Machine Learning

Agenda

Általános bevezető

- Fő típusok
- Üzleti problémák
- Módszertan

Adatelőkészítés

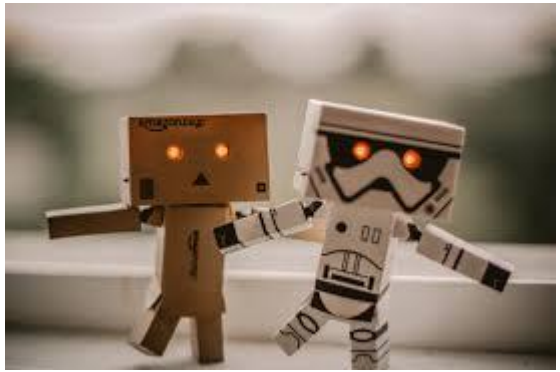
- Adattranszformációk
- Feature selection
- Tesztkörnyezet kialakítása

Klasszifikáció kiértékelése

- Accuracy
- ROC (AUC)

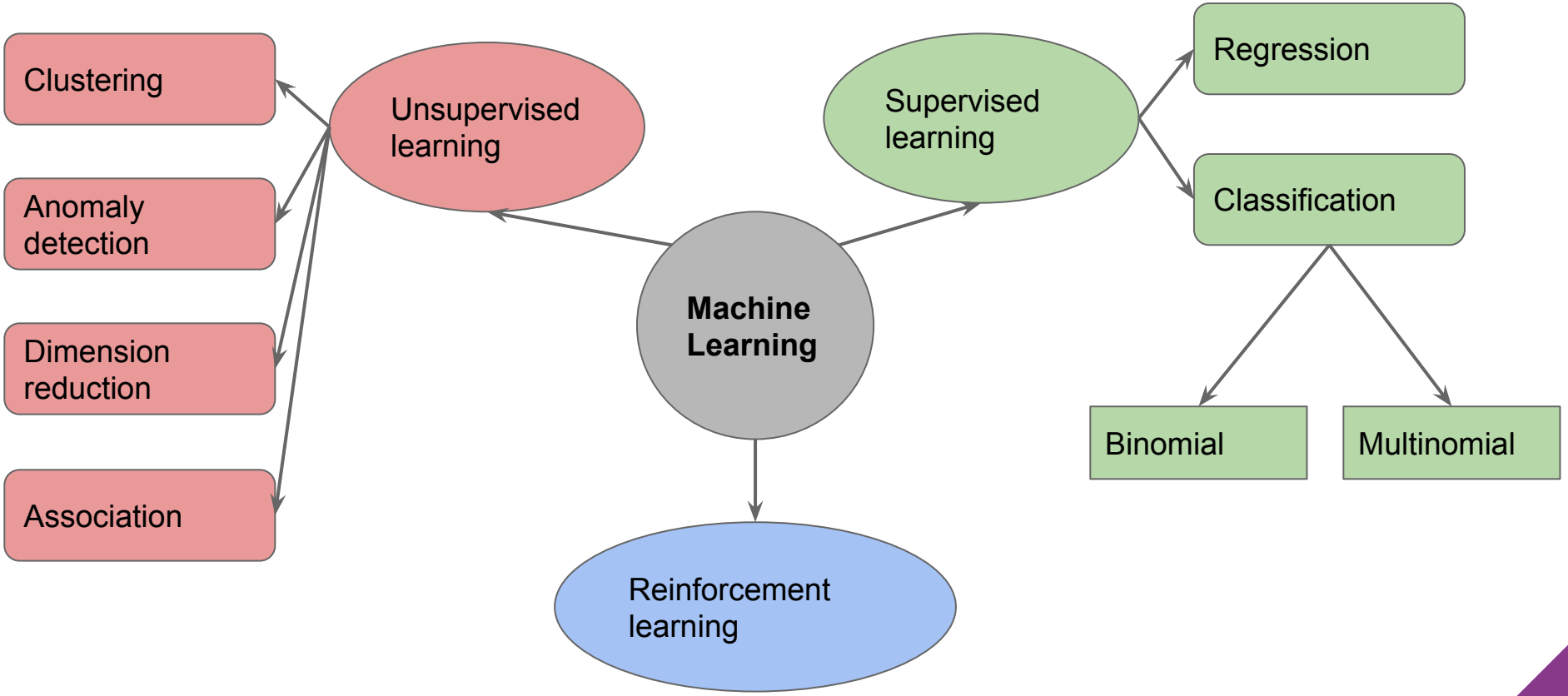
Machine learning

- előrejelzés
- összefüggések feltárása
- automatizálás



forrás:
https://www.pexels.com/search/machine%20learning/

Típusok


Clustering
Anomaly detection
Dimension reduction
Association
Unsupervised learning
Machine Learning
Supervised learning
Regression
Classification
Binomial
Multinomial
Reinforcement learning

