

Modellezés IBM SPSS Modelerrel

SPSS Nyári iskola 2019. július 11.

HOGYAN IS NEVEZZELEK?



betanews.com/2019/02/05/artificial-intelligence-or-data-science/

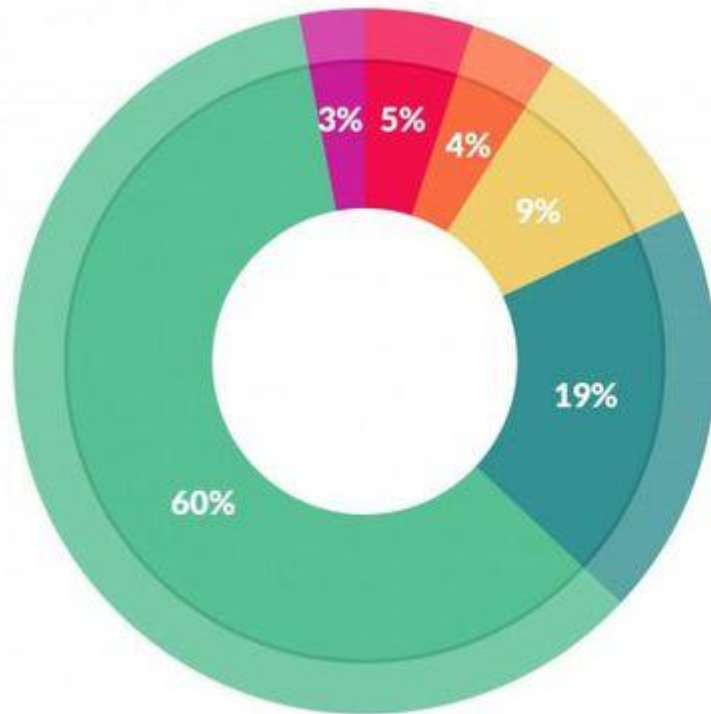
ADATBÁNYÁSZAT

- Tudásfeltárás
- Bányászat: a megmozgatott anyagnak csak kis része kerül hasznosításra
- Cél: olyan eszköz kifejlesztése, amely információszerzés céljából elemzi a nyers adatokat. Érvényes, újszerű, hasznos mintázatokat keresünk az adatokban.
- 1989: első konferencia



Is this the place to learn about mining?

MIT CSINÁL EGY ELEMZŐ?

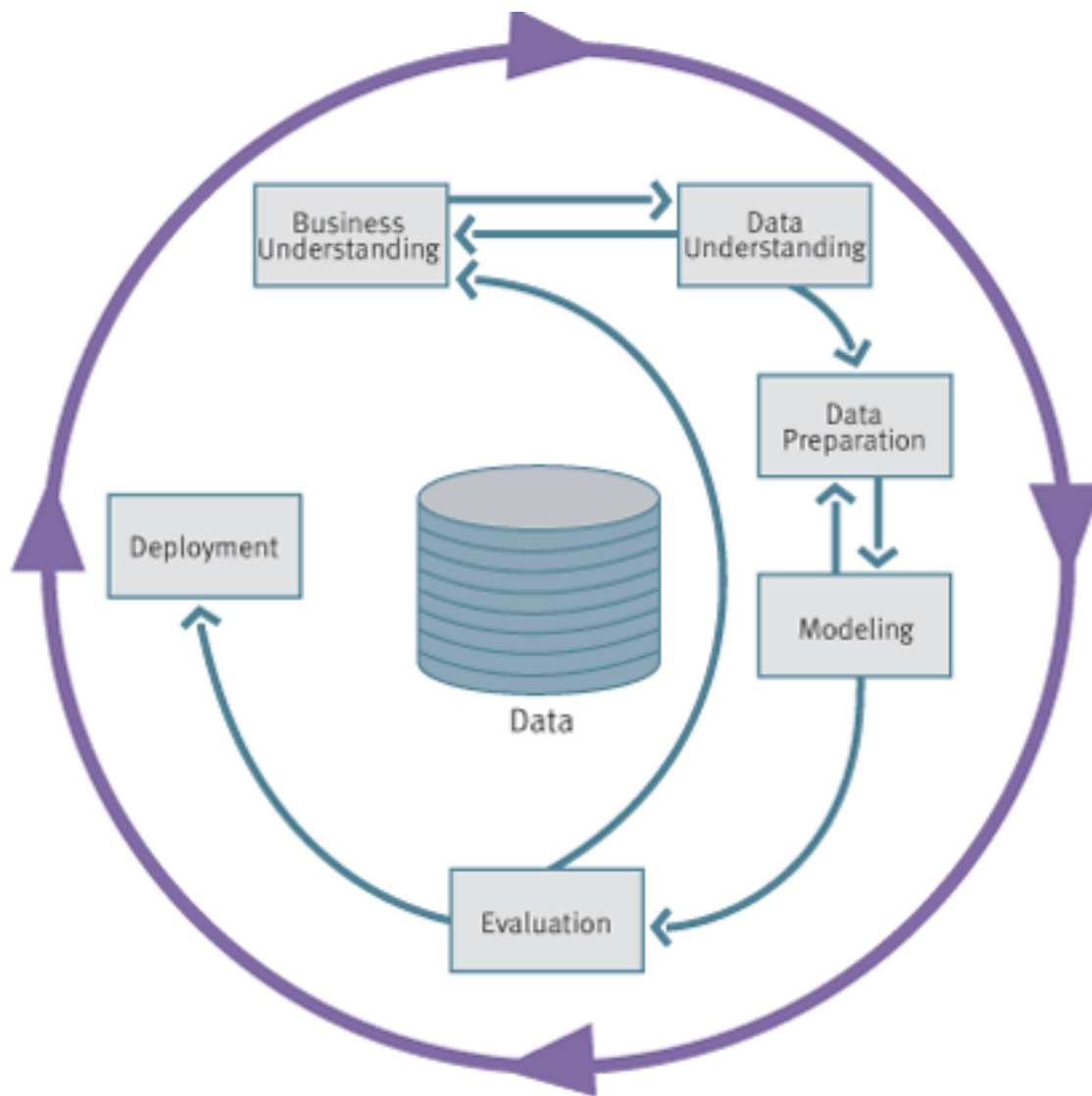


What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

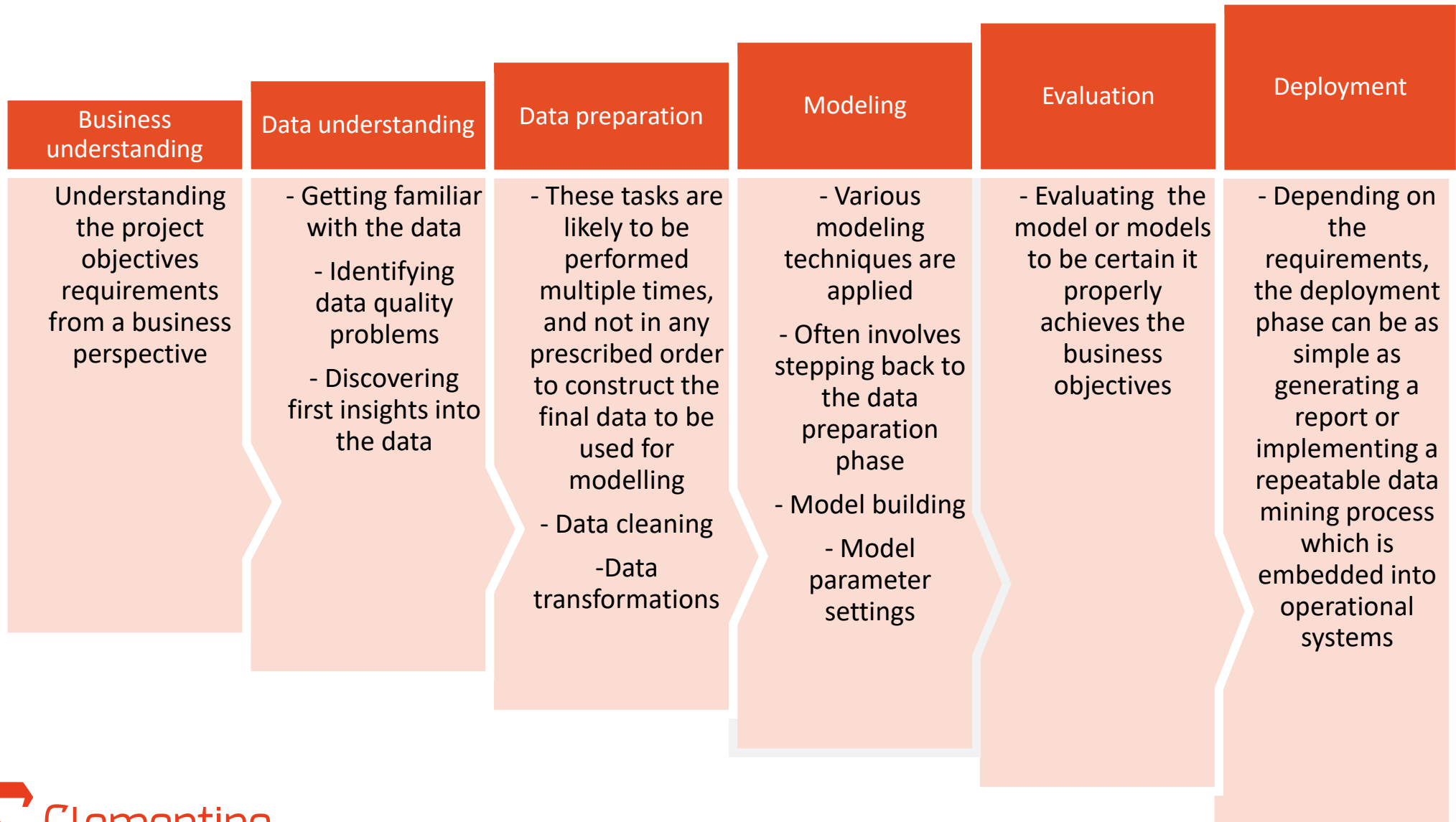
www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#45699ad66f63

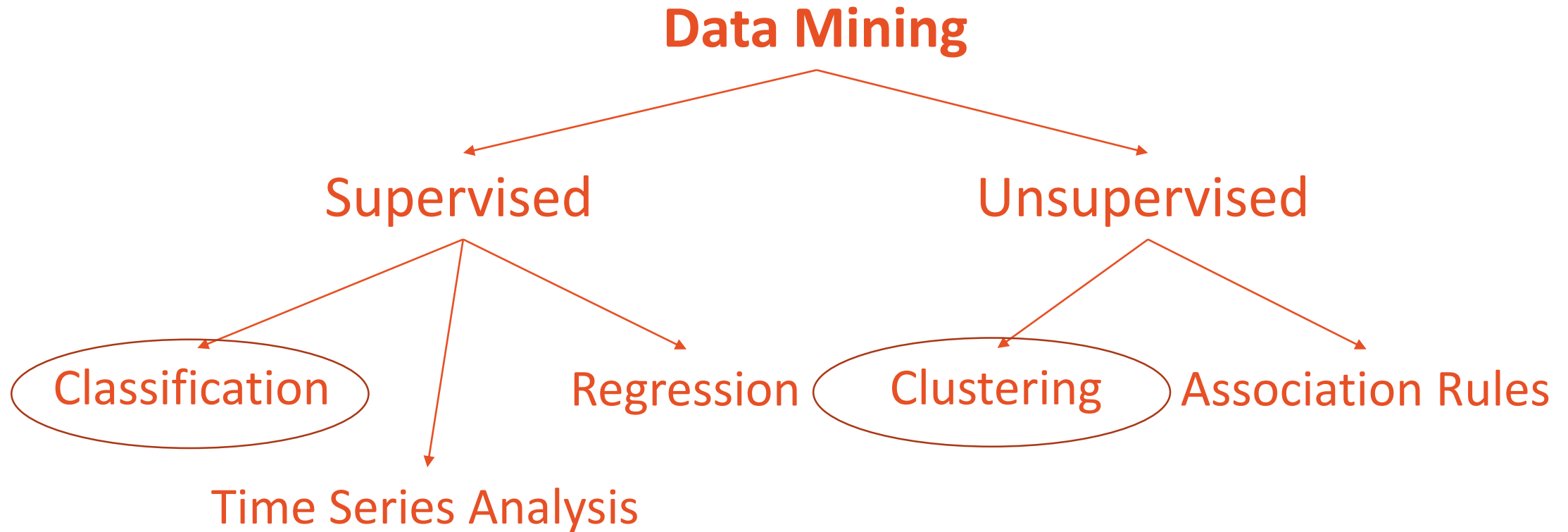
ADATBÁNYÁSZATI PROJEKT - MÓDSZERTAN



Cross-Industry
Standard Process
for
Data Mining

ADATBÁNYÁSZATI PROJEKT - MÓDSZERTAN



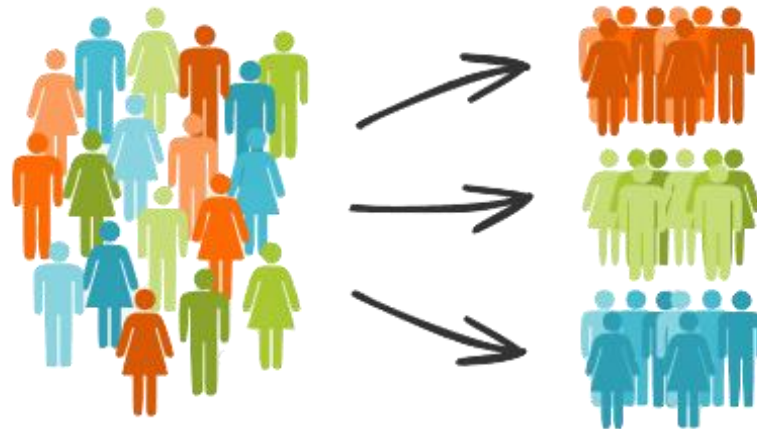


Klaszterezés

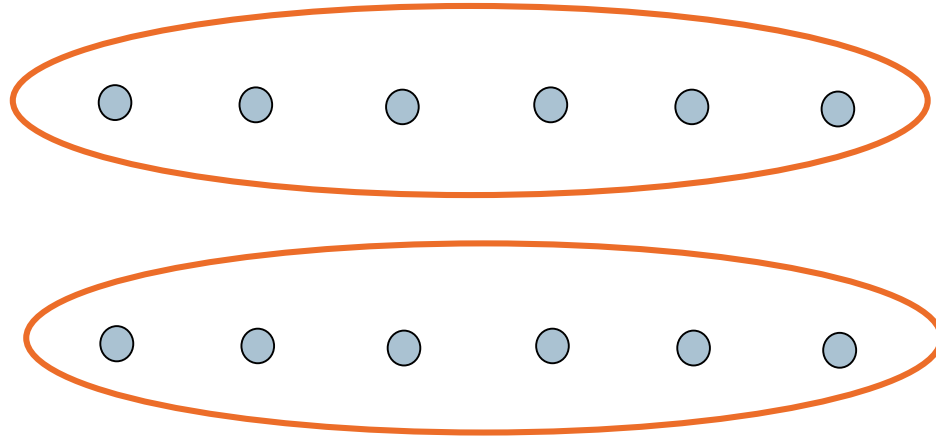
SPSS Nyári iskola 2019. július 11.

