Delež maščob - Uvod v odkrivanje znanj iz podatkov

Krištof Ocvirk, Tadej Kraševec

Uvod

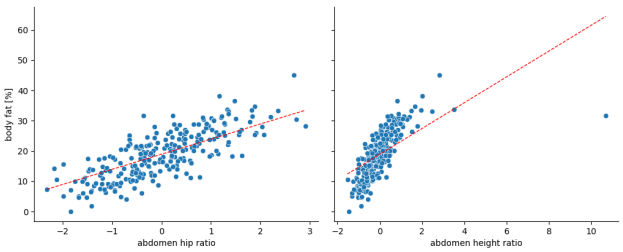
V tem projektu sva se ukvarjala z odkrivanjem najboljšega modela, ki bo na podlagi enostavnih meritev čim bolje napovedal delež maščob v telesu posameznika.

Analiza podatkov je bila narejena v Pythonu, kjer sva uporabila knjižnice *numpy*, *pandas*, *scikit-learn* in *seaborn*. Nekaj vizualizacij sva naredila tudi v programu Orange.

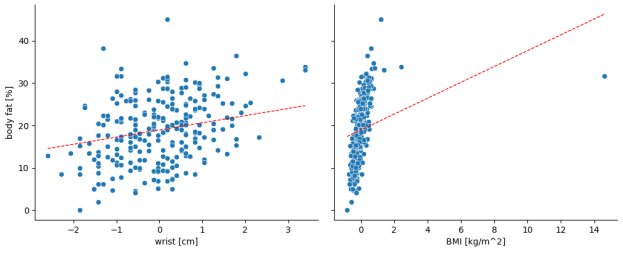
Model in podatki

Modelov za regresijo je veliko in po parih preizkušanjih nekaterih od njih sva se na koncu odločila za linearno regresijo z Lasso regularizacijo. Vse podatke razen ciljne spremenljivke sva tudi standardizirala.

Najino začetno razmišljanje je bilo, da teža in velikost sama po sebi ne povesta veliko, zato sva kot dodatne značilke vključila še tri nove značilke: index telesne mase, razmerje med obsegom trebuha in višino, razmerje med obsegom trebuha in bokov. Razlog za to je jasno viden na spodnjih slikah (Slika 1, 2 in 3),



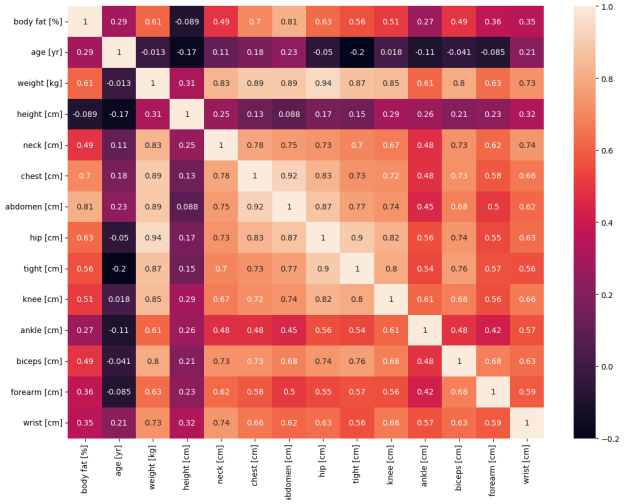
Slika 1: *Grafa: razmerje trebuha in bokov(levo), razmerje trebuha in višine(desno)*



Slika 2: *Grafa: obseg zapestja(levo), indeks telesne mase(desno)*

ki prikazujejo kako posamezne značilke vplivajo na delež maščob v telesu. Število primerov v podatkih je 252, kjer ni manjkajočih podatkov. So pa kljub temu v podatkih iztopajoči primeri, ki so tudi očitno razvidni na Slikah 1, 2, zato je model morda manj točen pri napovedi kot v situaciji brez iztopajočih primerov.

Na sliki 4 vidimo so značilke kolerirane med seboj. Kjer je korelacija velika (blizu 1), bi se lahko eno izmed značilk odstranilo brez vidne izgube v točnosti napovedi modela (npr. stegno ali bok, trebuh ali prsni koš, ipd).