Reinforcement Learning

- AlphaGo
- robotok mozgása

Supervised Learning

Regression
(numerikus célváltozó)

- Árak előrejelzése
- Kereslet előrejelzése
- Ügyfélérték becslés
- Várható élettartam
- Idősor előrejelzés

Classification
(kategorikus célváltozó)

- esemény bekövetkezése
  - hitel bírálat
  - lemorzsolódás
- keresztértékesítés

Supervised Learning Algoritmusok

Regression

- Lineáris regresszió
- Neurális hálózat
- Döntési fa / Random forest
- stb.

Classification

- Logisztikus regresszió
- Döntési fa / random forest
- Neurális hálózat
- Support Vektor Machine
- Naive Bayes
- Bayes hálózat
- K-nn
- stb

Supervised learning vs unsupervised learning

- Adatok fel vannak címkézve (célváltozó)
- Cél: címke minél pontosabb előrejelzése
- $f(x_i)=y_i+\mathcal{E}$
- Training - teszt partíció

- Nincsen címke
- Cél: mintázatok keresése az adatokban
- osztályozzuk x_i-ket
- Csak training adat van

Unsupervised Learning

Dimenzió csökkentés

- Összefüggések feltárása
- Zaj csökkentése az adatban
- Futásidő csökkentése

Klaszterezés

- Ügyfélcsoportok azonosítása
- Termékek csoportosítása

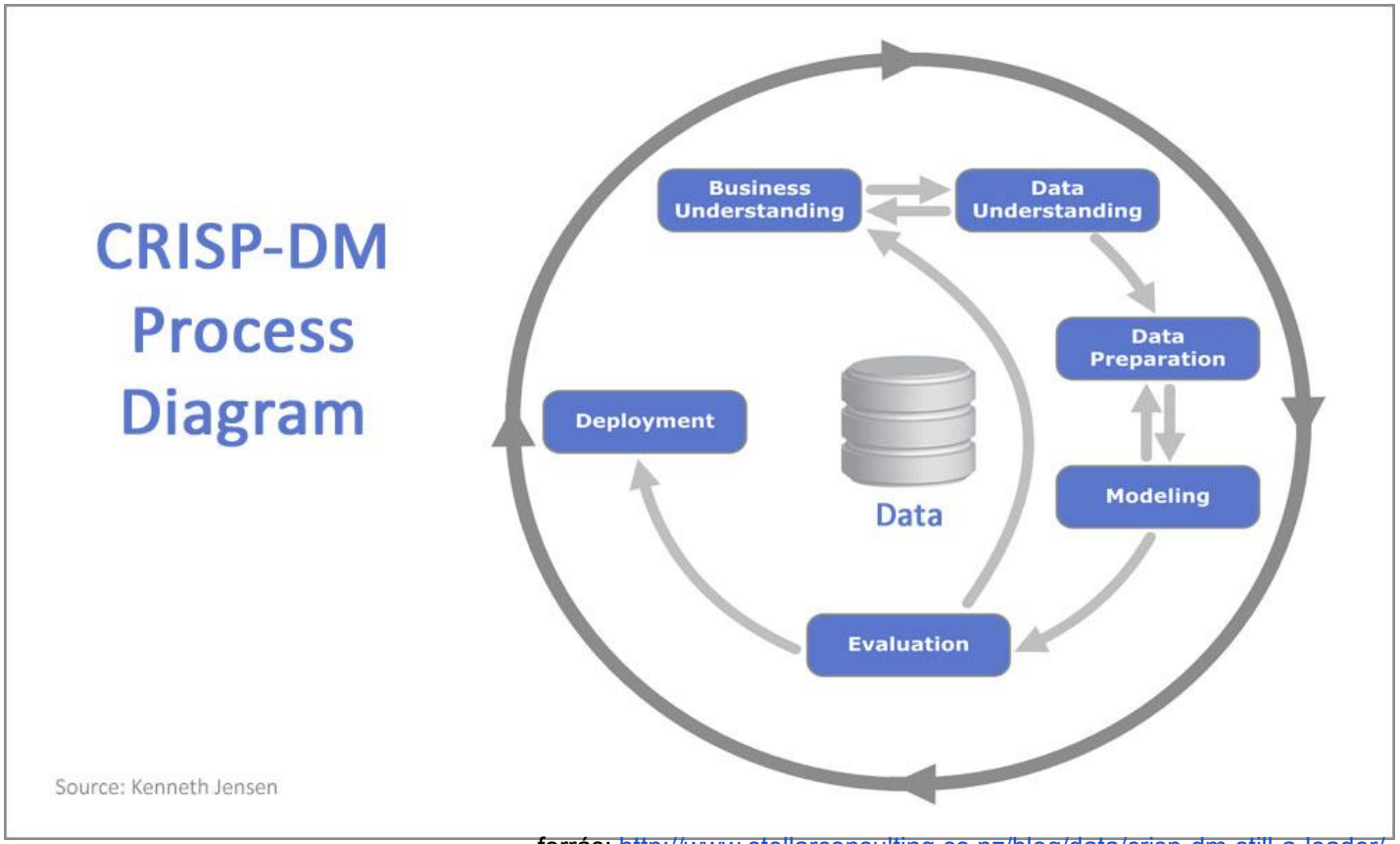
Anomália detekció

- Csalás detektálás
- Hibakeresés

Asszociációs szabályok

- Vásárlói kosárelemzés
- Keresztértékesítés

Módszertan



forrás: http://www.stellarconsulting.co.nz/blog/data/crisp-dm-still-a-leader/

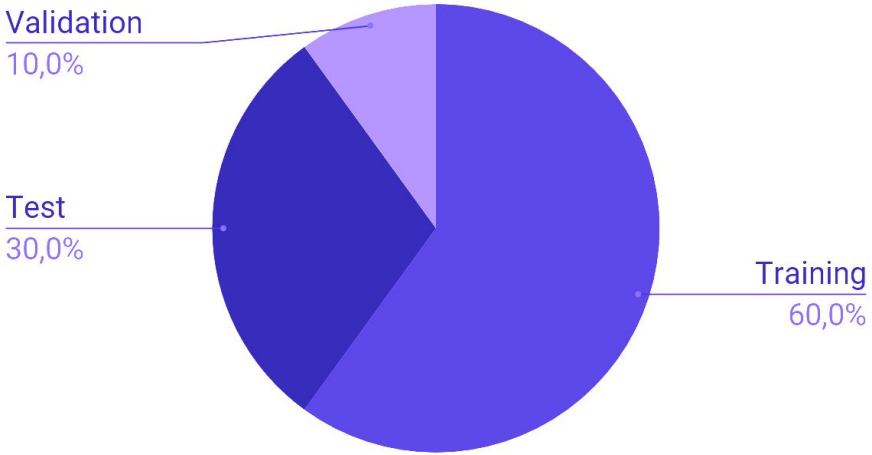
Adatok előkészítése

Adatok előkészítése és megismerése

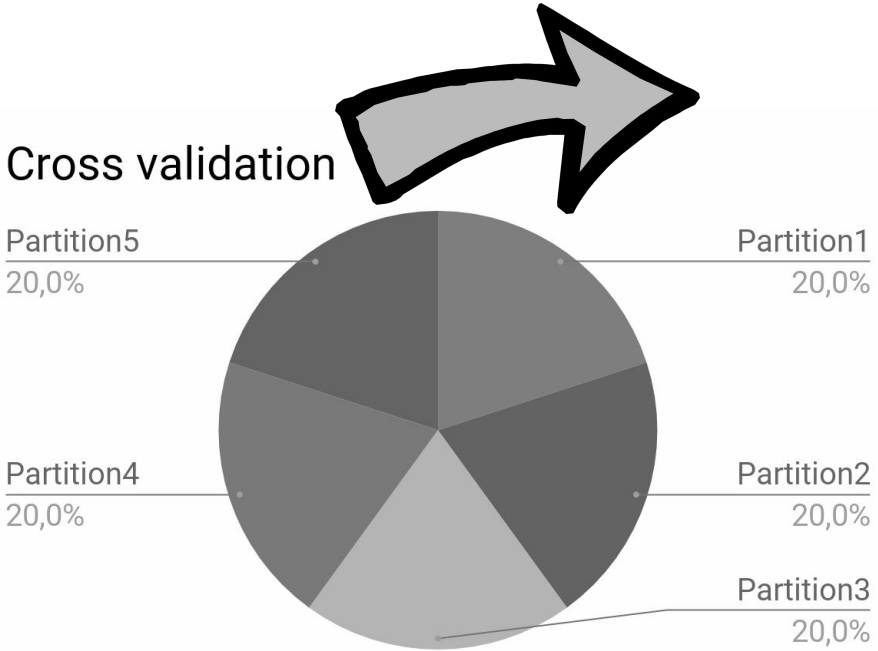
- Filterezés - rekordok kiválogatása
- Hiányzó értékek kezelése
- Kiugró értékek kezelése
- Leíró statisztikák
- Vizualizáció
- Változók létrehozása
- Feature selection
- Training - teszt (- validáló) adatok leválogatása vagy cross-validation környezet kialakítása

