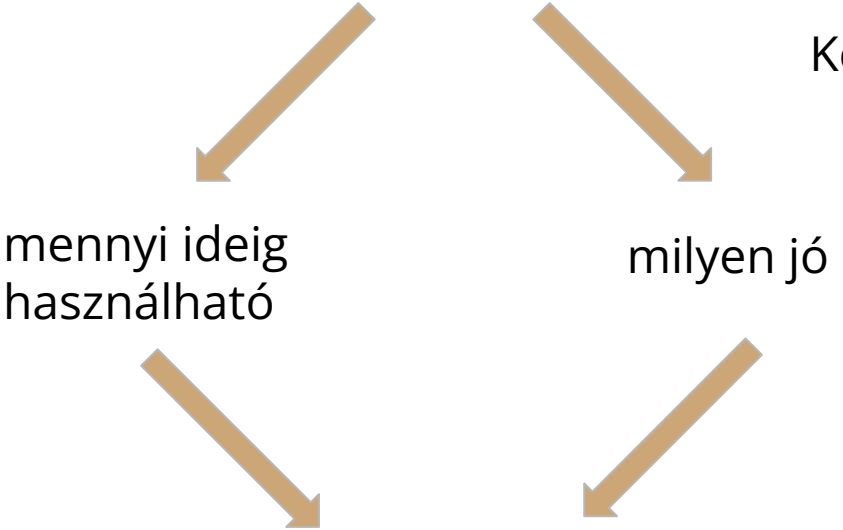
# Használt autó adatok elemzése

Balázs Frigyes, Nagy Balázs

# Elméleti keret

Használt autók értéke → jövőbeli hasznosság jelenértéke

Kooreman és Haan, 2006




tökéletes piac hiánya
(információs asszimetria)

# Változók

Kor

Megtett km-k száma

Teljesítmény

Üzemanyag

Kivitel

Állapot

Tömeg

# Adatok bemutatása

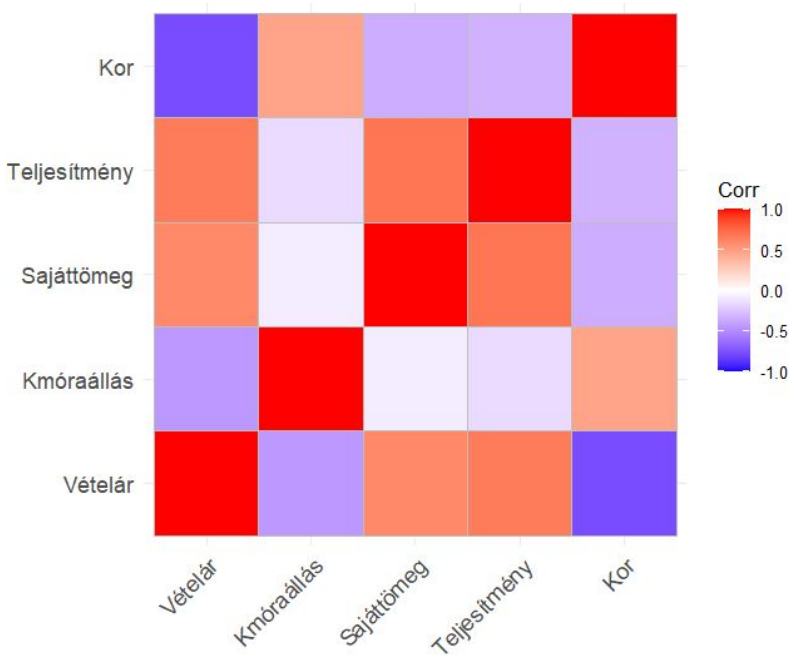
Március végén scrapelt adatok, nagyjából 89 ezer

Szűrés után ~65 ezer

Outlierek kiszűrése

Esetenként adatok kisebb részét használtuk (futási idő, memória limit)

# Változók korrelációja



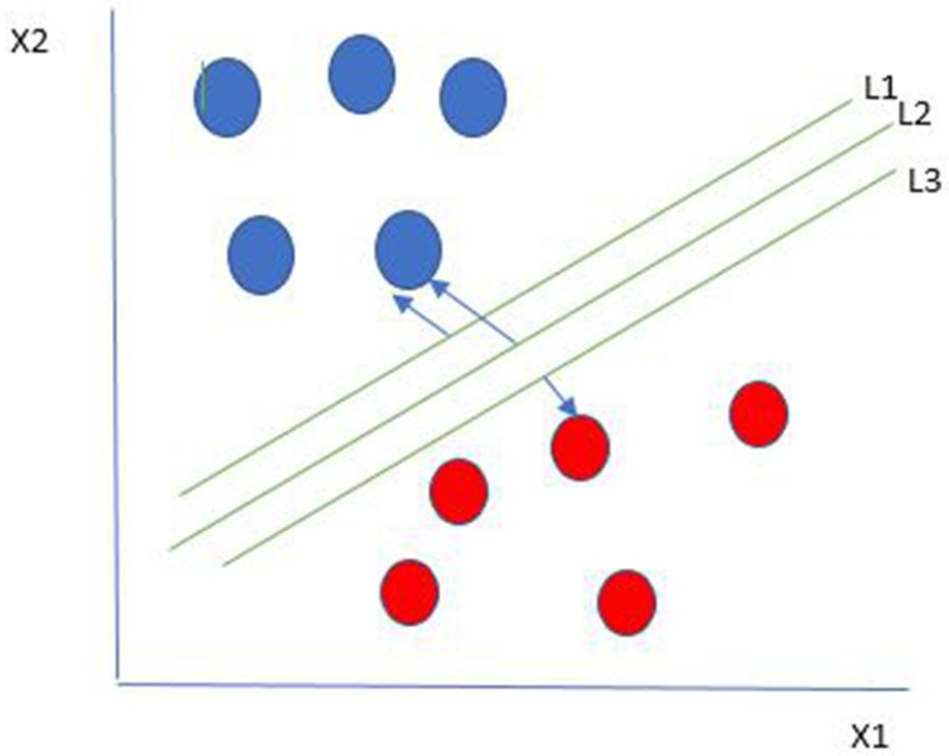
# Adatok forrása

- hasznaltautok.hu
- Március végi scrapelés
- Vpn, rvest, Bash használata

Kihívások:

- Rate limiting
- IP banning
- Adathibák

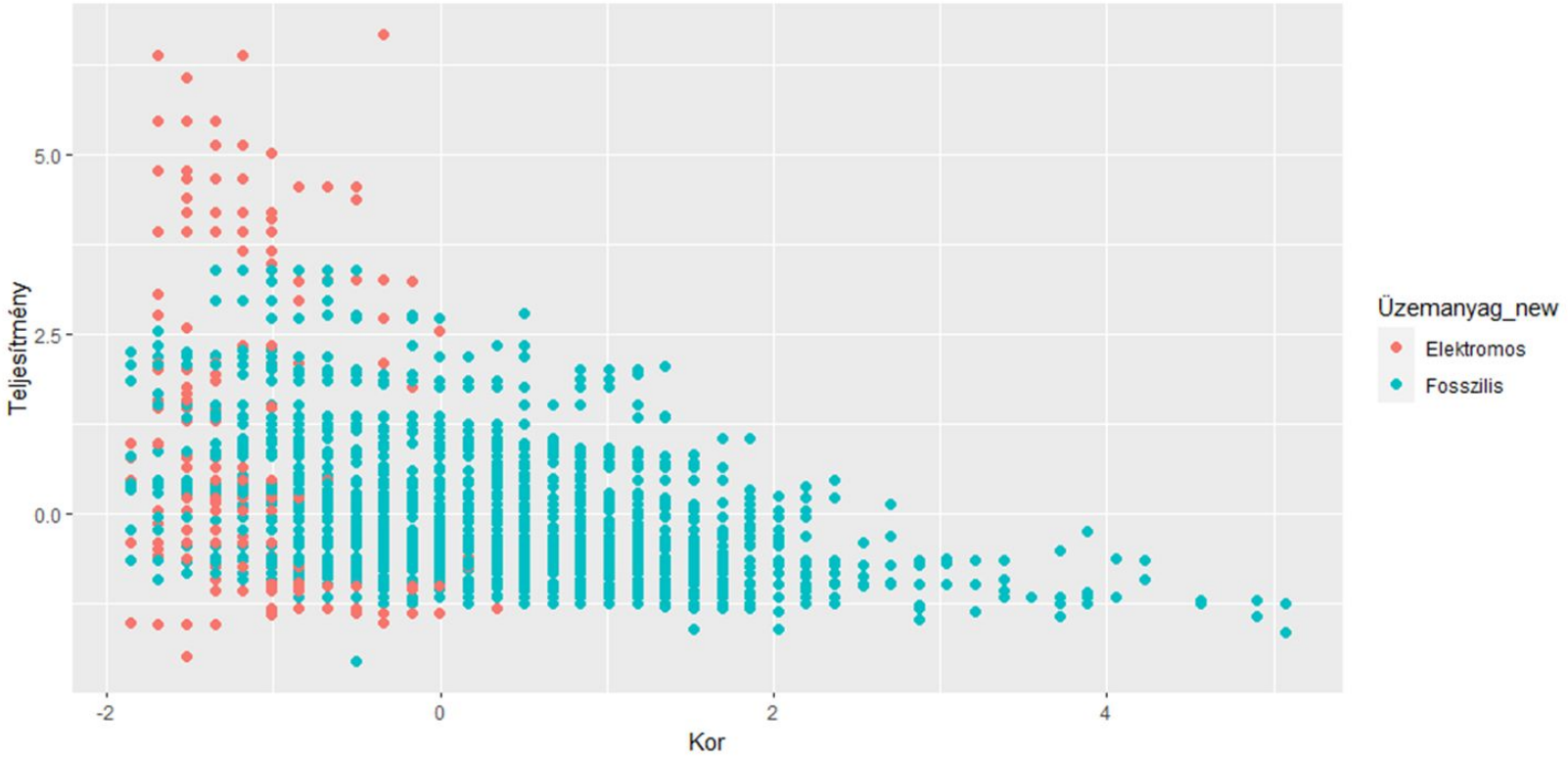
# SVM



# Magyarázott változó: üzemanyag

- SVM-el klasszifikációt végeztünk
- Eredeti adatok hiányosak
- Adatokban több üzemanyagtípus
- Leszűkítettük 2-re
  - Elektromos
  - Fosszilis (gáz, benzin, dízel)
- Ami egyéb azt kidobtuk
- Magyarázó változók: Kilométeróra állás, Kor, Teljesítmény, Saját tömeg