KLASZTEREZÉSI ELJÁRÁSOK

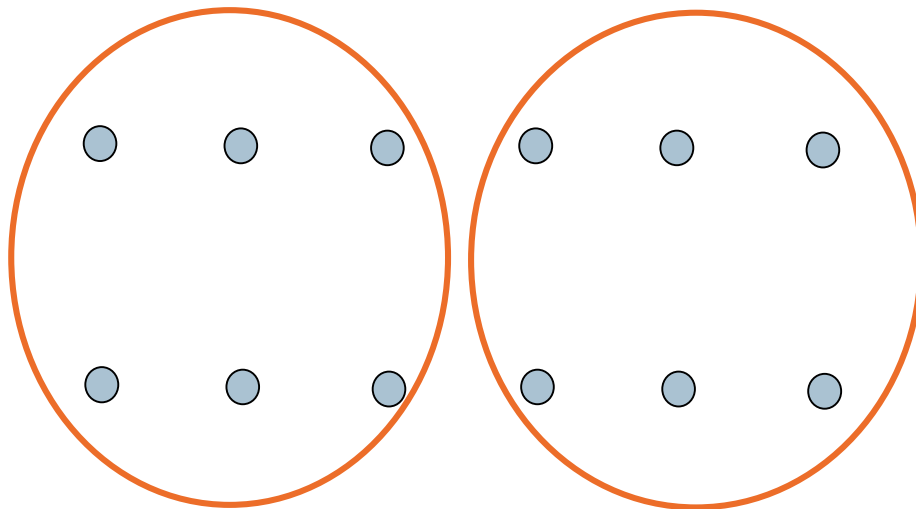
- **Klaszterezés ~ szegmentáció:** elemek csoportosítása úgy, hogy a valamilyen közös tulajdonsággal rendelkező elemek azonos, az egymástól jelentősen eltérő elemek különböző csoportba kerüljenek. Feltáró technika. Dimenziószám-csökkentő eljárás.
- **Alkalmazási terület:**
 - Ügyfél-szegmentáció
 - Viselkedési profilok
 - Csillagászat – csillagok csoportosítása mozgásuk alapján
 - Orvostudomány – génkutatás, pszichiátriai betegek csoportosítása
 - Biológia – állat- és növényközösségek csoportosítása



SZUBJEKTIVITÁS



SZUBJEKTIVITÁS



A klaszterezés szubjektív!

HASONLÓSÁG -> TÁVOLSÁGFÜGGVÉNY

Pontok távolsága

- Első lépés: változók normalizálása
- Egy numerikus változó (A) esetén:
 $\text{Távolság}(X,Y) = A(X) - A(Y)$
- Több numerikus változó esetén:
Euklideszi távolság: $d(X,Y) = \sqrt{\sum_i (X_i - Y_i)^2}$
- Kategorikus változó esetén:
dummy változók, megfelelő skálázással

KLASZTEREZÉSI ELJÁRÁSOK

▪ Particionáló eljárás

A végső, megfelelő klaszterezést a pillanatnyi eredményként kapott klaszterezés folyamatos, iteratív pontosításával érjük el. A klaszterek száma előre adott. Egyszerre határozza meg a klasztereket.

Pl.: K-Means, GMM

▪ Hierarchikus eljárás

Egymásra épülő lépések sorozata, amely a korábbi lépés klasztermegoldásai alapján hozza létre a következő klasztereket. Lehet összevonó (Agglomerative) és felosztó (Divisive). Az összevonáson alapuló algoritmus először minden egyes elemet külön klaszternek tekint és összekapcsolja őket egyre nagyobb klaszterekbe, míg a végén egyetlen, az összes elemet tartalmazó klasztert kapunk. Az elemeket egy hierarchikus adatszerkezetbe (fa, dendrogram) rendezi.

Pl.: TwoStep, HDBSCAN

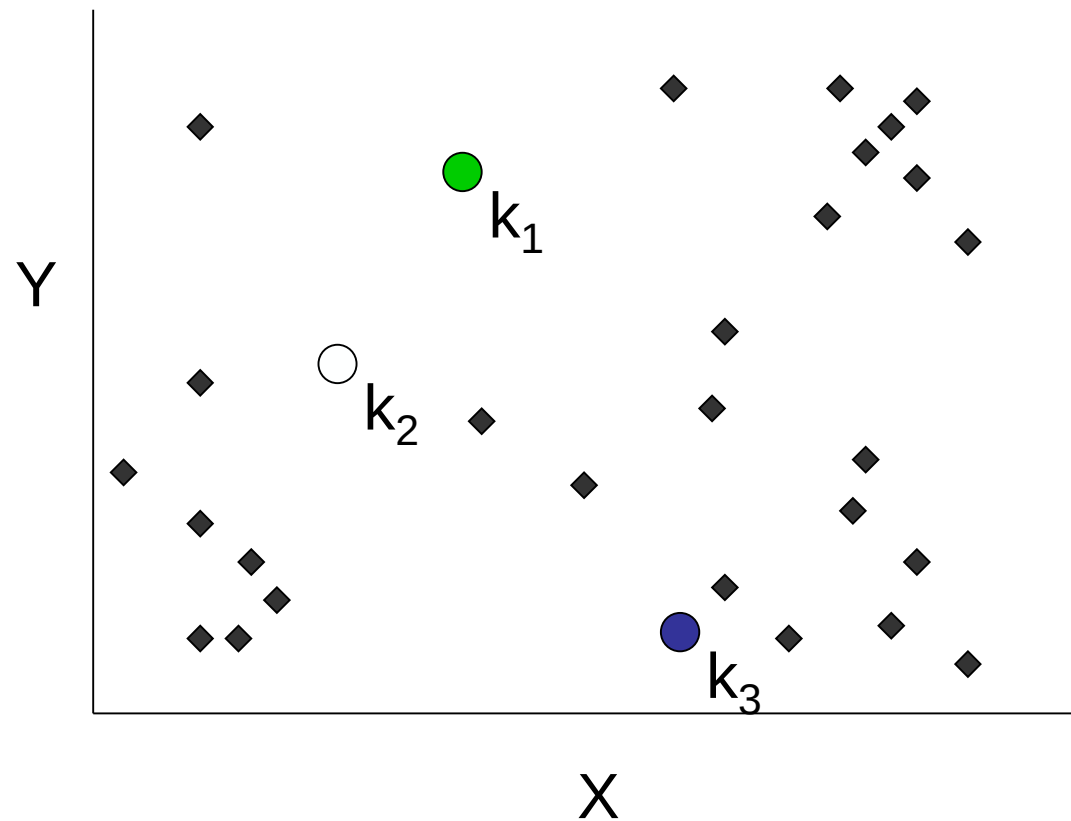
ITERÁCIÓS ALGORITMUS - K-MEANS

Iterációs algoritmus

1. Kiválaszt k db kezdeti klaszterközéppontot (lehető legtávolabb egymástól)
2. Minden elemet hozzárendel a legközelebbi középponthoz (Euklideszi távolság alapján)
3. A klaszterközéppontokat az így létrejött klaszterek valódi középpontjába teszi át
4. A 2. és 3. lépést addig ismételi, míg eleget nem tesz valamilyen megszakítási kritériumnak (pl.: küszöbérték, időkorlát)