Tesztkörnyezet kialakítása

Training-test-validation



Cross validation



Feature selection

Miért fontos?

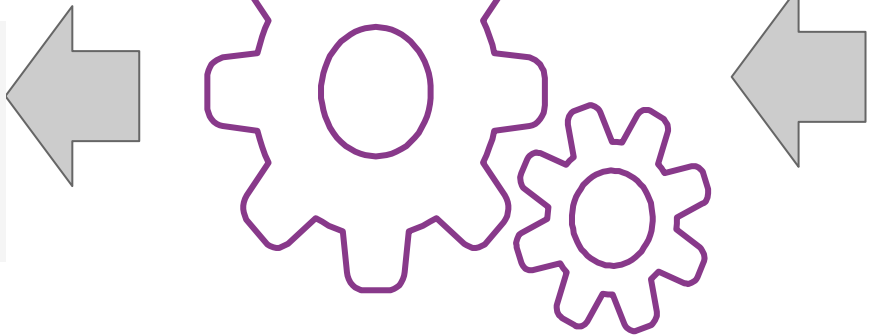
- zaj csökkentése
- túltanulás elkerülése
- gyorsabb modellépítés

- Célváltozótól független változók kiszűrése
- Összefüggő bemenő változók kezelése

Modell építése

model <- randomForest(Target ~ ., data=df)

algoritmus



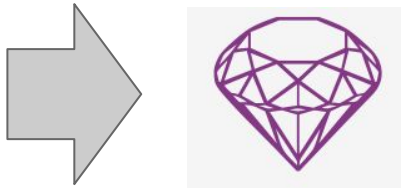
df (tanító adat)

Age	Salary	Ed	Target
21	405	M	**1**
58	587	H	**0**
42	100	L	**1**
19	256	M	**0**
33	800	H	**0**
...	...	...	**...**