# Kiinduló adatok



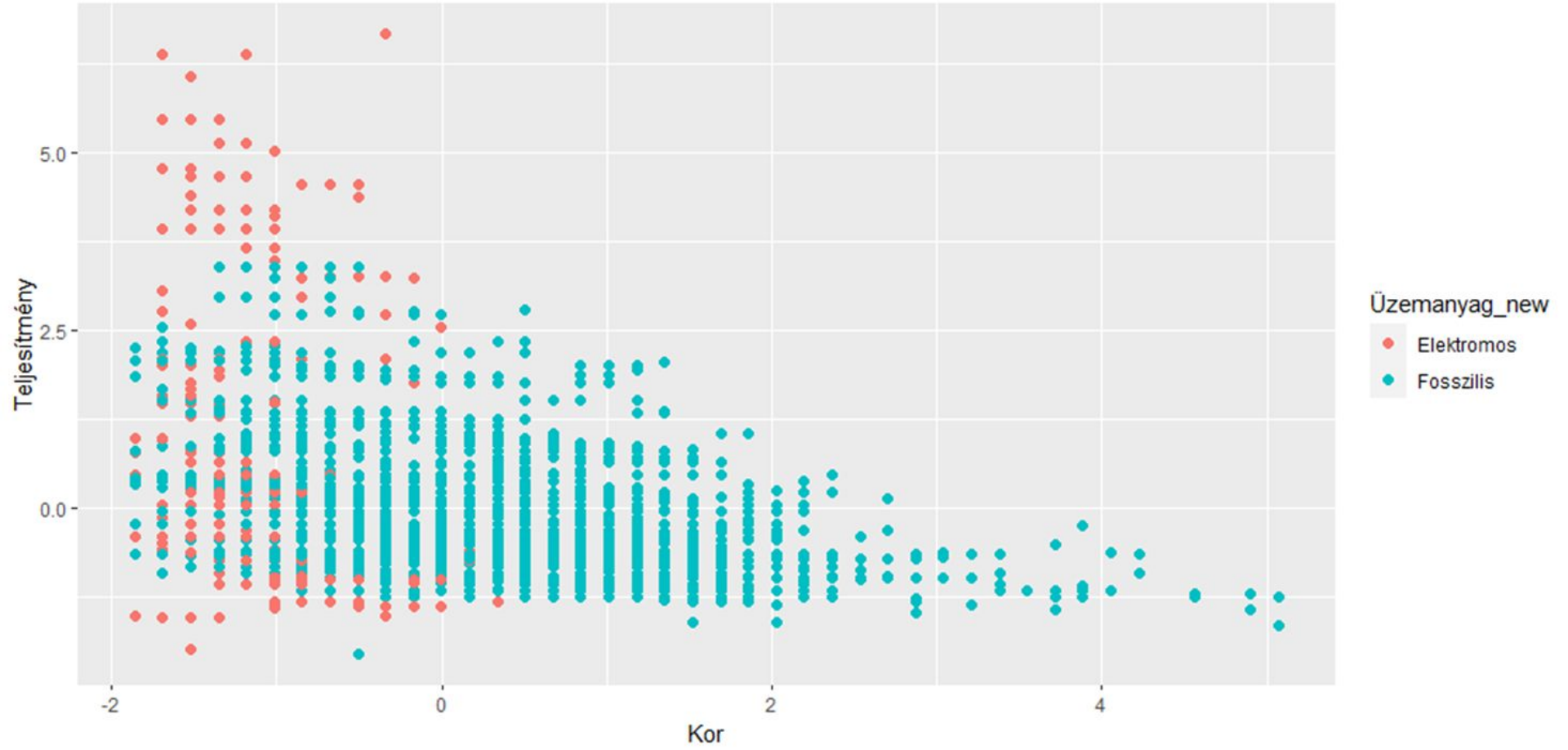
# SVM modellek

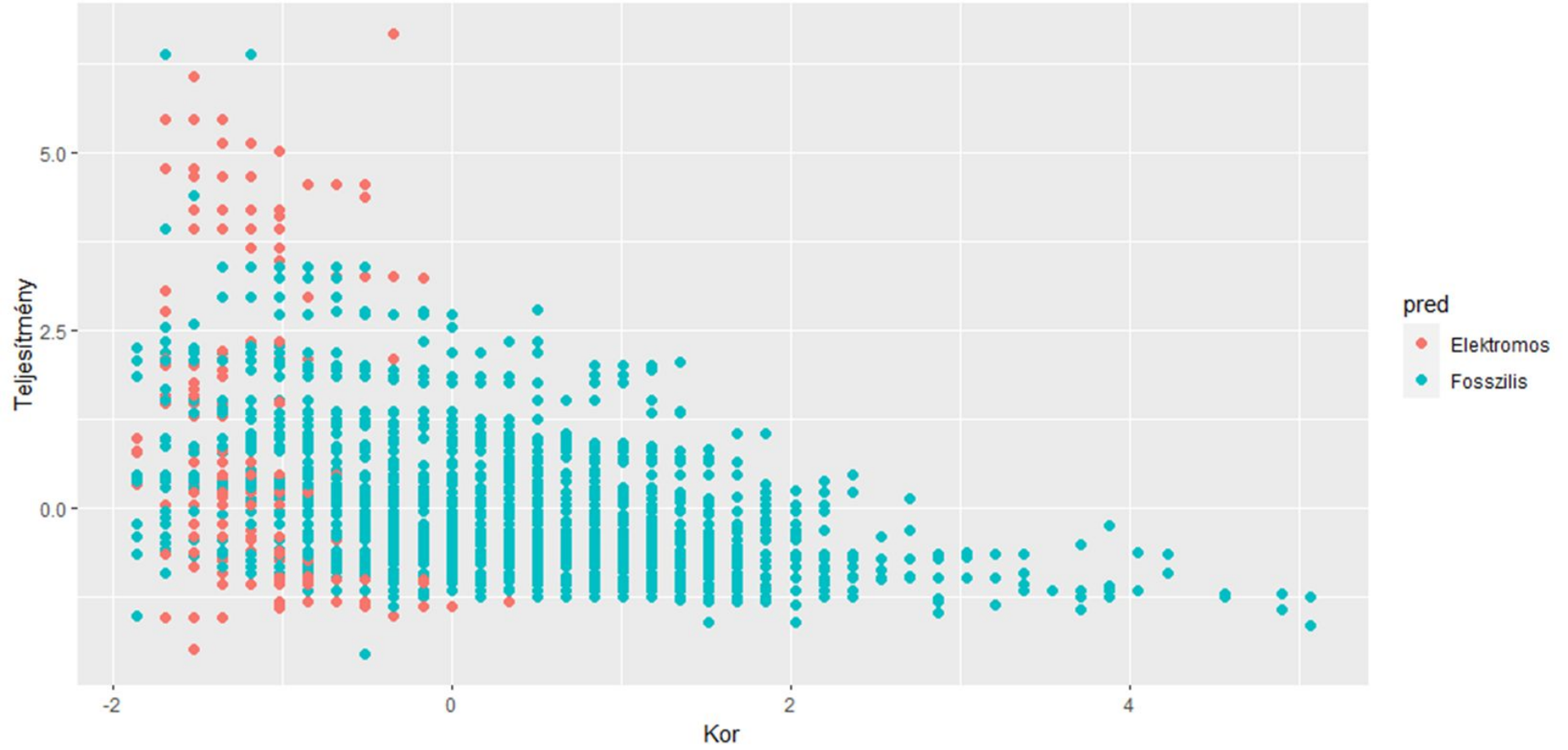
- Kipróbált specifikációk
- Lineáris
- Polinomiális
  - Több fokszámmal
- Radial
  - „határmegszegések büntetése” keresztvalidációval hangolva
  - Gamma paraméter keresztvalidációval hangolva
- Standardizált és nem standardizált adatokra