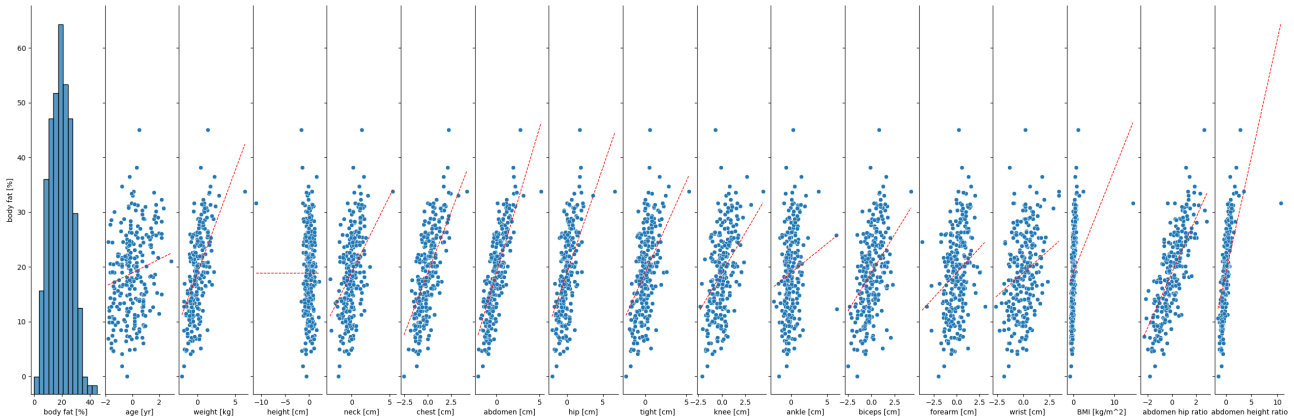


Slika 4: *Graf korelacij med značilkami*

Rezultati

Med analizo sva poskusila z linearno regresijo brez regularizacije, z Lasso regularizacijo, z Ridge regularizacijo, regresijska drevesa in regresijske gozdove. Poleg teh modelov pa sva še preverila, kakšno napako ima napoved s povprečno vrednostjo.

Modele sva ocenila s petkratnim prečnim preverjanjem. Za oceno natančnosti modela sva se zanašala na MSE (Mean Squared Error), ki je bila pri večini modelov okoli 4 (v večini zgibov). Izmed modelov, ki sva jih preverila sva se odločila za linearno regresijo z lasso regularizacijo. Model nama je dal najpreprostejši rezultat z napako, ki ni dosti odstopala od drugih modelov. Izpis koeficientov modela z linearno regre-



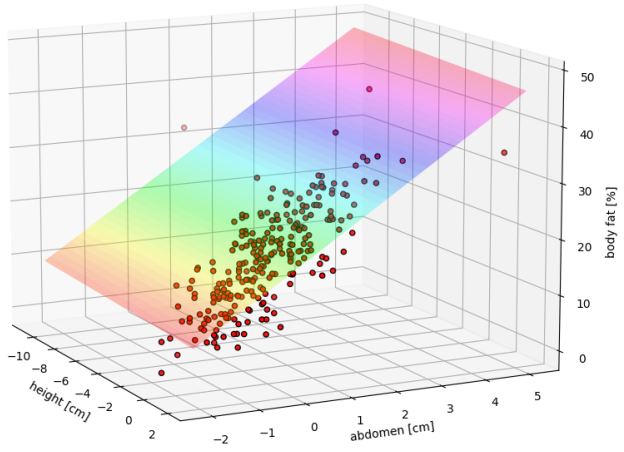
Slika 3: *Razpršeni grafi, ki prikazujejo delež maščob glede na posamezno značilko. Rdeča premica predstavlja linerno funkcijo, kjer sta parametra dobljena z linearno regresijo z Lasso regularizacijo*

sijo z Lasso regularizacijo:

```
[0.  0. -0.15263204 -0.  0.  5.31630572 0.  0.  0.
 -0.  0.  0. -0.]
```

kjer prvi neničelni koeficient predstavlja višino, drugi neničelni pa obseg trebuha.

Na koncu sva se odločila, da bova trenirala model samo na originalnih značilkah, ker sva tako dobila najpreprostejši model. S tem modelom lahko predvidimo delež maščob samo z višino in obsegom trebuha. Vizualizacijo modela se vidi na sliki 5.



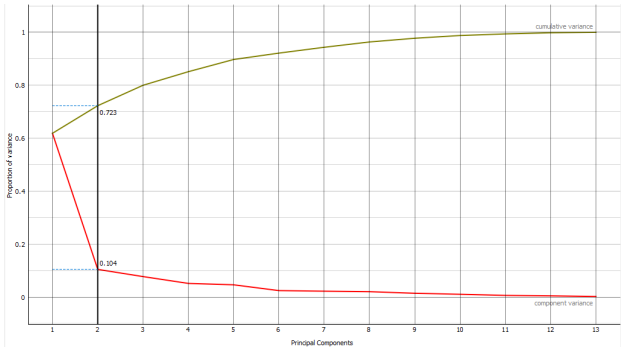
Slika 5: *Vizualizacija modela na značilkah: višina in obseg trebuha*

Vsi rezultati so ponovljivi in če zaženete katerega od modelov, boste dobili izpis oblike:

```
Error for linear regression:
MAX_ERROR: 72.15415907791083
MSE: 12.37822019730445
MAE: 4.613625744533458
Errors of cross valuation sets
MAX_ERROR: [321.61139751 8.73162504 10.04680942
 9.05815617 11.32280726]
MSE: [45.22145461 3.94914359 4.07080608
 3.72808884 4.92160787]
MAE: [9.74355465 3.20373238 3.22182206
 2.96775382 3.93126582]
Coefficients: [ 0.70367131 1.40552034 1.55173492
 -1.11065035 -0.4377729 -9.45159775
 4.81186704 1.0268425 -0.04882306 0.24524287
  0.43534886 0.78680295
 -1.40464212 -6.88995029 5.49318608
  11.93692931]
Intercept: 18.93849206349207
```

Diskusija

Zelo dober dokaz zakaj obseg trebuha največ doprinese k deležu maščob je tudi viden na Sliki 6, ki prikazuje koliko variance nam doprinese posamezna komponenta, kjer prva komponenta predstavlja lastni vektor, usmerjen proti značilki obsega trebuha.



Slika 6: *Graf analize glavnih komponent*