Modell alkalmazása

```
predict(model, df_test, type="prob")
```

Age	Salary	Ed
28	455	H

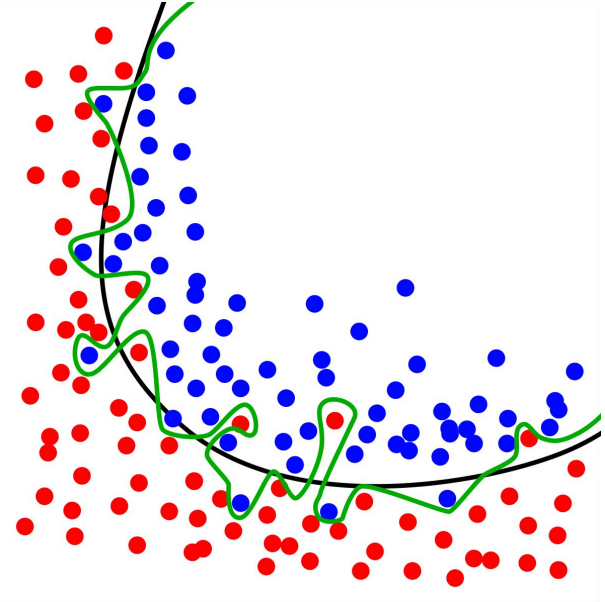


0.86

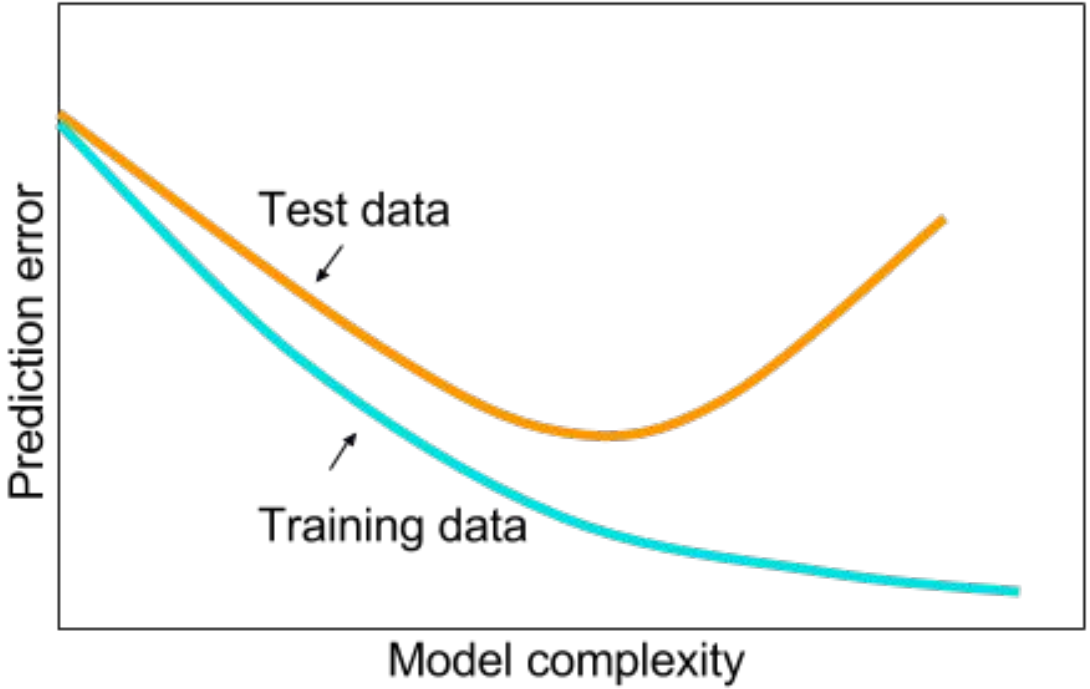
Modellek kiértékelése

1. Teszt adaton a modell alkalmazása (scoreolás)
2. A score-ok összehasonlítása a címkével
3. A legjobb modell kiválasztása a teszt adatokon
4. A kiválasztott modell kiértékelése a validáló adatokon

Overfitting (Túltanulás)



forrás: https://en.wikipedia.org/wiki/Overfitting



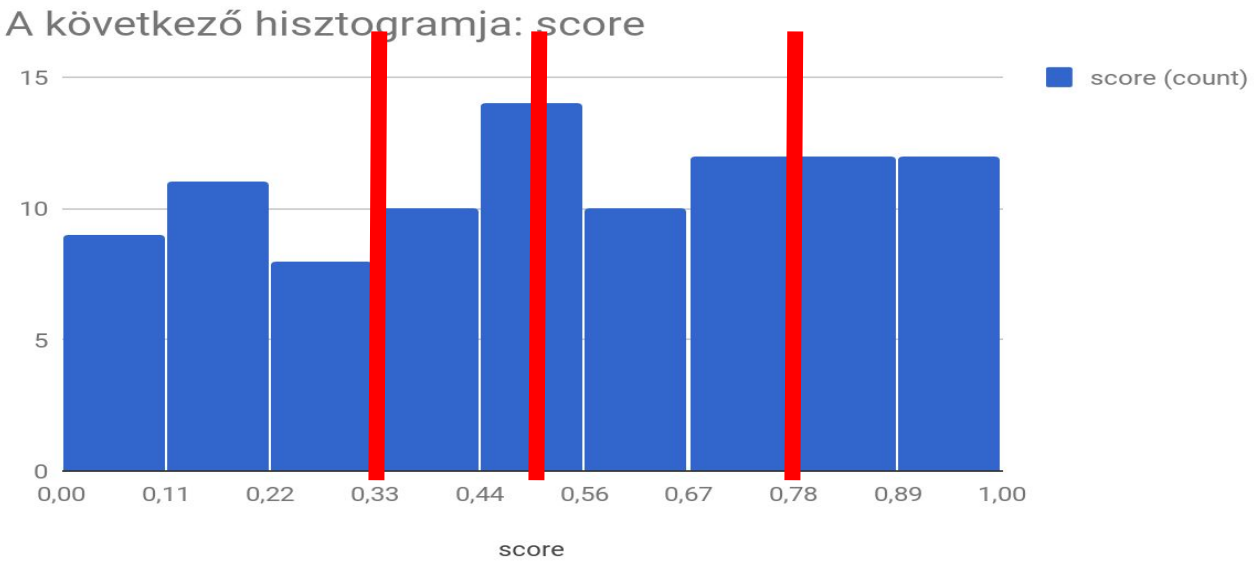
forrás: http://gluon.mxnet.io/chapter02_supervised-learning/regularization-scratch.html

