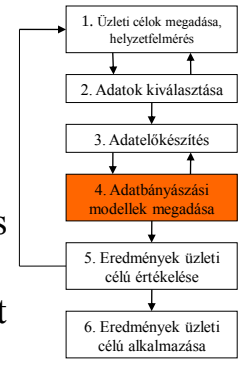
Kód	Név	Kor	Jövedelem	Autó	Ter. kód	Folyóirat
1001	Kiss János	31	220	1	1	HVG
1001	Kiss János	31	220	1	1	Autós élet
1001	Kiss János	31	220	1	1	Blikk
1023	Nagy József	21	90	1	1	Blikk
1045	Nagy János	26	150	0	2	Sport újság
1001	Kiss János	31	220	1	1	Lakáskultúra

A végső adatok

A változtatások:

1. Címek helyett területi kódok → tömörítés.
2. Születési dátum helyett a kor és a jövedelem helyett jövedelem/1000 → így ez a két érték közel azonos nagyságrendű → könnyebben értelmezhető és ábrázolható a kettő közötti kapcsolat.
3. Ha a különböző újságok olvasói közötti kapcsolatok az érdekesek, akkor az előfizetési dátum nem fontos, ezért elhagyjuk.

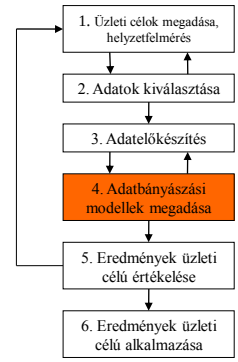
4. Adatbányászási modellek megadása



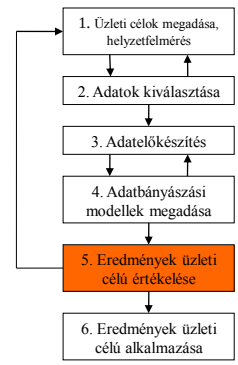
Modellező technika kiválasztása az elemezni kívánt adatok típusa alapján → másféle algoritmus használható numerikus adatok esetén, vagy ha pl. jó adós, átlagos adós, rossz adós kategória változók szerint kell elemezni.

Modell tesztelési módszerének meghatározása → adatokból tanuló mintát és teszt mintát kell létrehozni.

Modellalkotás: a kiválasztott algoritmus alkalmazását jelenti.
Az előzőleg kiválasztott technikával elvégezésre kerül a mintakeresés a rendelkezésre álló adatok között.
A *modell kiértékelése:* a feltárt ismeretek értelmezését, megjelenítését jelenti.
Ha az eredmények nem megfelelők, akkor az előző fázishoz való visszatéréssel kell folytatni a feladat megoldását.



5. A kapott eredmények üzleti célú értékelése
Kapott eredmények vizsgálata: mennyire felelnek meg az előzetes üzleti elvárásoknak.
Döntés a további teendőkről:
- befejezés esetén: eredményeket értelmezhető formában át kell adni a felhasználóknak ,
- újabb kérdések felvetődése esetén vagy ha a kapott eredmények nem használhatók → iterációs folyamat következik.

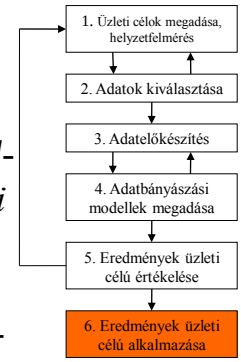


6. Az eredmények üzleti célú alkalmazása

Ebben a fázisban kell megtervezni az *eredmények felhasználását, beépítését az üzleti folyamatokba*

Biztosítani kell az alkalmazás *„frissítését”,* elkerülve a már nem aktuális adatokra épült modellek alkalmazását.

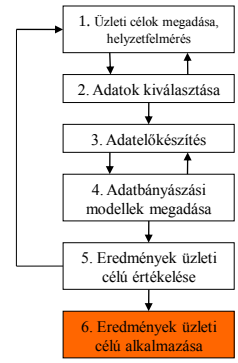
Összefoglaló jelentés készítése: tartalmaznia kell a végrehajtott lépéseket, a kapott eredményeket és a jövőre vonatkozó, más feladatoknál is használható tapasztalatokat.



Célszerű ismertetni a pozitívumok mellett a negatívumokat is: hol van szükség felülvizsgálatra, módosításokra.

Megismerve a felhasználók tapasztalatait, összehasonlítható a valóság és a modell.

A különbségek feltárása új modellek megalkotását ösztönözheti, vagy az összehasonlítás akár olyan tény(eke)t is kimutathat, amely(ek) a továbbiakban üzleti célként is használható(k).



Vezetői információs rendszerek