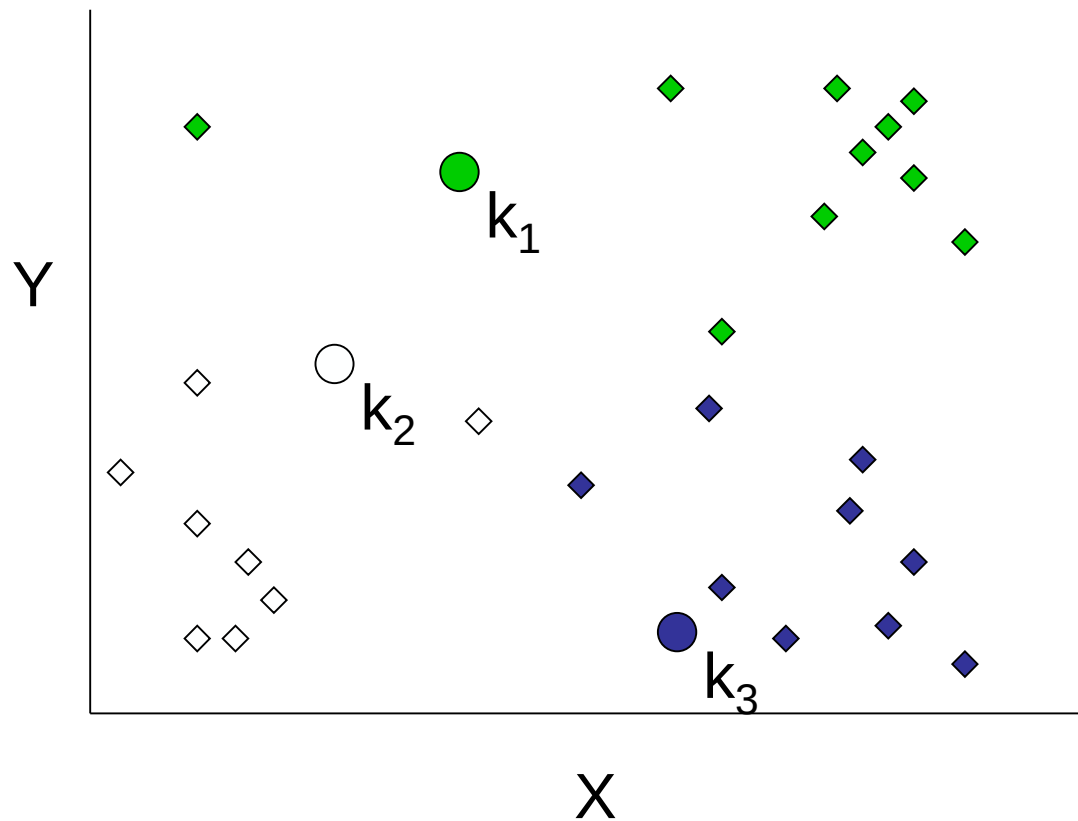
K-MEANS – PÉLDA 1. LÉPÉS

Kijelöl három
klaszter-
középpontot
véletlenszerűen



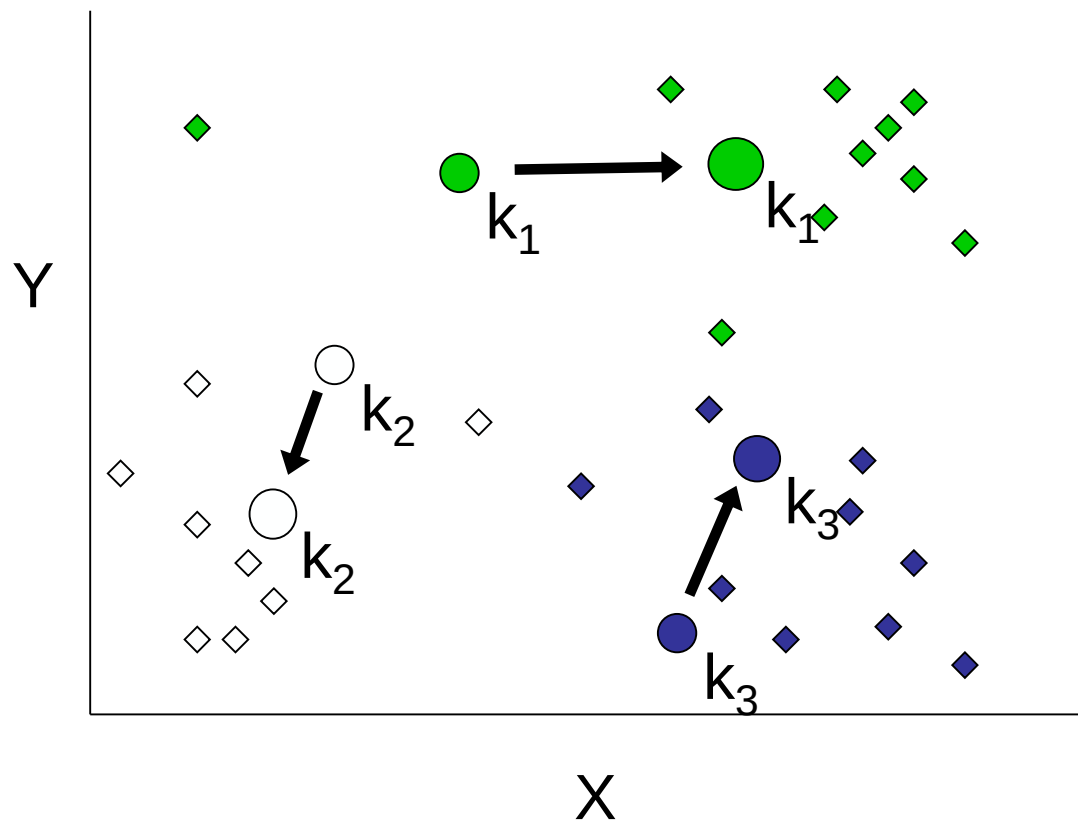
K-MEANS – PÉLDA 2. LÉPÉS

Minden elemet
a hozzá
legközelebbi
középponthoz
rendel

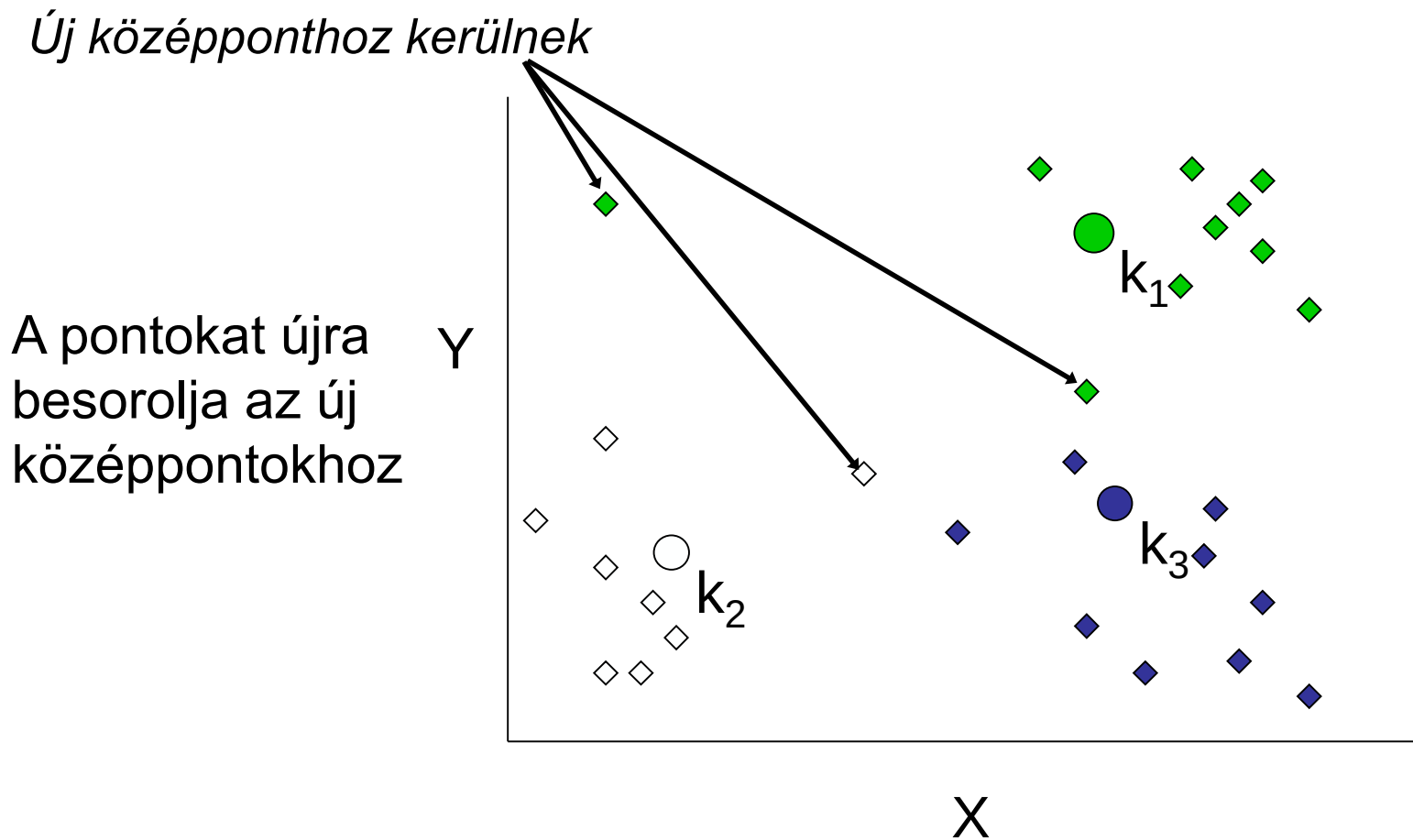


K-MEANS – PÉLDA 3. LÉPÉS

Áthelyezi a
középpontokat
a klaszterek
tényleges
középpontjába

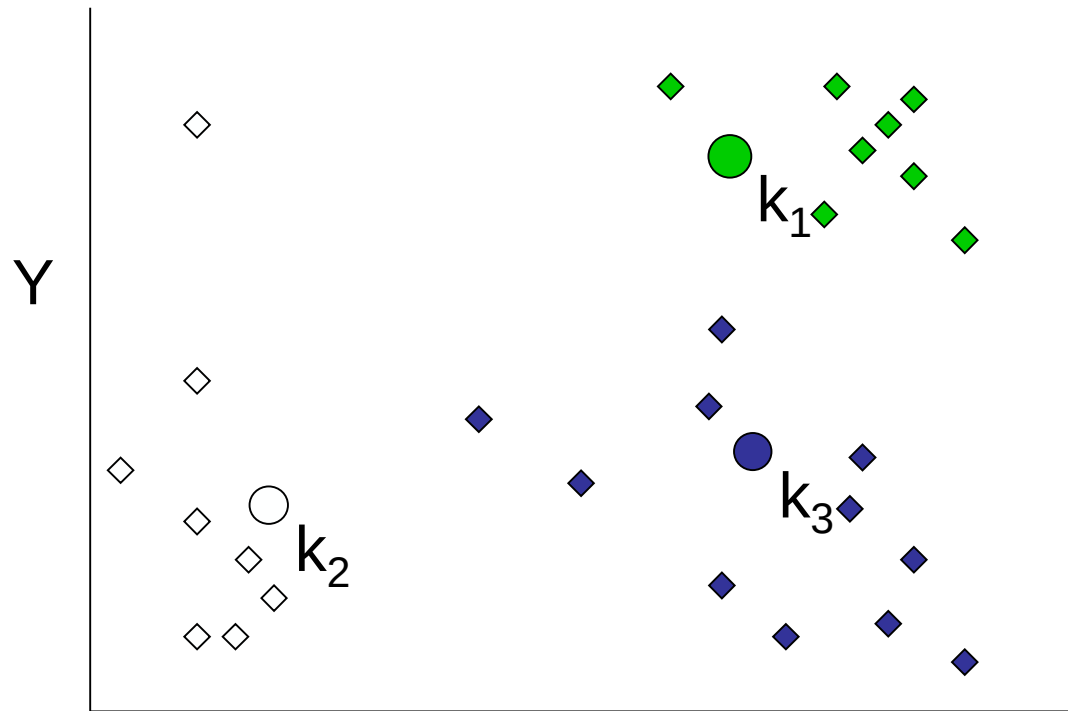


K-MEANS – PÉLDA 2. LÉPÉS ÚJRA



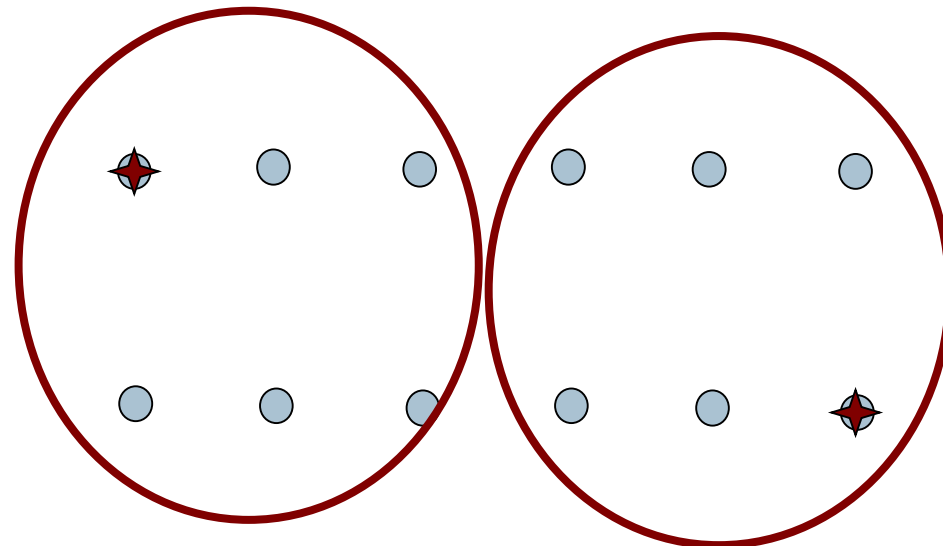
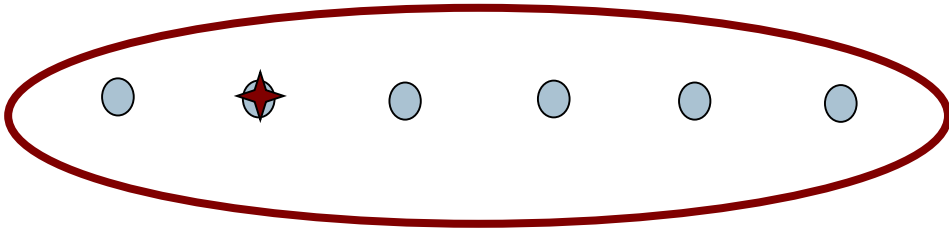
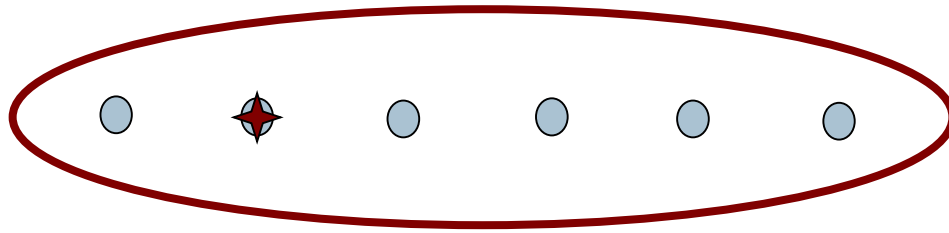
K-MEANS – PÉLDA 2. LÉPÉS ÚJRA

A pontokat újra
besorolja az új
középpontokhoz

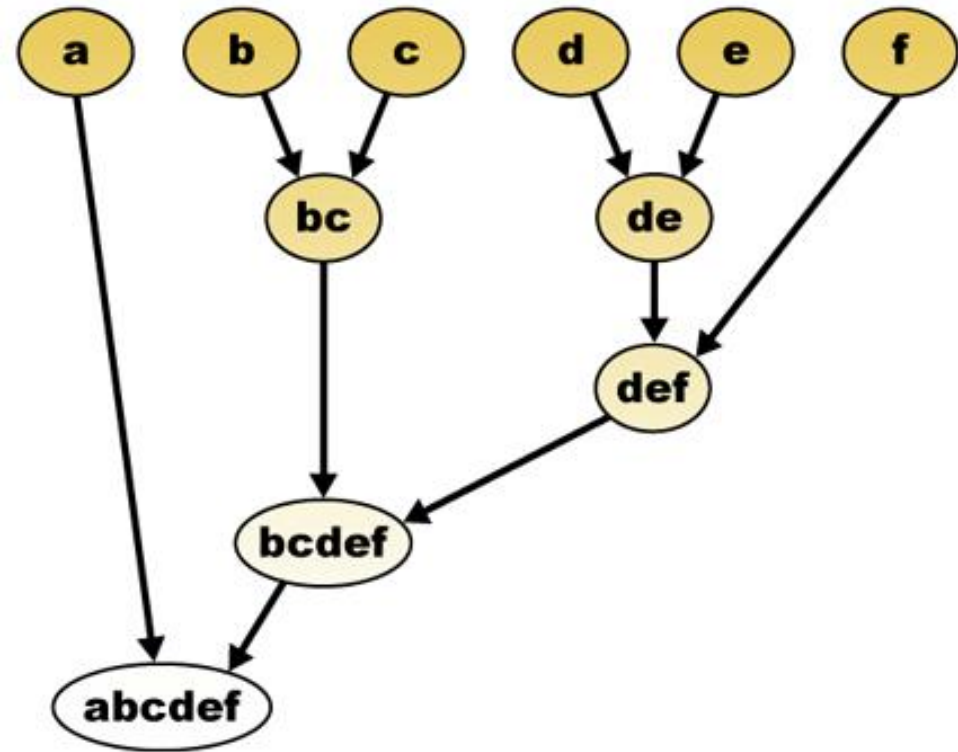
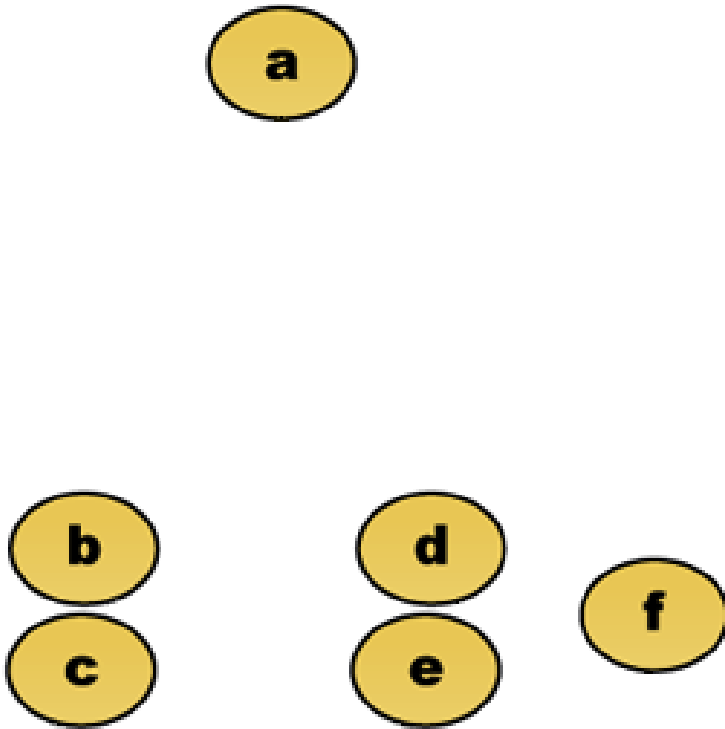


Majd újra áthelyezi a középpontokat...

KEZDŐPONTOK MEGVÁLASZTÁSA



HIERARCHIKUS ALGORITMUS



KLASZTEREZÉSI ELJÁRÁSOK

	Hierarchikus eljárás	Nem hierarchikus eljárás
Előny	•Segít a klaszterek számának meghatározásában.	•Nagy számú mintavételi egységek esetén is használható. •Gyorsabb •Rugalmasabb •Megbízhatóbb eredményeket ad
Hátrány	•Érzékeny a kiugró értékekre. •Lassú folyamat.. •Összevonást nem lehet szétbontani.	•A klaszterek számát előre kell meghatározni. •Klaszterközeppont kiválasztása. •Függ a megfigyelések sorrendjétől. •Lokális optimumot talál meg.