Bináris klasszifikáció kiértékelése

Accuracy vs. ROC (AUC)

Cut-off érték

type="response" vs. type="prob”



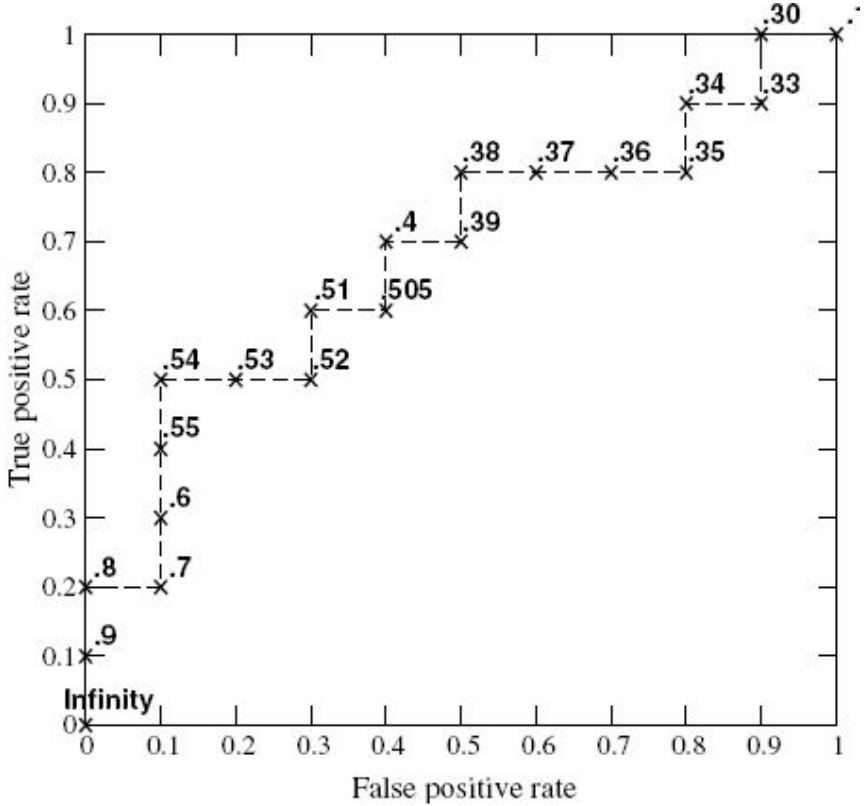
Pontosság - accuracy

		Előrejelzett kategória	
		No	**Yes**
Valós kategória	**No**	True Negative	False Positive
	Yes	False Negative	True Positive

(TN+TP)/(TN+FP+FN+TP)

ROC chart

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



forrás: https://ccrma.stanford.edu/workshops/mir2009/references/ROCintro.pdf

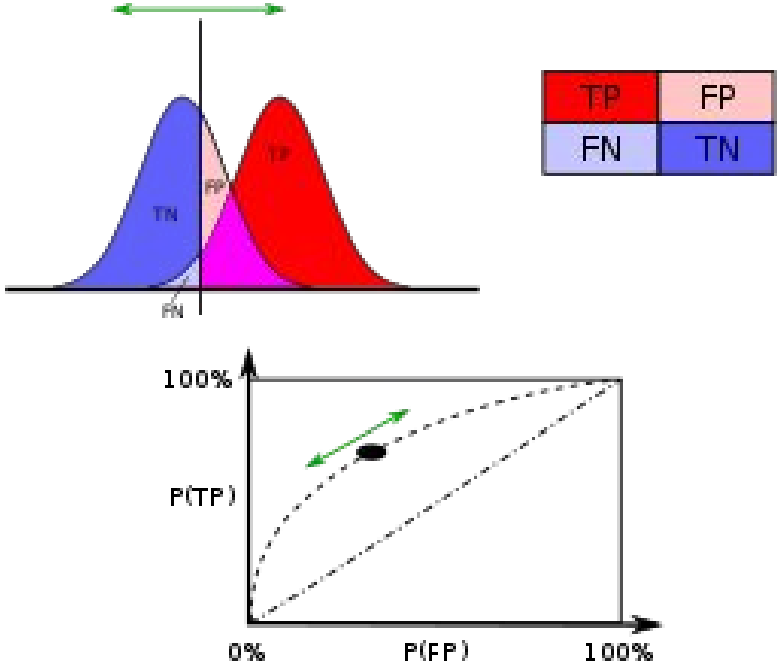
ROC chart

AUC: ROC görbe alatti terület

AUC ~ 0.5 Random score
AUC ~ 0.7 Gyengén szeparáló modell
AUC ~ 0.9 Jó szeparáció
AUC ~ 1 Tökéletes modell