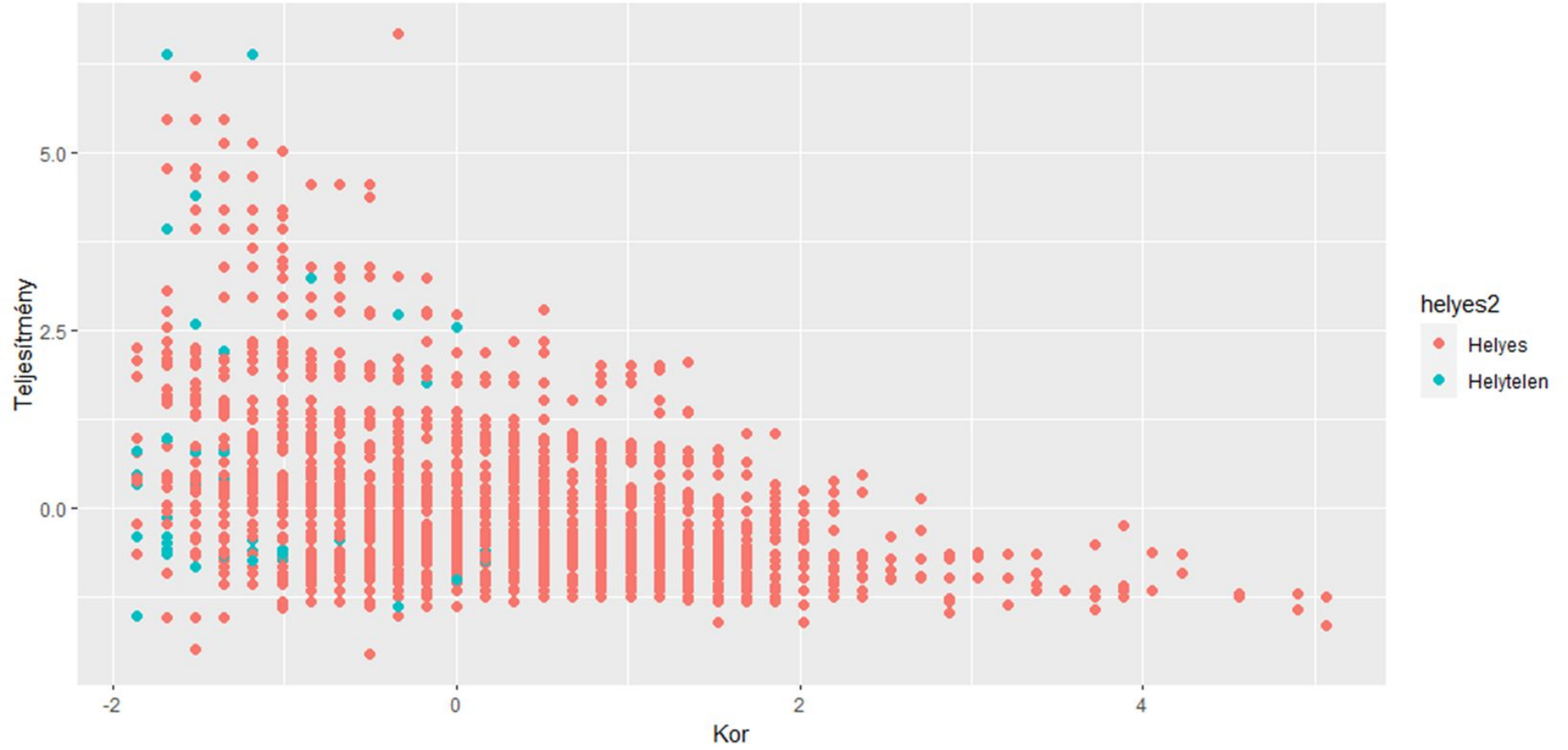
# Végleges modell:

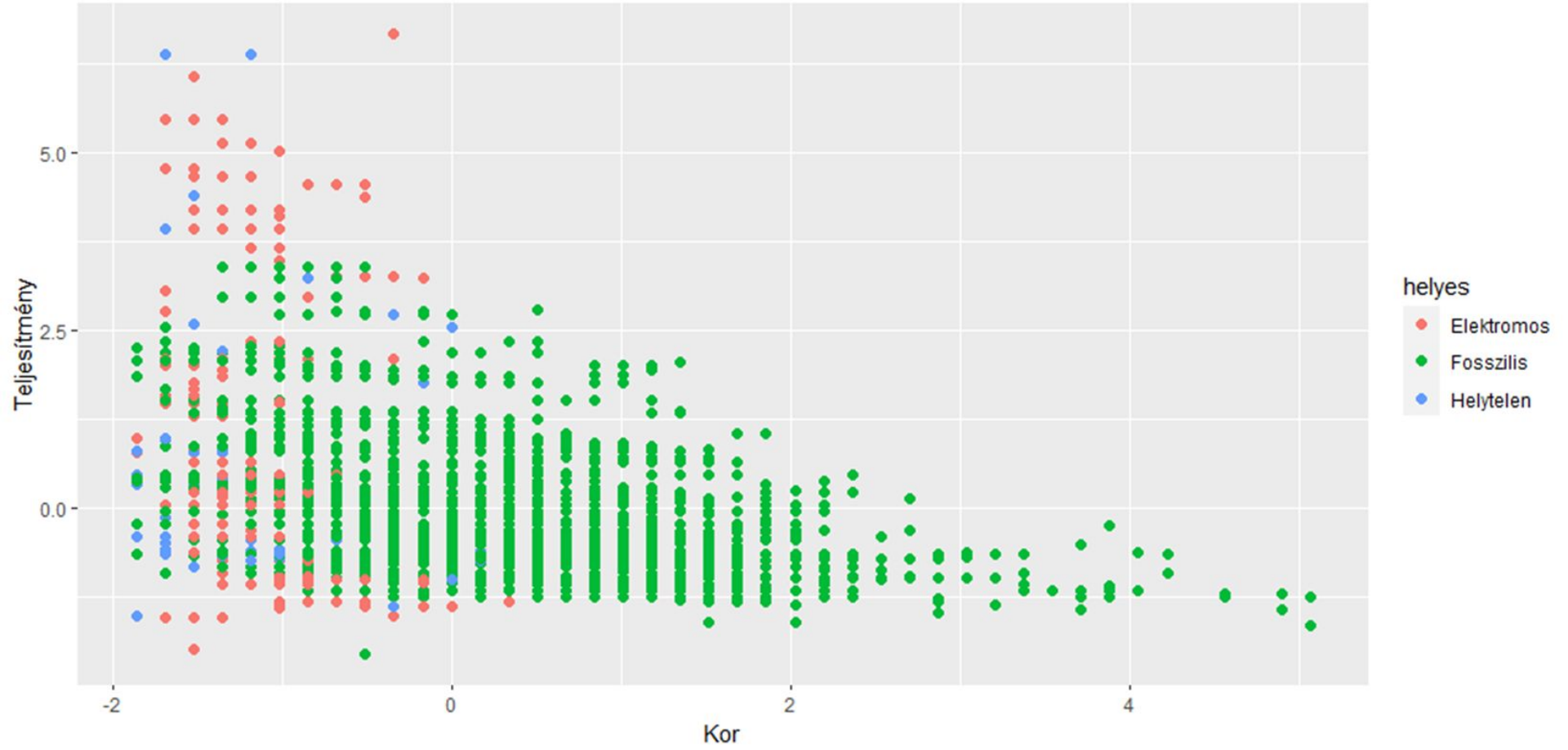
- 96,9%-os pontosság (nullmodell 82%-os)
- radial

| | Elektromos | Fosszilis |
|---|---|---|
| Elektromos | 407 | 52 |
| Fosszilis | 98 | 4347 |