K-MEANS NODE

Jellemzők

- Automatikusan elvégzi az input változók normalizálását
- Hiányzó értékek: 0.5 (kivéve set, ahol a dummy alapváltozó)
- Megadhatja az elemek távolságát a klaszterközeppontról illetve a klaszterek távolságát

Beállítási lehetőségek

- Leállási feltételek
 - Iterációszám
 - Klaszterközepponatok változásának mértéke (tolerance)
- *Set encoding value*: Kategorikus változók skálázása állítható

TWO-STEP NODE

Jellemzők

- A hiányzó értéket tartalmazó rekordokat eldobja
- Választható a változók standardizálása
- Kihagyhatjuk a kiugró értékeket
- Megadhatunk pontos klaszterszámot vagy intervallumot (általában a kisebb értékhez lesz közel az eredmény)
- Választhatunk távolságot: Log-likelihood, Euklideszi
- Választhatunk kritériumot az optimális klaszterszám meghatározásához: BIC, AIC (entrópia alapú)
- TwoStep-AS – Analytic Serveren futtatható változat (Big Data)

KLASZTEREZÉS KIÉRTÉKELÉSE - MUTATÓK

Silhouette

$$SC = \frac{1}{N} \sum_{i=1}^N \frac{\min_{j \in C_{-i}} \{D_{ij}\} - D_{iC_i}}{\max(\min_{j \in C_{-i}} \{D_{ij}\}, D_{iC_i})}$$

- Mennyire hasonló egy pont a saját klaszteréhez a többi klaszterhez viszonyítva – ezt átlagoljuk
- -1 és 1 közötti érték, 0.5 fölött jó, 0.2 alatt nagyon rossz

SSE ~ Sum of Squares Error (összetartás mértéke)

$$SSE = \frac{1}{N} \sum_j \sum_{i \in j} D_{ij}^2$$

- A klaszterközéppontoktól való távolságnégyzetek összegének átlaga
- K-Means erre optimalizál, ennek egy lokális minimumánál áll meg

D_{ij} - i rekord és j klaszterközéppont közötti távolság

D_j - a teljes sokaság középpontja és a j . klaszterközéppont távolsága

N_j - j klaszter elemszáma

C_{-i} - i rekordot nem tartalmazó klaszterek

C_i - i rekordot tartalmazó klaszter

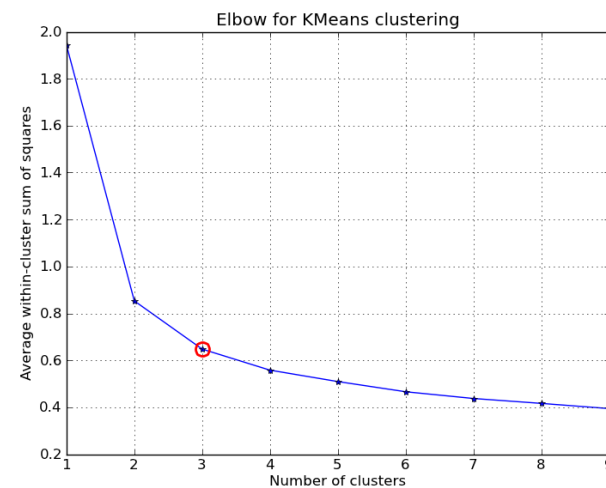
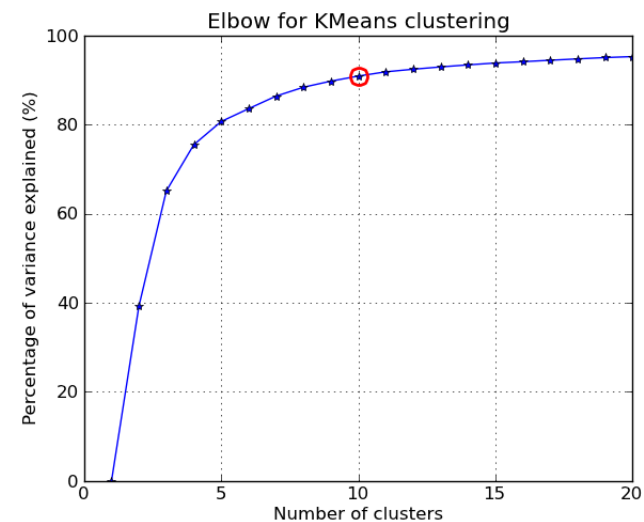
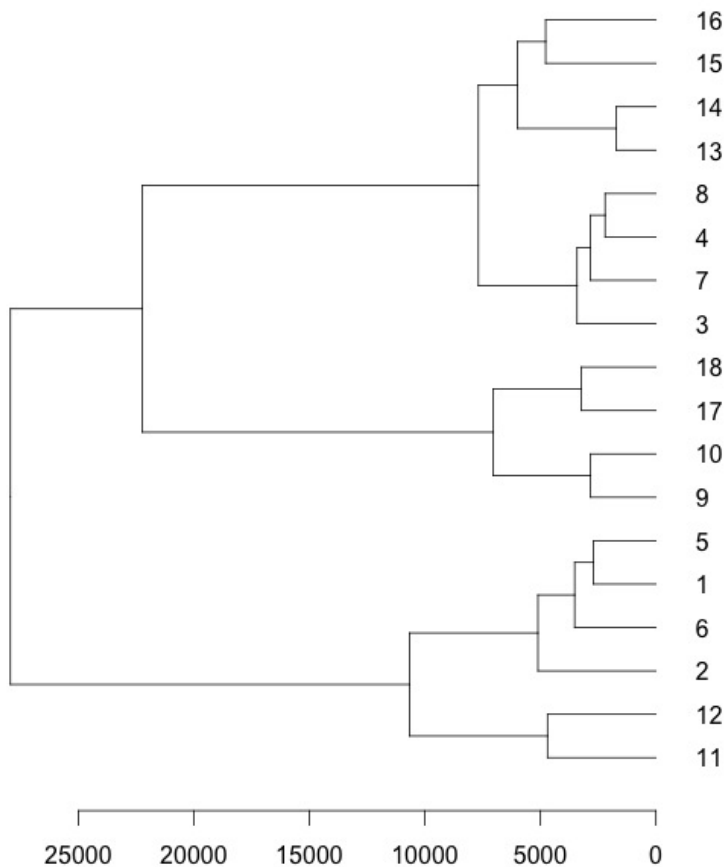
SSB ~ Sum of Squares Between (szeparálás mértéke) $SSB = \frac{1}{N} \sum_j N_j D_j^2$

KLASZTERSZÁM MEGHATÁROZÁSA

- A priori: korábbi tapasztalatok vagy külső elvárások alapján előre meghatározott
- Klaszterösszevonás távolsága hirtelen megnő (pl.: dendrogram alapján)
- Stabilitás: ha n és $n+1$ klaszter esetén $n-1$ db klaszter megegyezik és az n . bomlik fel két kisebbre
- Könyökszabály (Elbow-method)
- Hüvelykujj-szabály $k \approx \sqrt{n/2}$
- Klaszterek relatív mérete
- Klaszterek értelmezhetősége

KLASZTERSZÁM MEGHATÁROZÁSA

Dendrogram



IBM SPSS MODELER – ALGORITMUSOK

Modeler 18.0

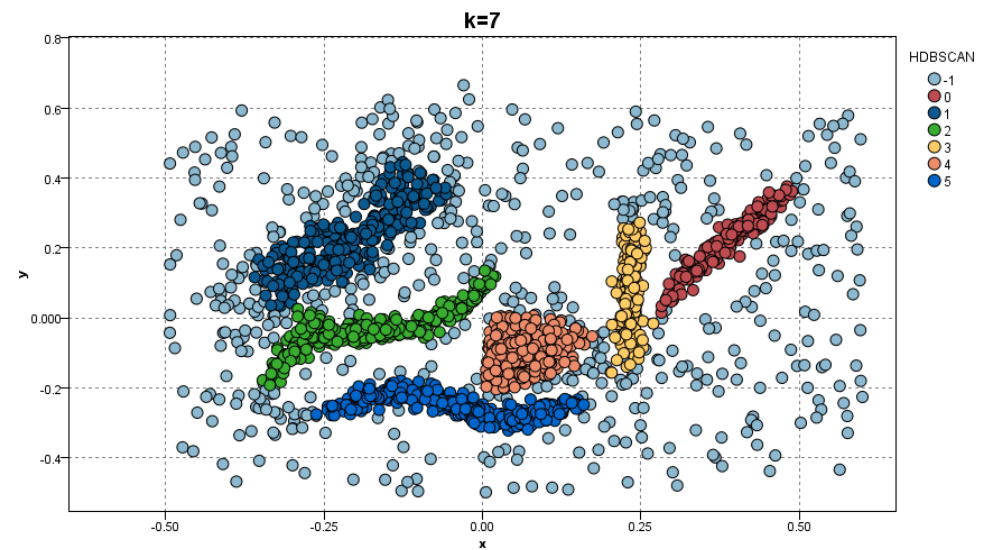
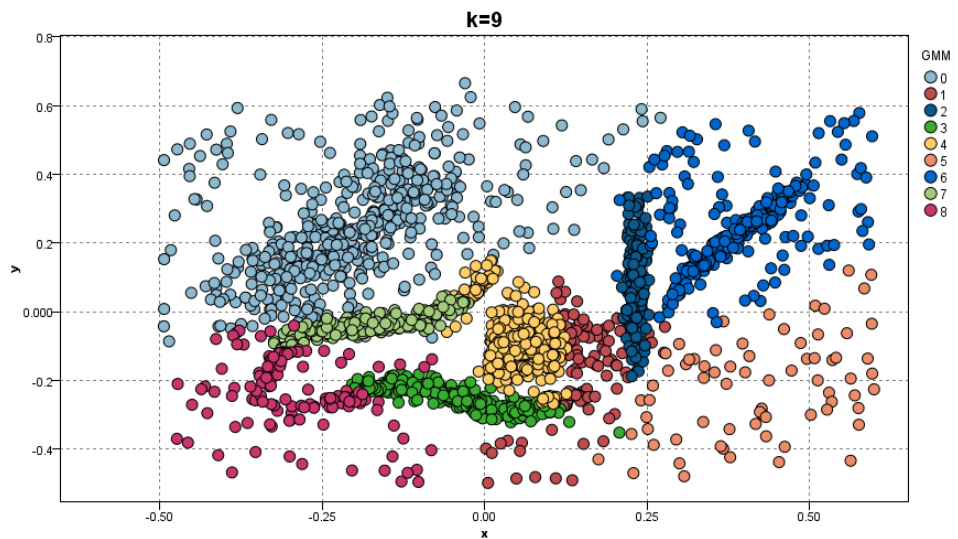
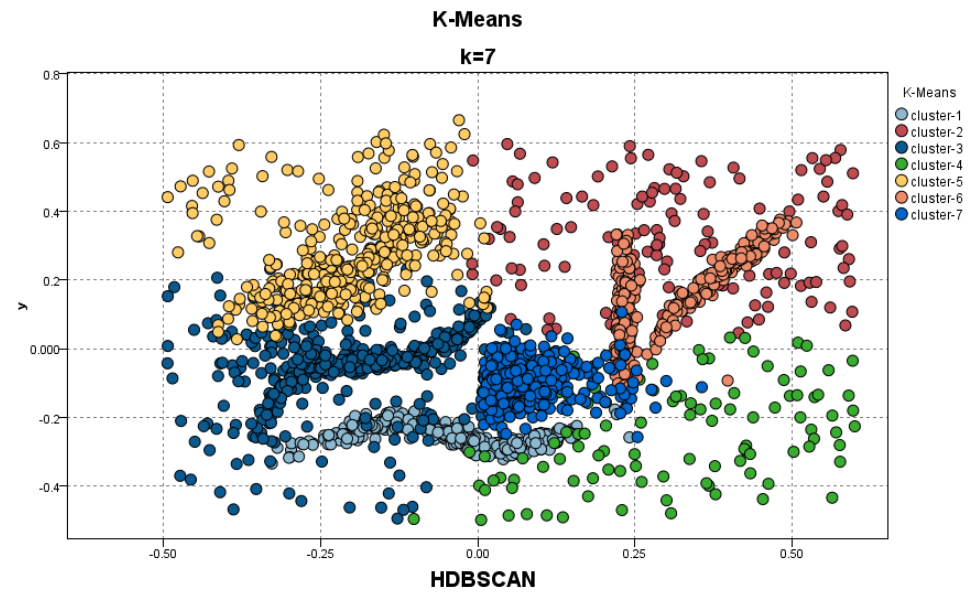
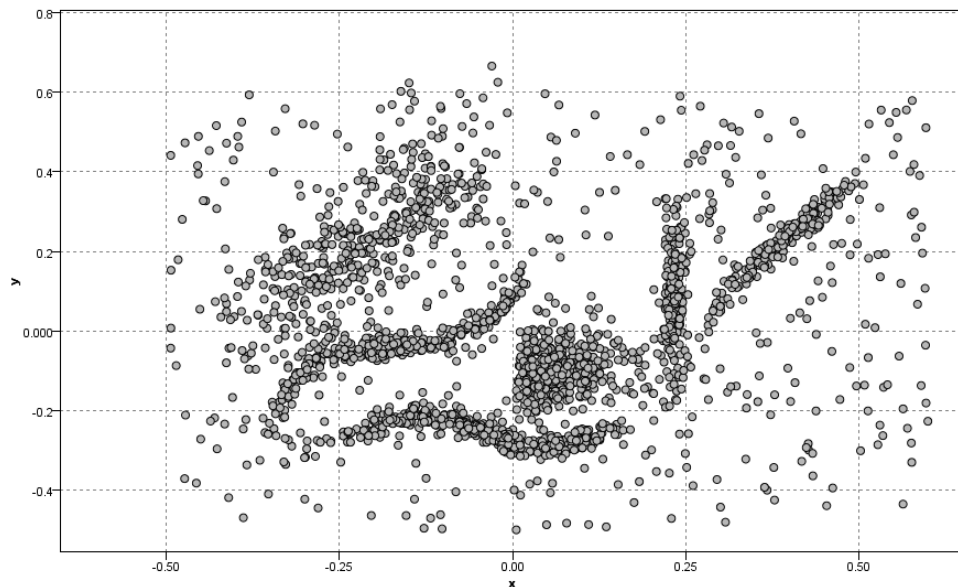
- K-Means
- Two-Step
- Kohonen