library(pROC)

plot(roc(df$label, df$score),print.auc=TRUE)



forrás: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

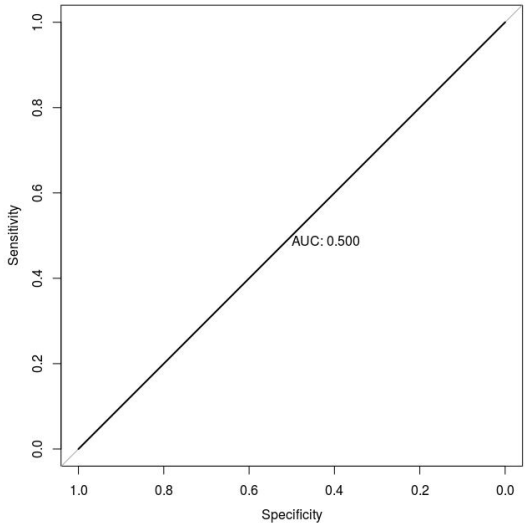
Példa - 1. modell

Beteg	Score	Előrejelzett kategória
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Nem	0.1	NEM
Igen	0.1	NEM

		Előrejelzett kategória	
		NEM	**IGEN**
Valós kategória	**Nem**	9	0
	Igen	1	0

Accuracy = 9/10 = **90%**

AUC = **0.5**



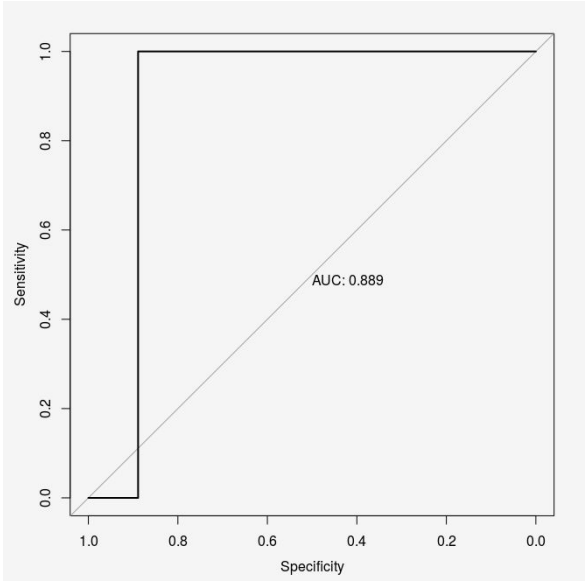
Példa - 2. modell

Beteg	Score	Előrejelzett kategória
Nem	0.1	NEM
Nem	0.2	NEM
Nem	0.2	NEM
Nem	0.3	NEM
Nem	0.3	NEM
Nem	0.4	NEM
Nem	0.4	NEM
Nem	0.6	IGEN
Igen	0.8	IGEN
Nem	0.9	IGEN