Teljesítmény
5.0
2.5
0.0
-2
0
2
4
Kor
Üzemanyag_new
Elektromos
Fosszilis


Teljesítmény
5.0
2.5
0.0
-2
0
2
4
Kor
pred
Elektromos
Fosszilis


Teljesítmény
5.0
2.5
0.0
-2
0
2
4
Kor
helyes2
Helyes
Helytelen

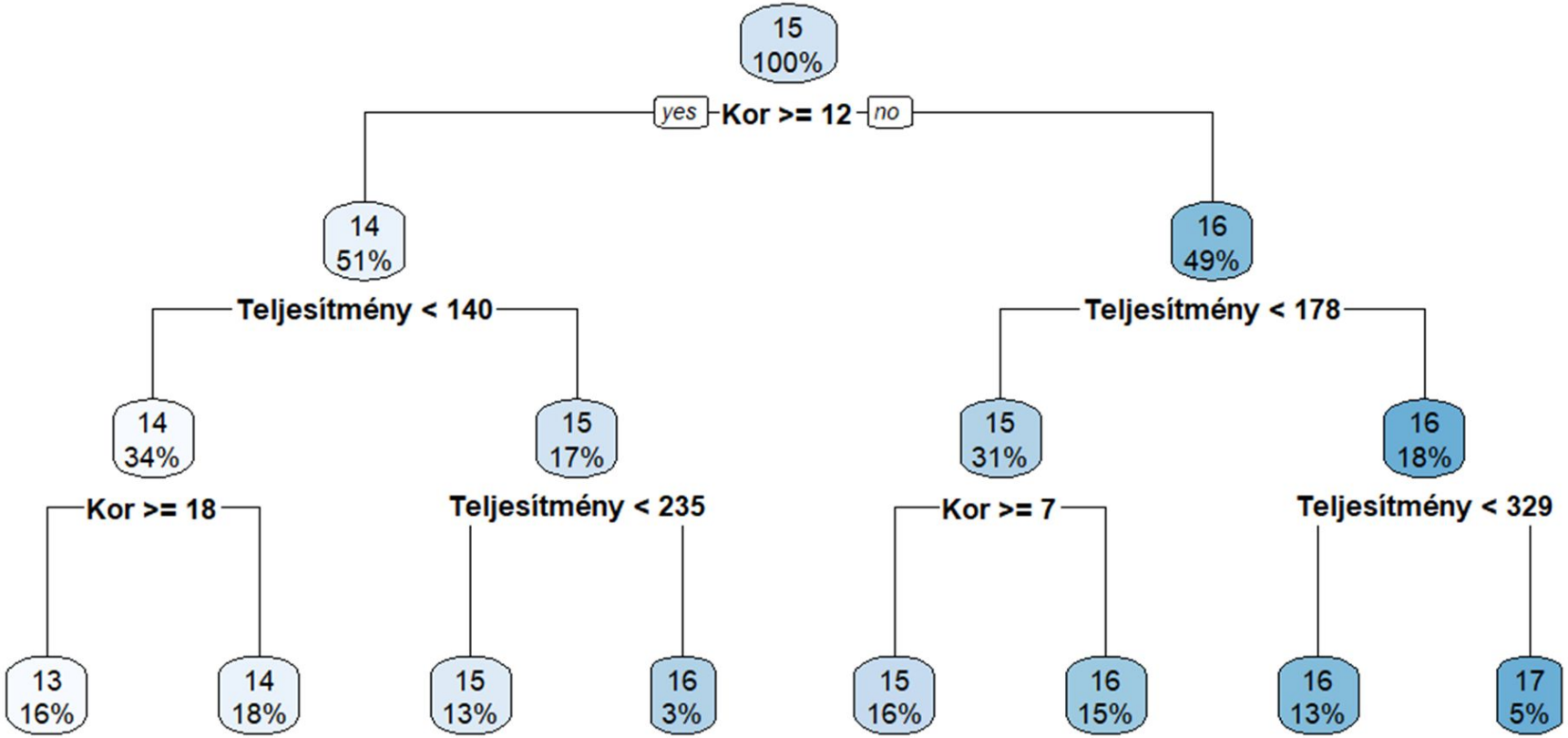

Teljesítmény
5.0
2.5
0.0
-2
0
2
4
Kor
helyes
Elektromos
Fosszilis
Helytelen

# Fa alapú modellek

döntési fa, véletlen erdő, xgboost

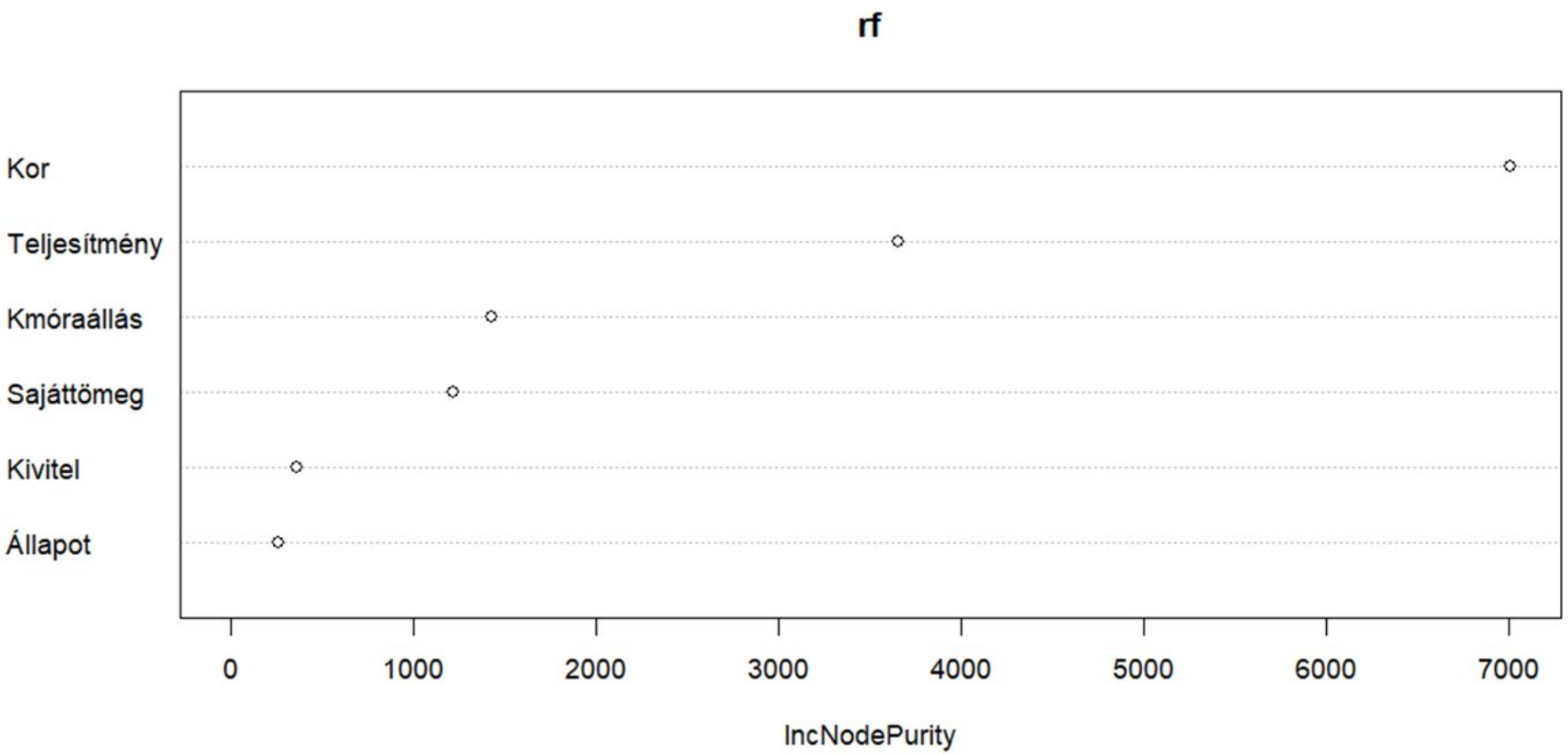
# Döntési fa

- Ár logaritmusát magyarázza
- Magyarázó változók: Állapot, Kivitel, Km óra állás, Saját tömeg, Teljesítmény, Kor
- Paraméter hangolással optimalizált
- 0,80 R^2 a tesztadaton
- Kor a legfontosabb