- Auto Cluster

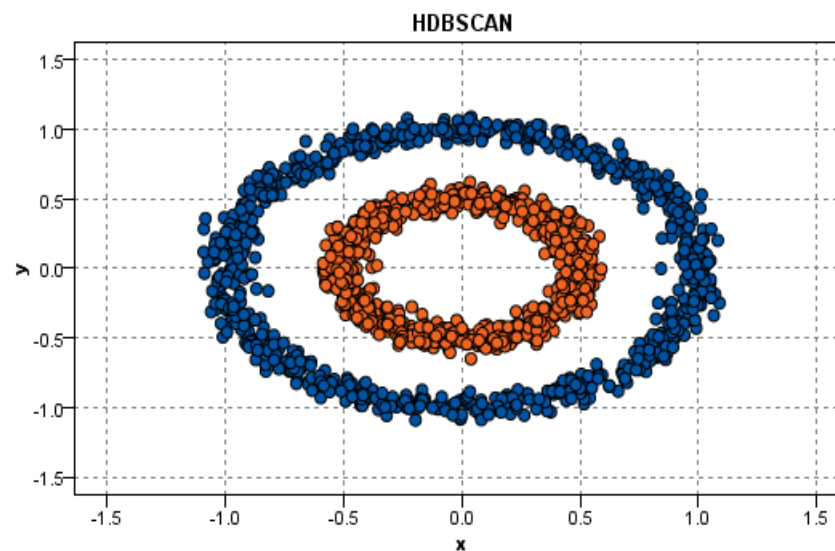
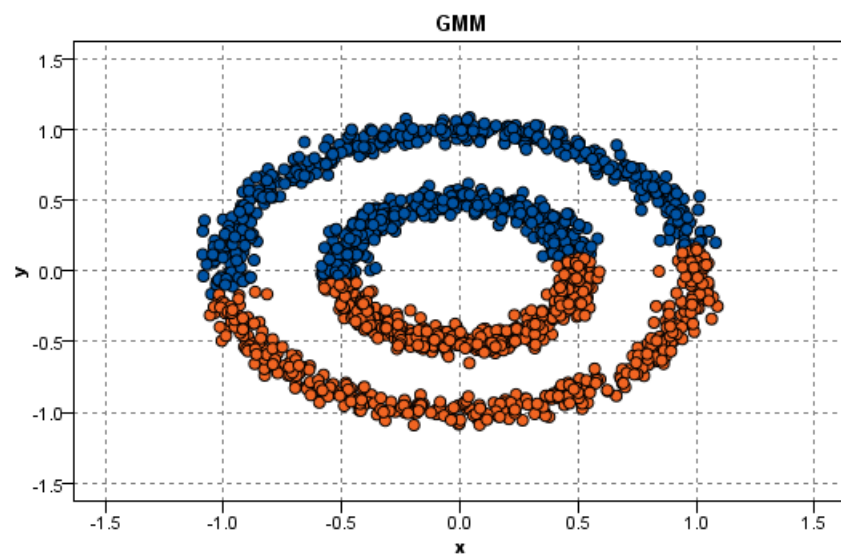
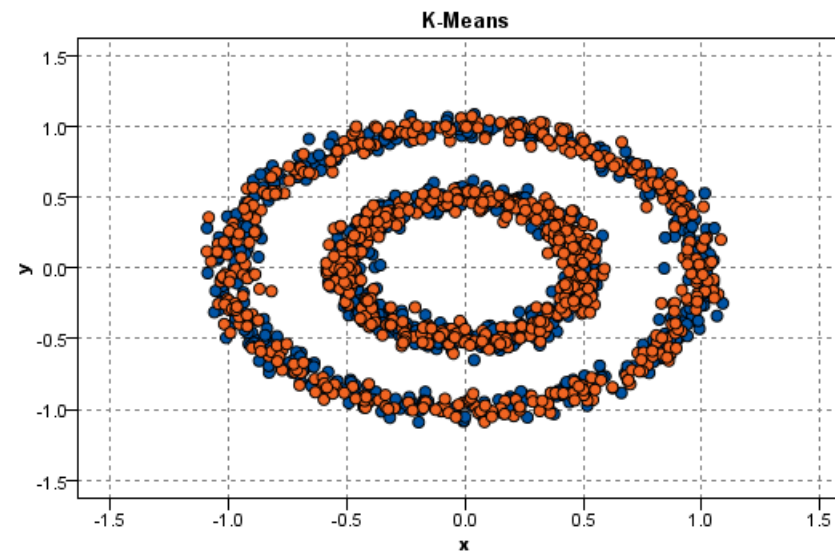
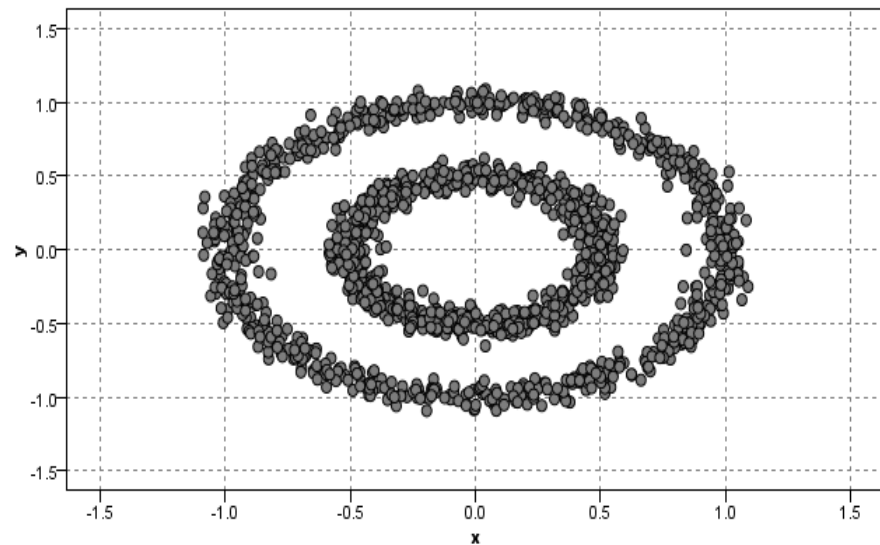
Modeler 18.2

- GMM
- HDBSCAN

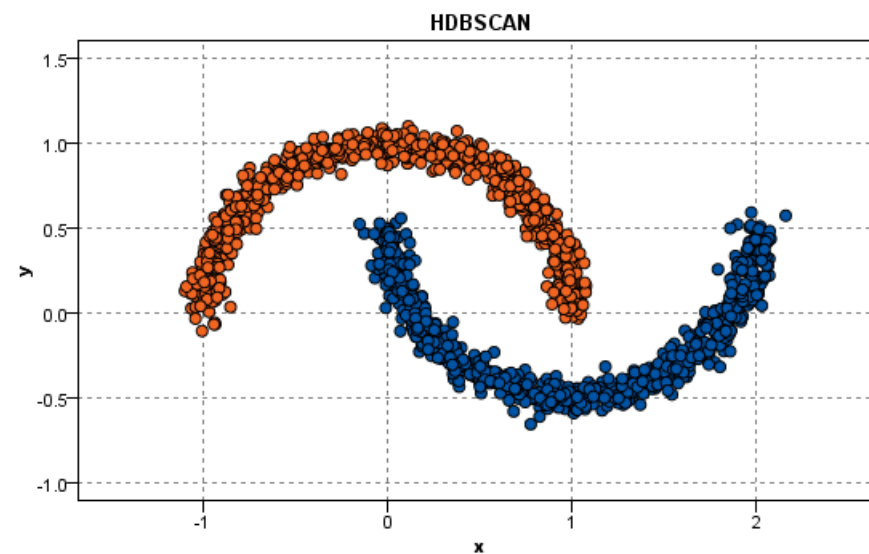
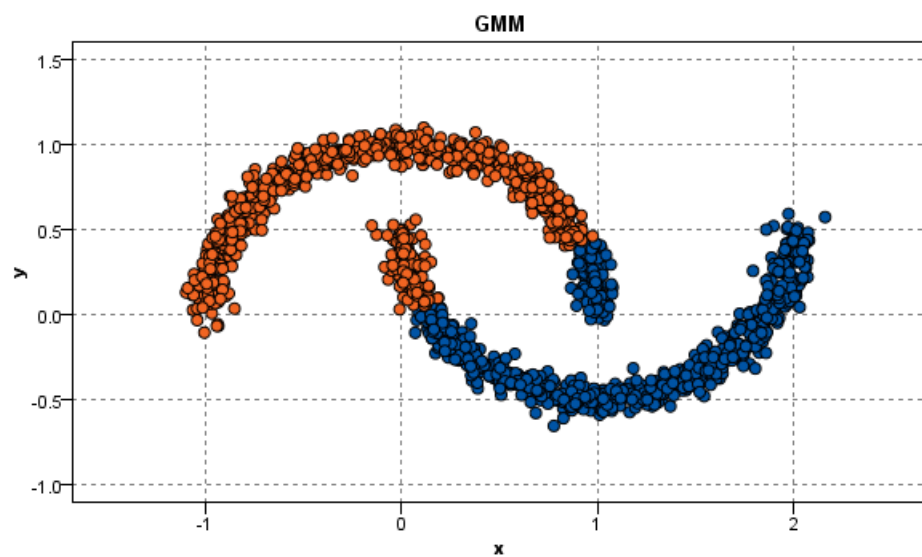
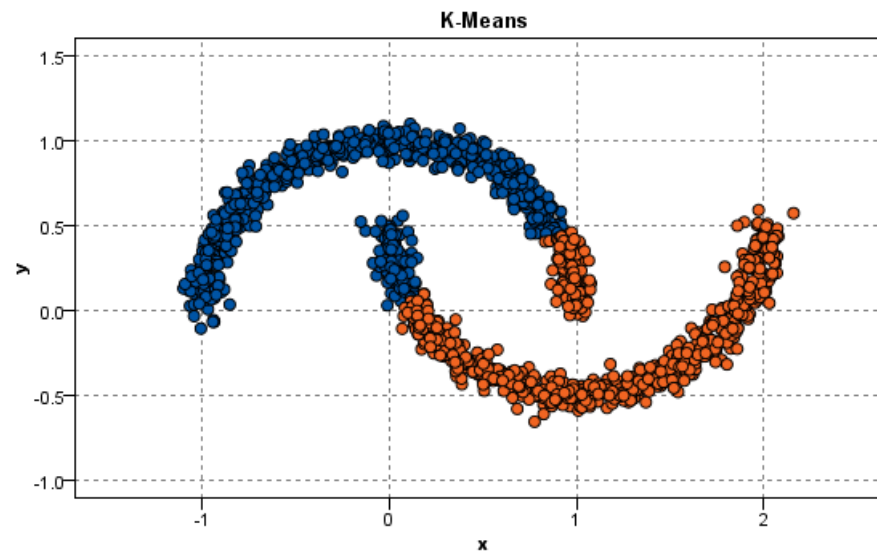
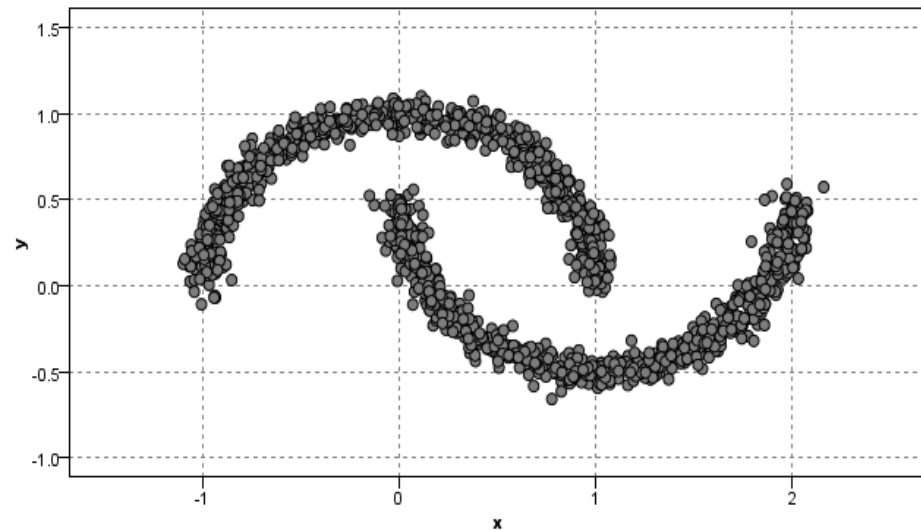
KLASZTEREZŐ ALGORITMUSOK ÖSSZEHAISONLÍTÁSA



CIRCLES



MOONS



CASE STUDY – AIRPORTS

- USA-beli repülőjáratok 2015-ös indulási és érkezési adatai

Adattáblák: www.kaggle.com/usdot/flight-delays

- flights.csv
- airports.csv
- airlines.csv

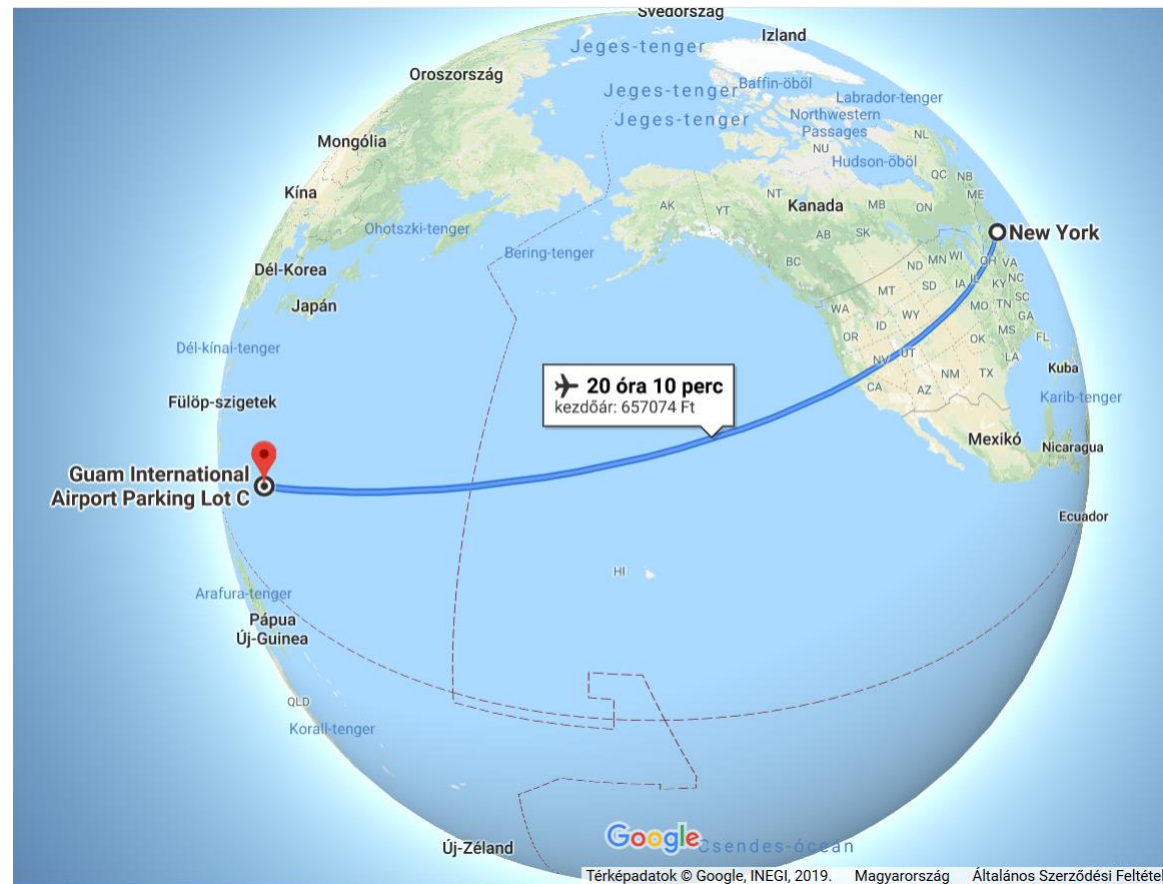
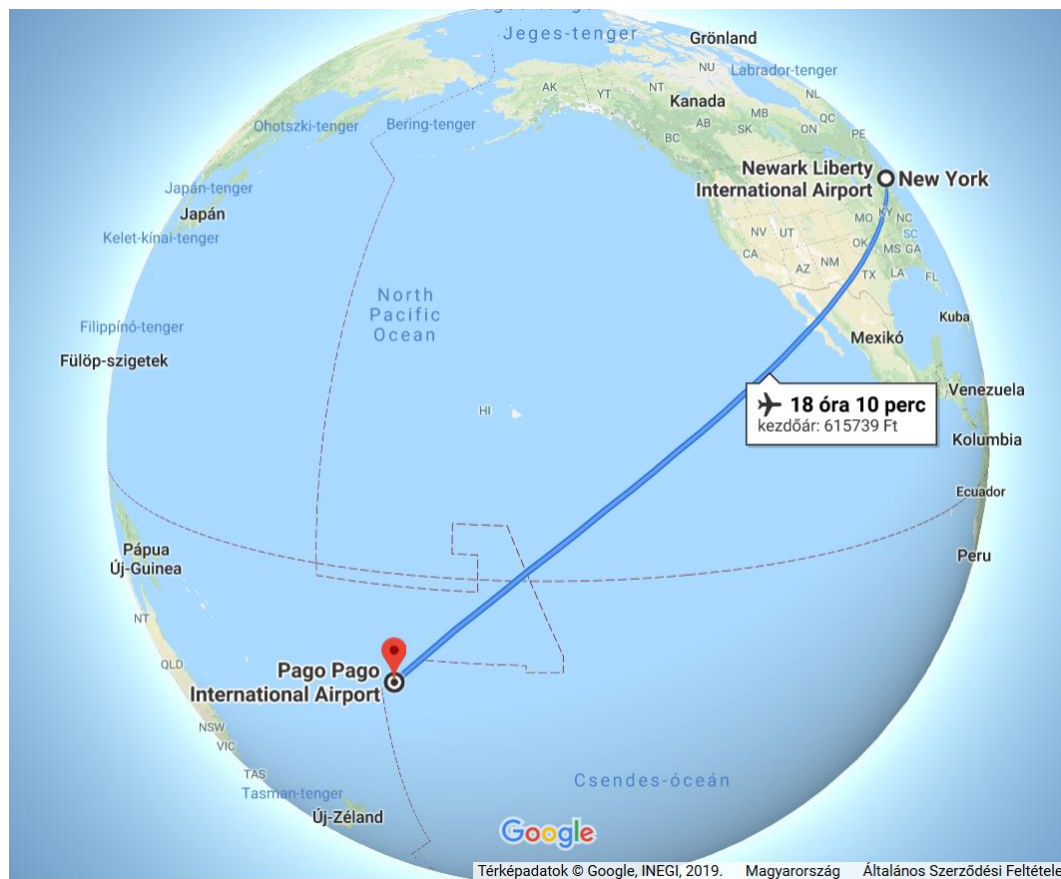
Adatok