		Előrejelzett kategória	
		NEM	**IGEN**
Valós kategória	**Nem**	7	2
	Igen	0	1

Accuracy = 8/10 = **80%**

AUC = **0.889**



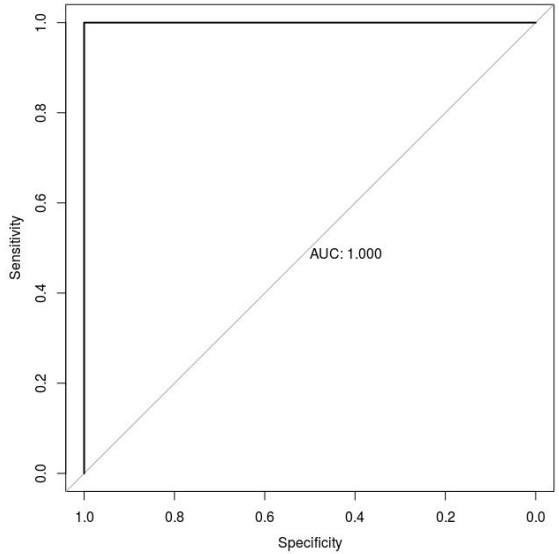
Példa - 3. modell

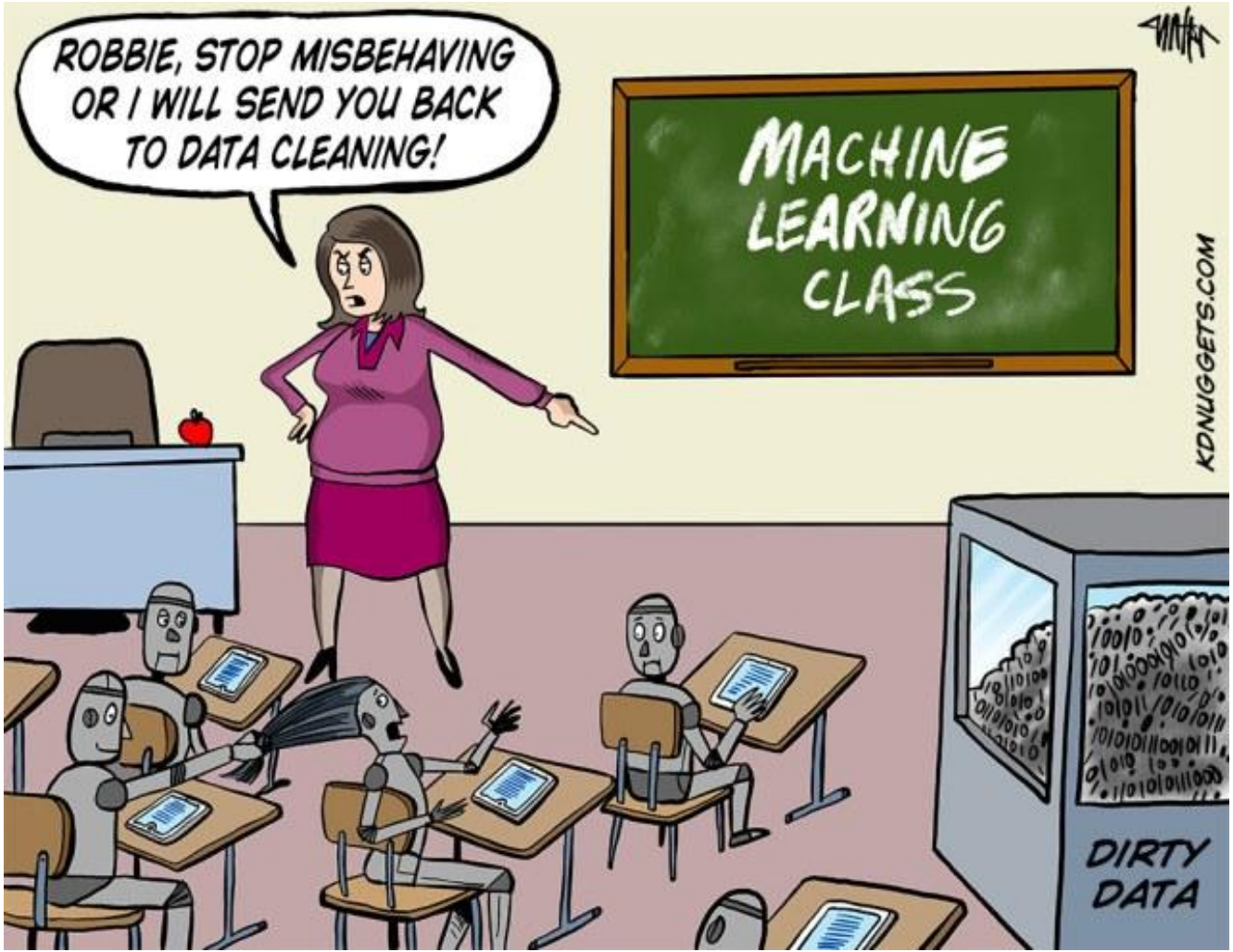
Beteg	Score	Előrejelzett kategória
Nem	0.52	IGEN
Nem	0.53	IGEN
Nem	0.54	IGEN
Nem	0.55	IGEN
Nem	0.55	IGEN
Nem	0.6	IGEN
Nem	0.6	IGEN
Nem	0.7	IGEN
Nem	0.7	IGEN
Igen	0.9	IGEN

		Előrejelzett kategória	
		NEM	**IGEN**
Valós kategória	**Nem**	0	9
	Igen	0	1

Accuracy = 1/10 = **10%**

AUC = **1.0**





forrás: https://www.kdnuggets.com/2017/09/cartoon-machine-learning-class.html