15
100%
yes
Kor >= 12
no
14
51%
16
49%
Teljesítmény < 140
Teljesítmény < 178
14
34%
15
17%
15
31%
16
18%
Kor >= 18
Teljesítmény < 235
Kor >= 7
Teljesítmény < 329
13
16%
14
18%
15
13%
16
3%
15
16%
16
15%
16
13%
17
5%

# Véletlen erdő

- Hosszú futási idő (csökkentett adatmennyiség)
- Limitált hiperparaméter hangolás
- Azonos magyarázó változók
- 0,93 R^2 a tesztadaton


rf
Kor
Teljesítmény
Kmóraállás
Sajáttömeg
Kivitel
Állapot
0
1000
2000
3000
4000
5000
6000
7000
IncNodePurity

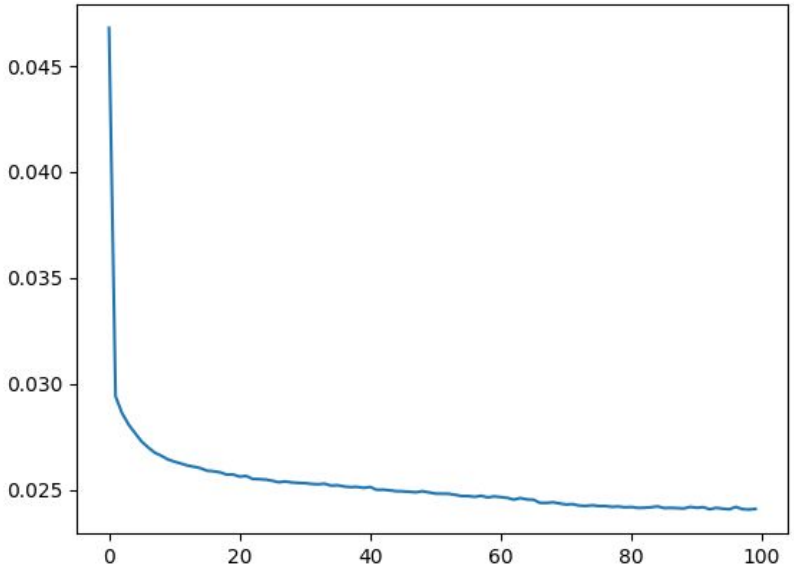
# XGBoost

- Azonos magyarázó változók
- Hiper paraméterek hangolása
  - Eta-tanulási ráta
  - Maxdepth
- 0,93 R^2 a tesztadaton
- Gyors futás

# Neurális hálók

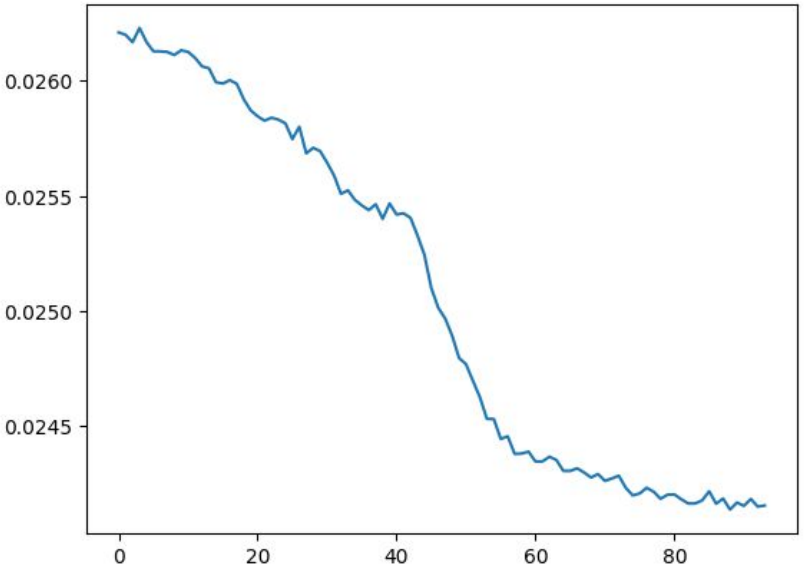
1x hidden layer ReLu, output sigmoid

R2: 0,910, MAE: 0,0242




2x hidden layer ReLu, output sigmoid

R2: 0,913, MAE: 0,0267



# Bónusz: Kolmogorov–Arnold Networks

Május elején publikált, Kolmogorov-Arnold reprezentációs tételen alapszik

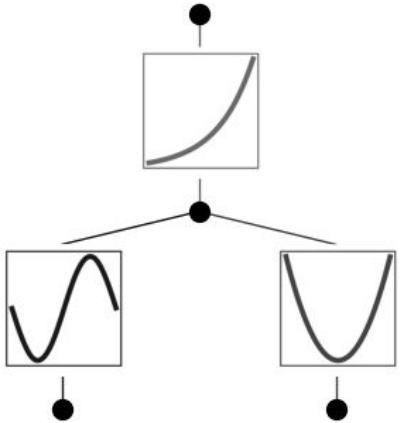
Komplex, több szintű képletek becslésére