- Honnan – Hova
- Tervezett indulás, érkezés
- Tényleges indulás érkezés

Feladat: Reptéri forgalom alapján milyen csoportokba sorolhatók a repterek?



KLASZTER - 2



Klasszifikáció

SPSS Nyári iskola 2019. július 11.

KLASSZIFIKÁCIÓ VS. KLASZTEREZÉS

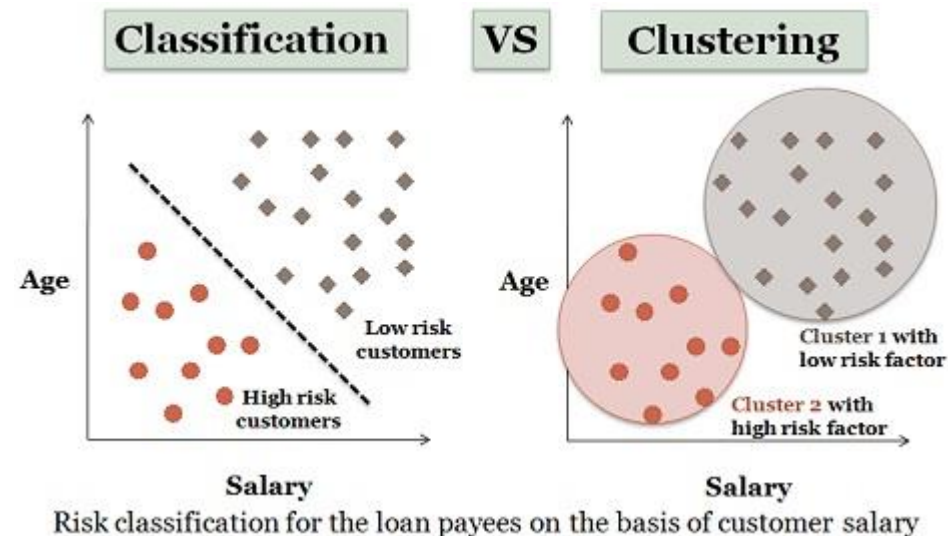
Klasszifikáció

- Osztályok előre adottak (van célváltozó)
- Felügyelt (supervised)
- Van tanító, tesztelő minta
- Leíró és prediktív feladatra

Klaszterezés

- Osztályok előre nem adottak (nincs célváltozó)
- Nem felügyelt (unsupervised)
- Nincs tesztelő minta
- Leíró feladatra

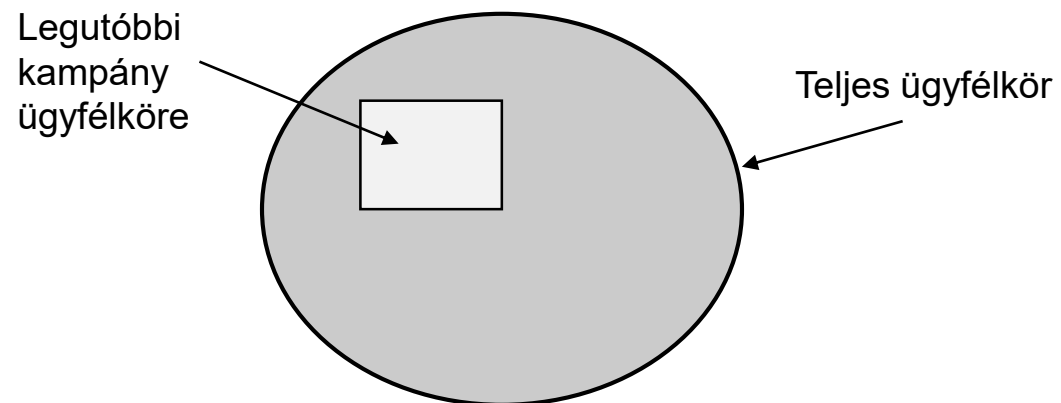
↓
Pontosság



MODELLTANÍTÁS

	Tanítás - Modelling	Alkalmazás - Deployment
Adat	múltbeli	jelenbeli
Célváltozó	ismert	nem ismert
Magyarázó változók	ismert	ismert
	Ugyanaz az adatkör	
Kampány példán		

Példa: bank DM kampány



CASE STUDY – FLIGHTS DELAY

- USA-beli repülőjáratok 2015-ös indulási és érkezési adatai

Adattáblák: www.kaggle.com/usdot/flight-delays

- flights.csv
- airports.csv
- airlines.csv

Adatok

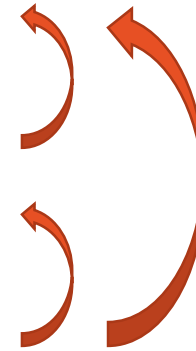
- Honnan – Hova
- Tervezett indulás, érkezés
- Tényleges indulás érkezés

Feladat: Mely járatokat ne válasszam, ha nem szeretnék késni?



KLASSZIFIKÁCIÓ MODELLALKOTÁS LÉPÉSEI

1. Tesztkörnyezet kialakítása
2. Változók modellben betöltött szerepének meghatározása
 - Célváltozó, magyarázó változók
3. Modellek tanítása
 - Többféle algoritmus, többféle beállítás
 - Csak a tanító mintán!
4. Modellek kiértékelése, összehasonlítása
 - Tanító és tesztelő mintán
5. Legjobb modell(ek) kiválasztása



TESZTKÖRNYEZET

- **Tanuló – tesztelő – validáló minta**

A mintát véletlenszerűen szétosztjuk tanuló, tesztelő, esetleg validáló mintára (általában 70% - 20% - 10%). Csak nagy elemszám esetén alkalmazható. Tanuló mintán létrehozunk többféle modellt, a tesztelőn kiválasztjuk a legjobbat, majd a validáló mintán értékeljük a modell jóságát.

- **Keresztvalidálás ~ k-fold Cross Validation**

A mintát véletlenszerűen k egyenlő részre osztjuk, minden körben az egyik részminta a tesztelő minta, a maradék k-1 pedig a tanító minta. A modell pontossága a felépített k modell pontosságának átlaga.

- **Leave-one-out ~ LOOCV**

N elemű mintából minden körben egy elemet kiveszünk, a többi n-1 elemet tanítunk. A kihagyott elemre kiszámítjuk a modell hibáját. A modell pontosságát ezen n db hiba átlaga adja.

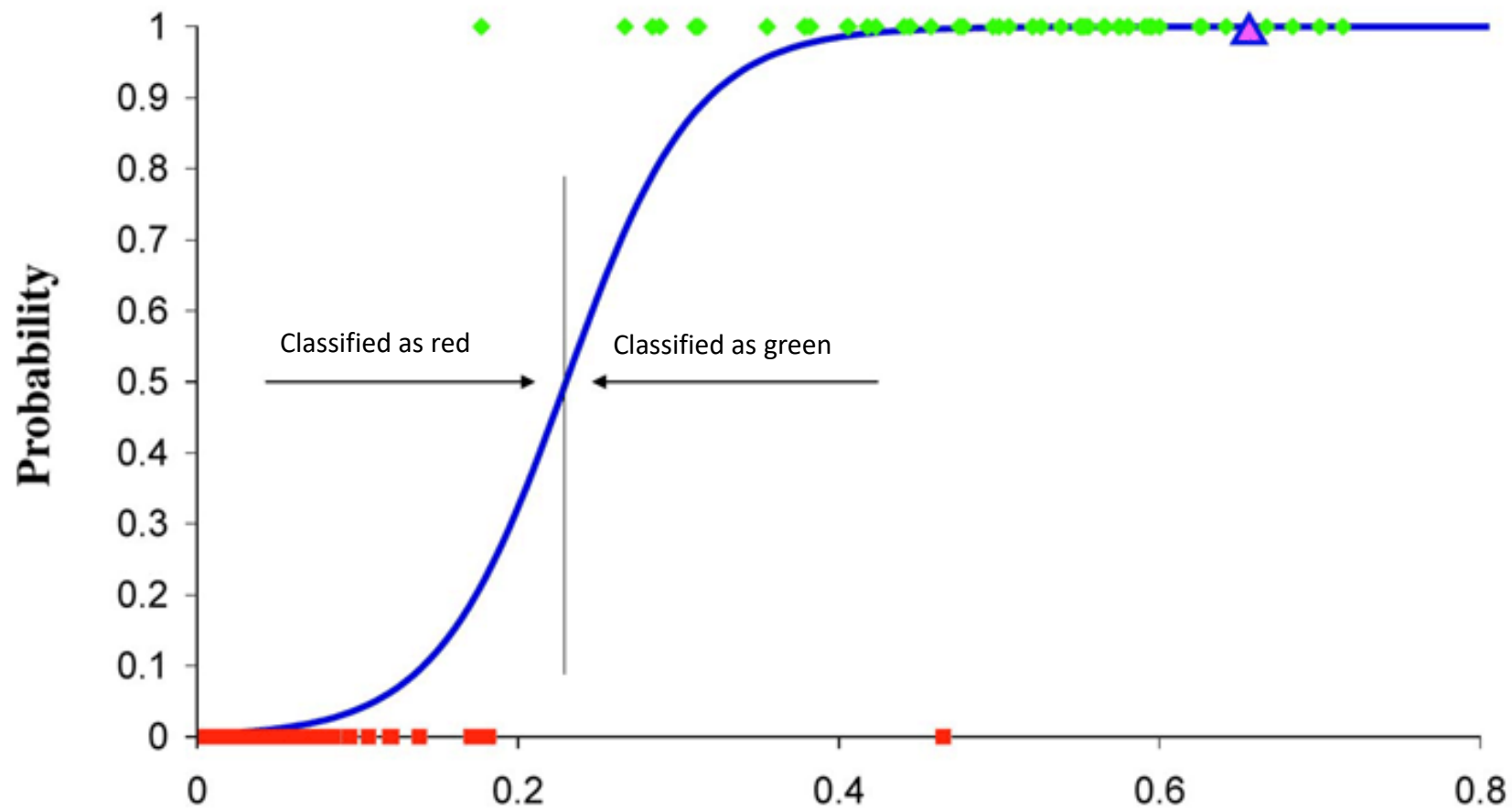
! Tesztelő mintát csak teszteléshez használjuk (tanítás).

! Véletlen mintavétel.

TESZTKÖRNYEZET

Módszer	Előny	Hátrány
Tanító – tesztelő – validáló minta	▪ egyszerű ▪ olcsó ▪ Modeler – Partition node	▪ adatvesztés ▪ nagy minta szükséges ▪ eredmény függ a minta kiválasztásától, nagy a varianciája
Keresztvalidálás ~ k-fold CV	▪ legmegbízhatóbb ▪ kis elemszám esetén is ▪ kisebb adatvesztés, mint a tesztelő mintánál	▪ számításigényes ▪ nincs a Modelerben, scripttel megoldható
Leave-one-out ~ LOOCV	▪ nincs benne véletlenszerűség ▪ CV-nél pontosabb, kisebb a tanításnál az adatvesztés	▪ számításigényes, ezért csak kis elemszám esetén ▪ tesztminta mindig egyelemű, így nem reprezentatív

LOGISZTIKUS REGRESSZIÓ



LOGISZTIKUS REGRESSZIÓ

- **Binomiális logisztikus regresszió:** két, egymást kölcsönösen kizáró kategória bekövetkezési esélyeinek egymáshoz viszonyított arányát, az odds mértékét modellezi a magyarázó változók értékének ismeretében.
- **Odds:** az egyes bekövetkezési esélyek egymáshoz viszonyított aránya

p – esemény bekövetkezésének valószínűsége ($Y=1$)

$$odds = \frac{P(Y=1)}{P(Y=0)} = \frac{p}{1-p}$$

- Feltételezés: az odds logaritmusa lineárisan függ a magyarázó változóktól

$$\ln(odds) = \text{logit}(p) = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_n \cdot x_n$$

- Az egyes kategória bekövetkezési valószínűsége az oddsból számítható:

$$p = \frac{odds}{1+odds} = \frac{e^{\beta_0 + \beta_1 \cdot x_1 + \dots + \beta_n \cdot x_n}}{1 + e^{\beta_0 + \beta_1 \cdot x_1 + \dots + \beta_n \cdot x_n}} \quad 1-p = \frac{1}{1+odds} = \frac{1}{1 + e^{\beta_0 + \beta_1 \cdot x_1 + \dots + \beta_n \cdot x_n}}$$

- Döntési szabály:

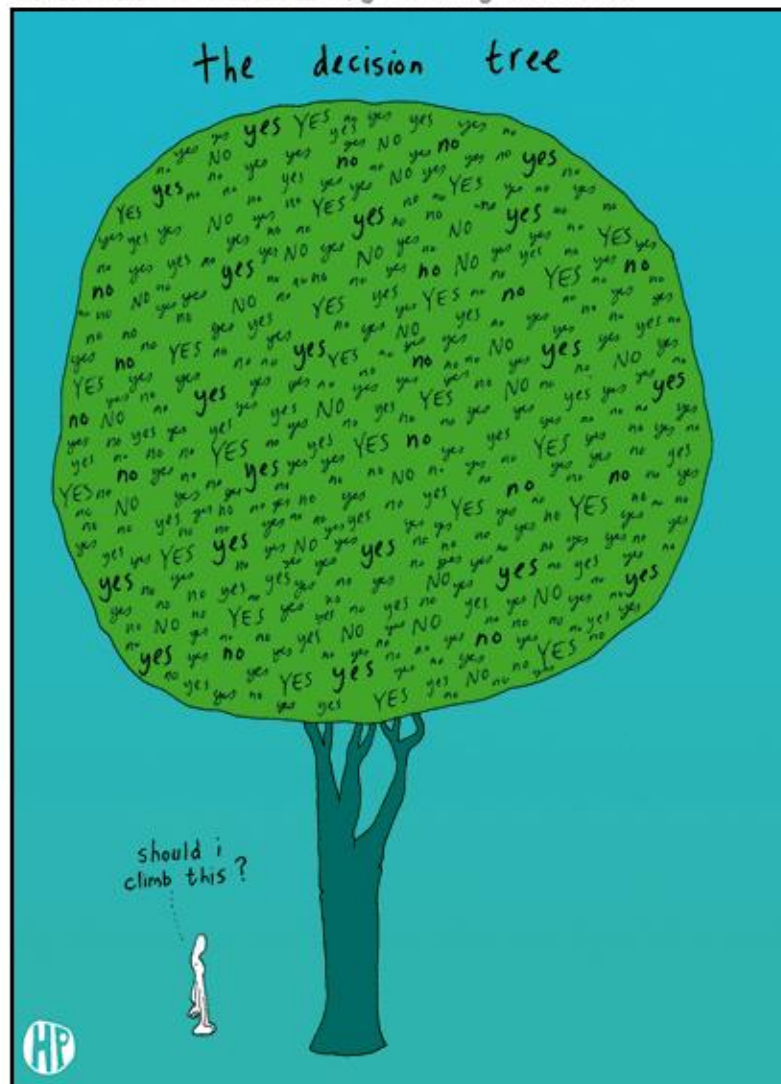
Modell (log. egyenlet) $\rightarrow p \rightarrow$ ha $p > \text{cut-off} \rightarrow \hat{Y}=1$

LOGISZTIKUS REGRESSZIÓ

Mutató	Jelentése	Jó érték	Teendő
Pseudo R^2 mutatók	A modell százalékosan hol helyezkedik el a két szélsőség, a null és a tökéletes modell között.	1-hez minél közelebb	
Likelihood-ratio test	Modell szignifikánsan eltér-e a konstans 0 modellettől (szignifikáns-e az illeszkedés)	0.05 alatt	
Wald statisztika	Az együtthatók szignifikánsan eltérnek-e 0-tól	0.05 alatt	A változó kivétele a modellből
B	Logisztikus regressziós egyenletben az együttható		
Exp(B)	A magyarázó változó oddsra gyakorolt parciális hatása		

DÖNTÉSI FA

HAROLD'S PLANET by Swerling and Lazar

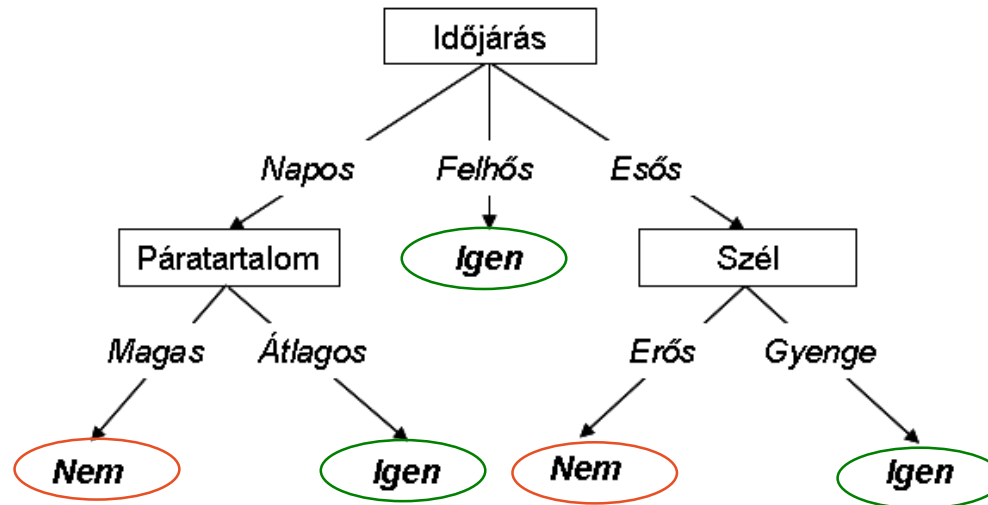


haroldspanet.com © 2006

DÖNTÉSI FA

Alapötlet: bonyolult összefüggéseket egyszerű döntések sorozatára vezet vissza

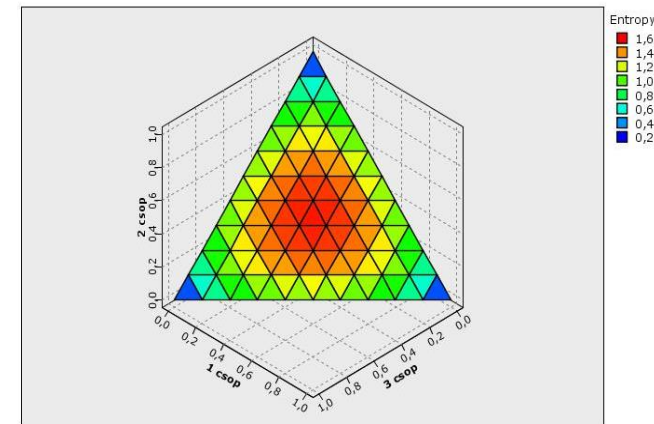
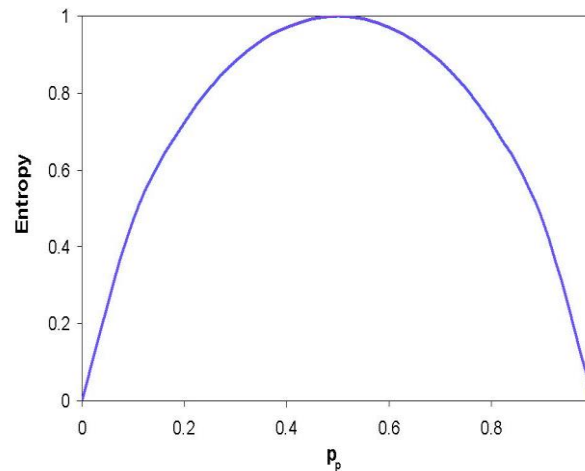
Példa: Alkalmas az időjárás teniszezésre?



- Minden csomópontban megvizsgálunk egy változót, X_i -t
- Minden csomópontból induló ág az X_i változó egy értékét jelöli
- Minden levél előrejelzi a célváltozó, Y értékét

DÖNTÉSI FA – VÁGÁSI SZABÁLYOK

- Minden csomópontban a rendelkezésre álló változók közül azt keressük, amellyel a tanulóhalmaz a célváltozóra nézve homogénebb csoportokat hoz létre, mint a vágás előtt.
- Vágási függvény például: Information gain (ratio) ~ entrópiacsökkenés (C5.0)
- **Entrópia:** a változó bizonytalanságát fejezi ki
$$H(Y) = - \sum_{i=1}^l p_i \cdot \log_2 p_i$$
 - ahol Y / különböző értéket p_i ($i=1\dots l$) valószínűséggel felvevő valószínűségi változó



DÖNTÉSI FA – INFORMATION GAIN

- **Information gain ~ entrópiacsökkenés:**
a vágással elérhető entrópiacsökkenés.
Megmutatja, hogy az X változó mennyi plusz információt hordoz Y -ról.
- A fát olyan változó mentén vágjuk tovább, amelynél az így elérhető entrópiacsökkenés maximális.

$$Gain(Y, X) = H(Y) - H(Y|X)$$

Entrópia a
vágás előtt

Entrópia a
vágás után

DÖNTÉSI FA PÉLDA: INFORMATION GAIN

Nap	Időjárás	Hőmérséklet	Páratartalom	Szél	Tenisz?
1	Napos	Meleg	Magas	Gyenge	Nem
2	Napos	Meleg	Magas	Erős	Nem
3	Felhős	Meleg	Magas	Gyenge	Igen
4	Esős	Enyhe	Magas	Gyenge	Igen
5	Esős	Hűvös	Átlagos	Gyenge	Igen
6	Esős	Hűvös	Átlagos	Erős	Nem
7	Felhős	Hűvös	Átlagos	Erős	Igen
8	Napos	Enyhe	Magas	Gyenge	Nem
9	Napos	Hűvös	Átlagos	Gyenge	Igen
10	Esős	Enyhe	Átlagos	Gyenge	Igen
11	Napos	Enyhe	Átlagos	Erős	Igen
12	Felhős	Enyhe	Magas	Erős	Igen
13	Felhős	Meleg	Átlagos	Gyenge	Igen
14	Esős	Enyhe	Magas	Erős	Nem

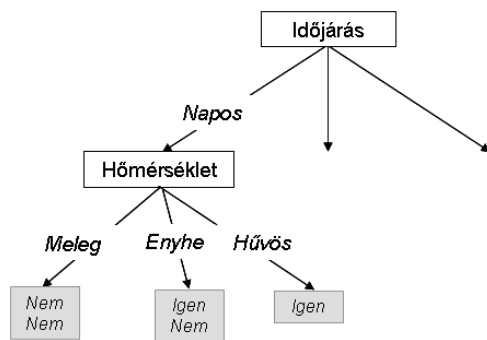
$$H_0 = H(9+, 5-) = -\frac{9}{14} \cdot \log_2 \frac{9}{14} - \frac{5}{14} \cdot \log_2 \frac{5}{14} = 0,940$$

$$Gain = H_0 - H_{idő} = 0,940 - \left(\frac{5}{14} \cdot H_{napos} + \frac{4}{14} \cdot H_{felhő} + \frac{5}{14} \cdot H_{esős} \right) = 0,247$$

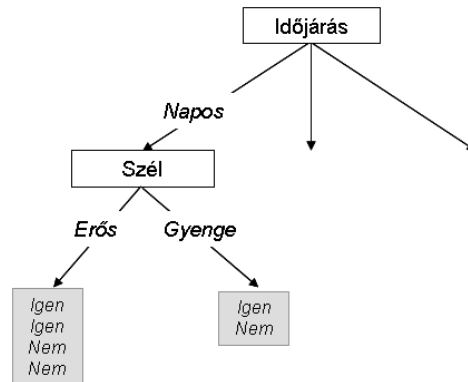
$$H(3+, 2-) = -\frac{3}{5} \cdot \log_2 \frac{3}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5} = 0,971$$

Változó	Kategória	Igen	Nem	Esetszám	Entrópia	Gain
Kiindulás		9	5	14	0,940	0,000
Időjárás	napos	2	3	5	0,971	0,247
	felhős	4	0	4	0,000	
	esős	3	2	5	0,971	
Hőmérséklet	meleg	2	2	4	1,000	0,029
	enyhe	4	2	6	0,918	
	hűvös	3	1	4	0,811	
Páratartalom	magas	3	4	7	0,985	0,152
	átlagos	6	1	7	0,592	
Szél	erős	6	2	8	0,811	0,048
	gyenge	3	3	6	1,000	

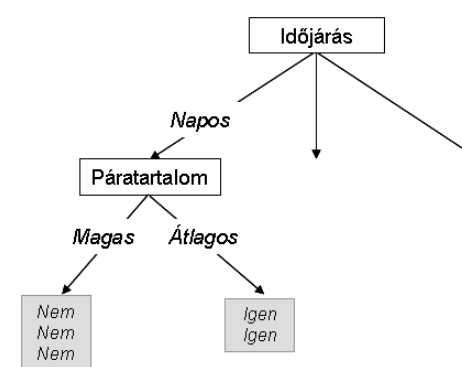
DÖNTÉSI FA PÉLDA



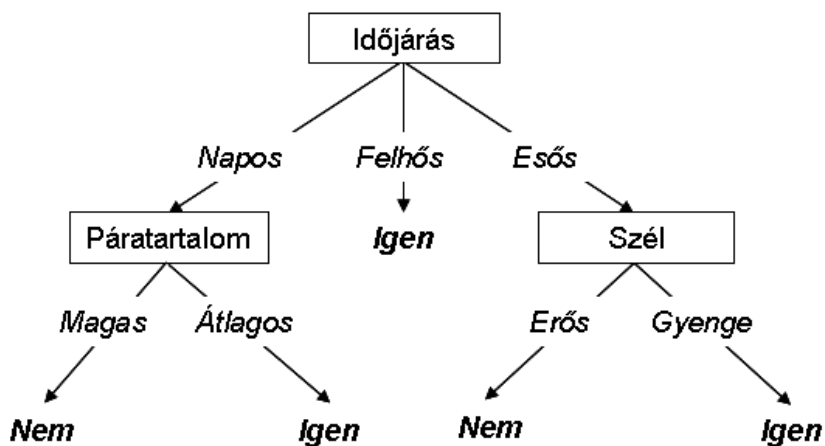
$Gain(Hőmérséklet)=0,571$



$Gain(Szél)=0,020$



$Gain(Páratartalom)=0,971$



DÖNTÉSI FA

Előny

- Könnyen értelmezhető, vizuális megjelenítés
- Automatikusan felismeri a lényeges változókat
- Nagy adatbázisokon is hatékonyan működik
- Misclassification cost használata
- Hiányzó értékek kezelése
- Extrém értékek nem okoznak gondot
- Interaktív mód – szakértői szabályok bevonása

Hátrány

- Túltanulás (de kezelhető)

KLASSZIFIKÁCIÓ OUTPUT

- \$L-célvált: **Előrejelzett érték** (folytonos célváltozó esetén csak ez)
- \$LP-célvált: **Konfidencia intervallum** (0.5 – 1 közötti érték)
- \$LRP-célvált: **Raw propensity score** (0-1 közötti érték) – flag célváltozó esetén

L helyett az adott modell betűjele (pl: N- NeuralNet, C-C5.0, B- Bayes stb)

KLASSZIFIKÁCIÓ - MODELLÉRTÉKELÉS

		Predicted class		
		Yes	No	
Actual class	Yes	TP: True positive	FN: False negative	← Elsőfajú hiba
	No	FP: False positive	TN: True negative	← Másodfajú hiba

Pontosság: helyes osztályozás aránya

$$\rightarrow \frac{TP + TN}{N}$$

Érzékenység ~ sensitivity (Recall): helyesen osztályozott pozitív minták aránya

$$\rightarrow \frac{TP}{TP + FN}$$

Sajátosság ~ specificity: helyesen osztályozott negatív minták aránya

$$\rightarrow \frac{TN}{TN + FP}$$

Megbízhatóság ~ precision: helyesen pozitív osztályba sorolt minták aránya

$$\rightarrow \frac{TP}{TP + FP}$$

KLASSZIFIKÁCIÓ - MODELLÉRTÉKELÉS

1. A modell alapján minden esethez egy pontszámot rendelünk (score érték)
2. A rekordokat a pontszám alapján csökkenő sorba rendezzük
3. A lista elején több találatot várunk

No	Score	Target	CustID	Age	
1	0.97	Y	1746	...	
2	0.95	N	1024	...	
3	0.94	Y	2478	...	
4	0.93	Y	3820	...	
5	0.92	N	4897	...	
...	...		...	...	
99	0.11	N	2734	...	
100	0.06	N	2422		

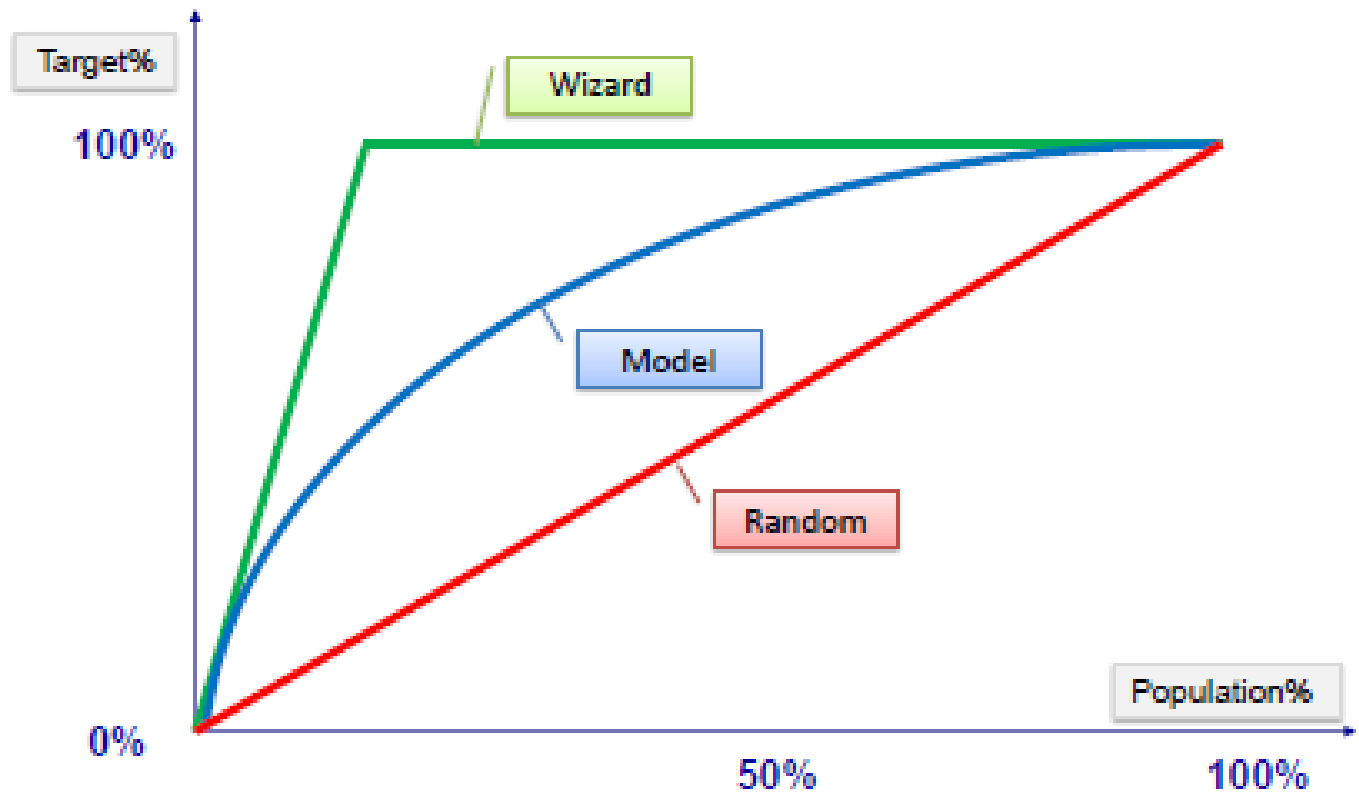
3 találat a lista első 5%-ában

Ha a mintában összesen 15 találat van, akkor a top 5%-kal a találatok 20%-át ($3/15=0,2$) találjuk meg.

KLASSZIFIKÁCIÓ – MODELLÉRTÉKEKELÉS – GAIN GÖRBE

**Gains ~ CPH ~
Cumulative % Hits**

$$\frac{\text{Nr of hits in top P\%}}{\text{Total Nr of hits}} \cdot 100$$

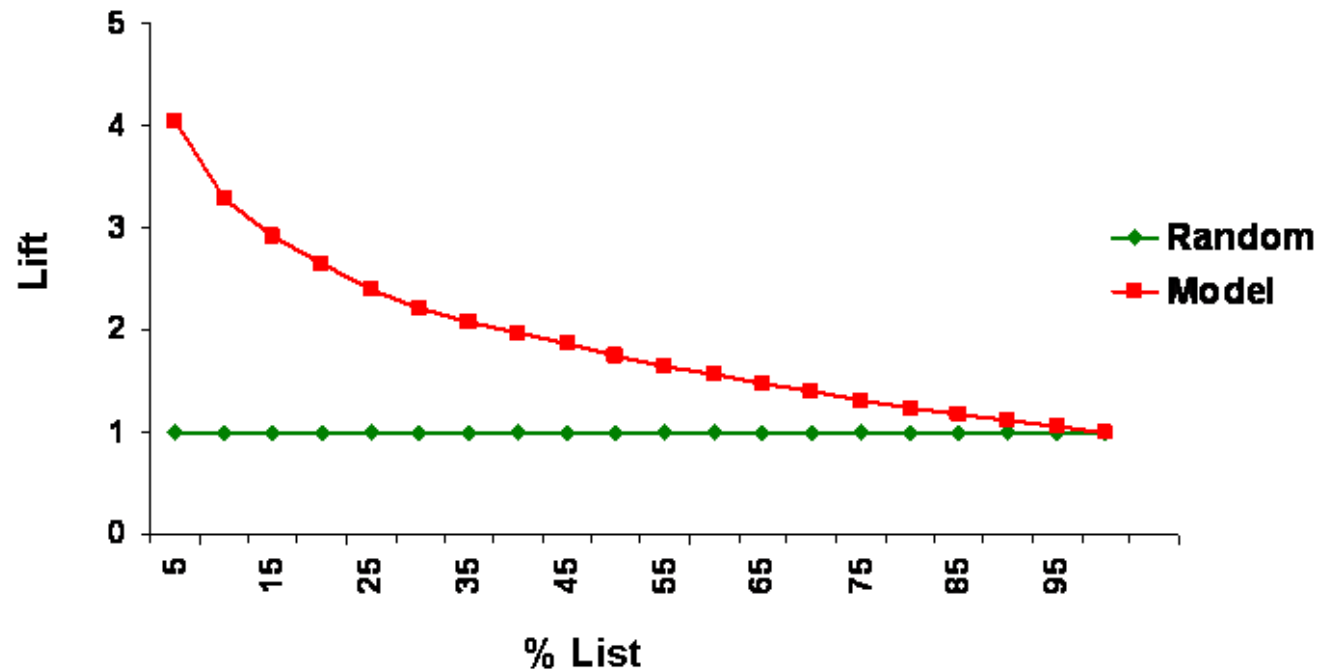


KLASSZIFIKÁCIÓ – MODELLÉRTÉKELEÉS – LIFT GÖRBE

Lift

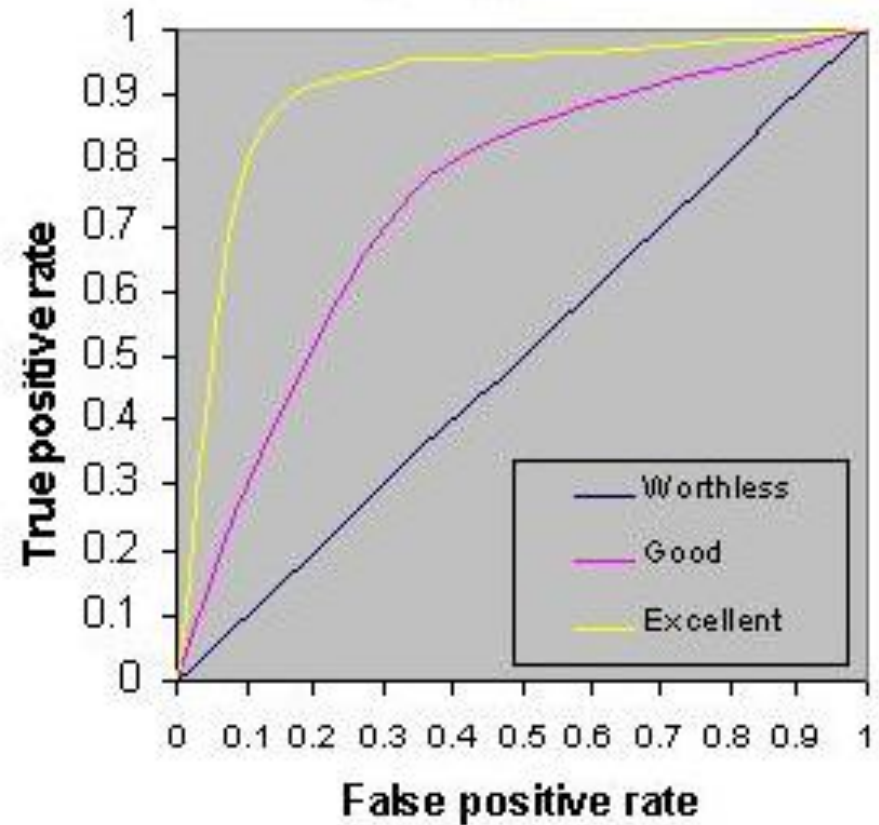
$$\frac{\text{Rate of hits in top } P\%}{\text{Total rate of hits}}$$

$$\text{Lift}(P\%) = \text{Gain}(P\%) / P$$



KLASSZIFIKÁCIÓ – MODELLÉRTÉKELÉS – ROC GÖRBE

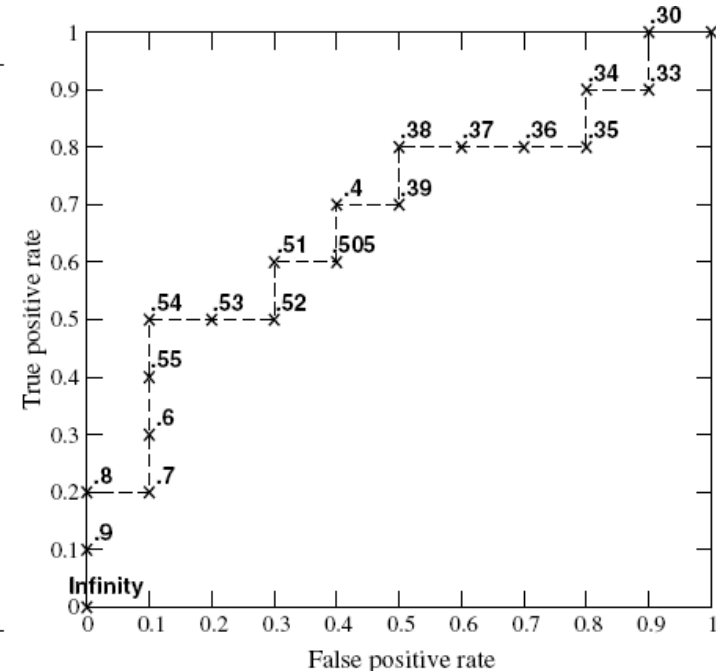
- Bináris célváltozó esetén a rangsorolás minőségének mérésére
- True positive / false positive
- Görbe alatti terület (AUC)~ modell jósága
 - Egy véletlen pozitív példa milyen valószínűséggel van előrébb a rangsorban, mint egy véletlen negatív példa.
- Véletlen modell: 0.5
- Tökéletes modell: 1
- Jó: 0.8 fölött



KLASSZIFIKÁCIÓ – MODELLÉRTÉKELES – ROC GÖRBE

- A tesztadatokat pontérték szerint sorba rendezzük (→ rangsor)
- Operációs küszöb (cut-off): vágási érték, amely feletti pontértékkal rendelkező adat az 1-es (pozitív) osztályhoz tartozik, alatta pedig a negatívba.
- Minden küszöbértéknek megfelel egy TPR (y-tengely) és egy FPR (x-tengely) érték.
- A küszöböt a max-tól a min-ig mozgatva megkapjuk az összes lehetséges TP/FP pontot. Ezen pontokat összekötő görbe a ROC görbe.

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



KLASSZIFIKÁCIÓ – MODELLÉRTÉKELÉS

1. Pontosság ~ accuracy
Helyes osztályozások aránya $\frac{TP + TN}{N}$

2. Találati mátrix
TP és TN arány ellenőrzése külön is

		Predicted class	
		Yes	No
Actual class	Yes	TP	FN
	No	FP	TN

3. Túltanulás ellenőrzése

Egy modell túltanul, ha rátanul a tanító minta sajátosságaira, így nem lesz általánosan alkalmazható.

Jelei:

- A modell pontossága a tesztelő mintán jelentősen rosszabb, mint a tanító mintán
- A tesztelő mintán a kiértékelő görbe (pl.: Gain-görbe, ROC-görbe) jelentősen rosszabb, mint a tanító mintán

4. Összehasonlítás a többi modellel
5. Legjobb modell kiválasztása

A tanított modelleket a fenti szempontok szerint sorban kiértékeljük, és minden lépésnél eldöntjük, hogy az adott modell megfelel-e az elvárásoknak, tovább léphetünk-e a következő szempontra.

CÉLVÁLTOZÓ ELOSZLÁSA

- Pl.: célváltozó eloszlása: 98% N, 2% Y
- Ekkor egy konstans N modell 98%-os pontosságú, viszont használhatatlan

!!! Nem elég a pontosságot nézni, nézzük meg a találati mátrixot!!!

- Megoldás:
 - Nem döntési fa modellek: minta átsúlyozása (Balance) → torzítást eredményez
 - Döntési fák esetén: **misclassification cost**

Kérdések?

Molnár-Edőcs Eszter
eedocs@clementine.hu