Éleken tanítható aktivációs függvények (b-spline)

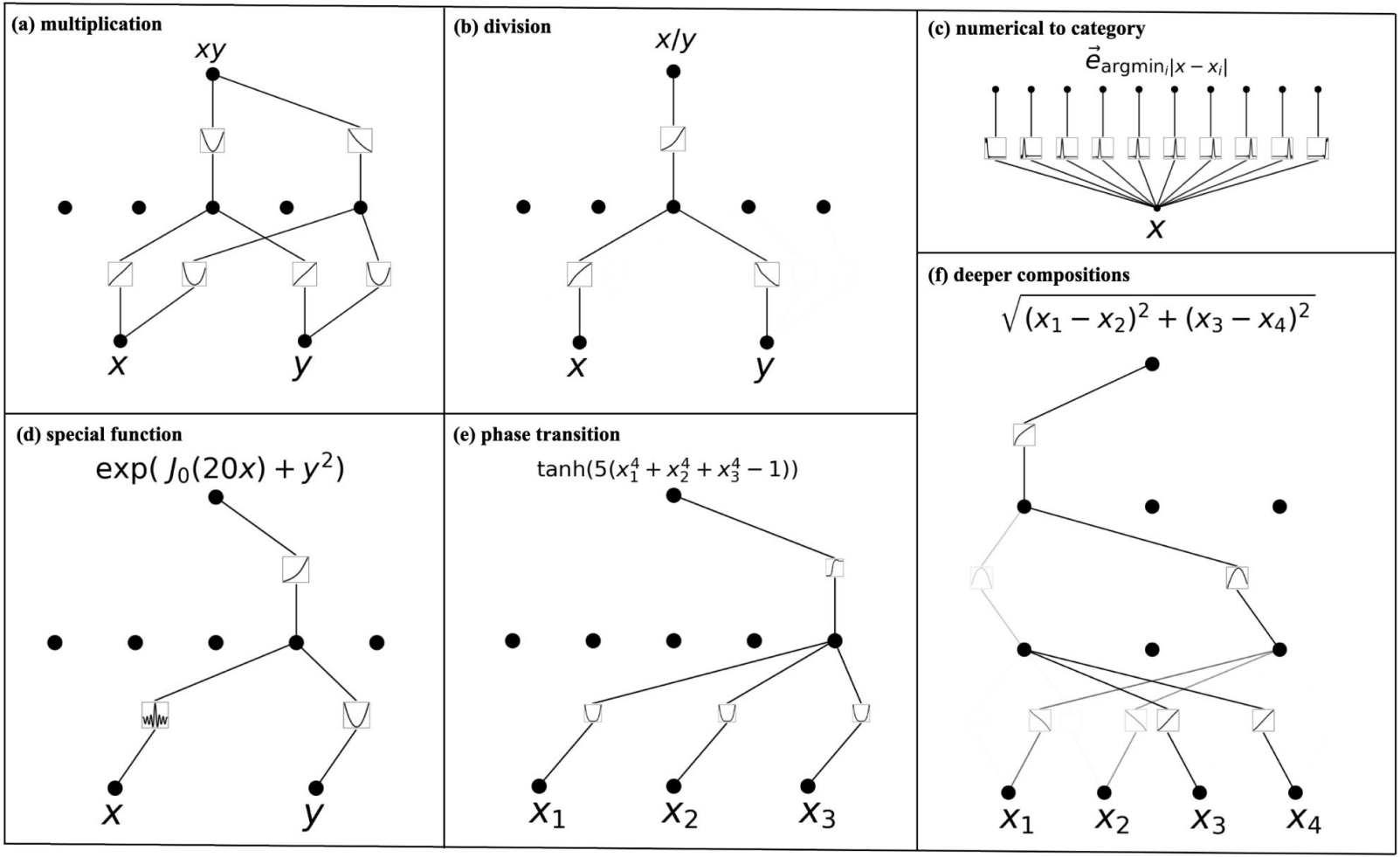
Csúcsok összeadják az élek értékeit

Pozitívum: Jobban, könnyebben értelmezhető

Negatívum: Lassabban tanítható, magas memóriaigény



Kép forrása: https://github.com/KindXiaoming/pykan




Kép forrása: https://github.com/KindXiaoming/pykan

# KAN eredmények

Nem jó - Collabnak nincs elég ram-ja

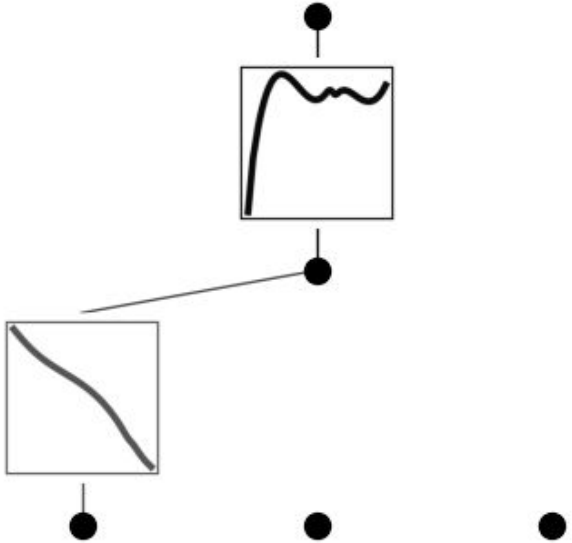
➙ nagyob butított modellt kellett futtatni

Adatok és változók kis részét felhasználva

(3 input 1 output változó)

Pruning nem futott le

Megtett kilométerek, Teljesítmény, Kor változók



Köszönjük a figyelmet!