Velika domača naloga: Delež maščob

Nika Čelan, Aleks Stepančič

[1] Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Abstract

Cilj naloge je bil narediti čim boljši model za napovedovanje deleža maščob pri odraslih moških. Dodala sva tri značilke, ki so se izkazale za smiselne, ko sva preizkušala pomembnost značilk z različnimi metodami. Preizkusila sva tudi več modelov in za končno napoved izbrala linearno regresijo z elastic net regularizacijo. Napoved sva testirala s prečnim preverjanjem in uporabila meri RMSE in R^2. Dobila sva rezultate in sicer .....

Uvod

Natančne meritve deleža maščob so zamudne in drage, zato želimo narediti model, ki bo delež napovedal na podlagi enostavnejših meritev, kot so starosti, višina, teža in obsegi različnih delov telesa. Dani so podatki o deležu maščob in različnih meritvah za 252 moških teles.

Za analizo sva izbrala ogrodje Python in uporabljala knjižnjice pandas, scikit-learn in matplotlib za vizualizacije.

Metodologija

Obdelava podatkov

Vse podatke razen ciljne spremenljivke sva najprej normalizirala z min-max normalizacijo po formuli $X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$.

Odstranila sva dva primerka. En ima ciljno spremenljivko enako nič. Pri drugem sva pri začetnih poskusih z linearno regresijo in prečnim preverjanjem opazila, da je za vsak poskus na eni izmed testnih podmnožic R^2 zelo negativen. Primer je bil osebek, ki je tehtal 93.0 kg in bil visok 74.9 cm.

Nato sva naredila še vložitev v $\mathbb{R}^2$ z PCA in t-SNE, da bi videla, če imajo podatki kakšno posebno obliko.

Dodajanje značilk

Ob pregledu podatkov, se nama je zdelo smiselno dodati še nekatere značilke interakcij, ki jih je bilo mogoče izpeljati iz danih atributov. Dodala sva razmerje med višino in težo, indeks telesne mase ter po formuli 2, ki računa približen delež maščob. Izračunan delež se je izkazal za zelo pozitivno koreliranega s ciljno spremenljivko Vir: `https://www.calculator.net/body-fat-calculator.html` (0.83).

Table 1: *Rezultati preizkušenih metod*

	RF	KNN	Lin. reg.
avg R^2	0.66	0.62	0.71
std R^2	0.05	0.05	0.02
avg RMSE	4.4	4.7	4.1
std RMSE	0.5	0.6	0.3

$$r = \frac{height}{weight}$$

$$ITM = \frac{weight}{(height * 100)^2} \tag{1}$$

$$BodyFatPercentage = \frac{495}{1.0324 - 0.19077 \cdot \log_{10}(waist - neck) + 0.15456 \cdot \log_{10}(height)} - 450 \tag{2}$$

Modeli

Med raziskovanjem, sva preizkusila nekaj napovednih modelov, najprej linearno regresijo z Elastic net regularizacijo, nato naključne gozdove in KNN(k-nearest neighbors).

Pri vseh je bil uporabljen Grid Search pri iskanju najboljših hyperparametrov modelov ter dvakrat ponovljeno petkratno prečno preverjanje na vseh podatkih, da se zmanjša disperzija napake. Pri iskanju smo maksimizirali R^2. Nato sva za vsakega izmed najboljših modelov izračunala tudi RMSE (root mean squared error) in R^2 ter njune standardne odklone pri prečnem preverjanju. Dobila sva rezultate prikazane v tabeli 1. Vsi preizkušeni modeli so bili infomativni.

Ko sva opravila začetno analizo z zgornjimi modeli sva se odločila, da bova za nadaljno analizo uporabila linearno regresijo, saj je ta model dal najboljša RMSE in R^2 rezultata in je poleg tega dobro razložljiv.

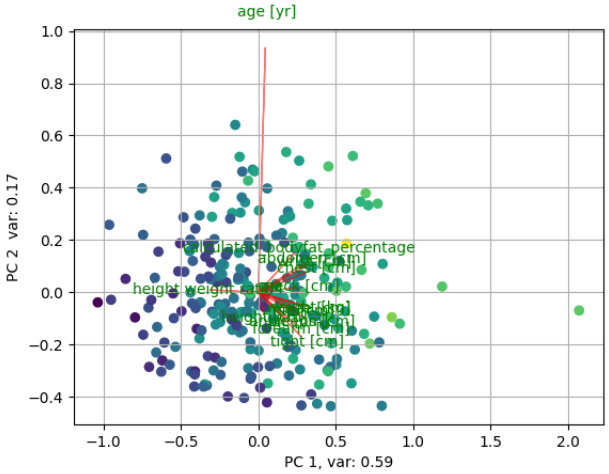

Figure 1: *Vizualizacija PCA metode*

Linearna regresija Uporabila sva Elastic net regularizacijo. To je metoda, ki linearni regresiji združi Lasso in Ridge regularizaciji in tako kopenzira pomankljivosti, ki jih imata vsaka za sebe. Pri Elastic net želimo minimizirati 3, pri čemer sta λ_1 in λ_2 hiperparametra.

$$\hat{\beta} \equiv \underset{\beta}{argmin}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1) \quad (3)$$

Uporaba regularizacije je rahlo izboljšala R^2 iz 0.71 na 0.72.

Rezultati in diskusija

Ugotovili smo, da se z najboljšim modelom v povprečju zmotimo za 4.1%. V Tabeli 2 vidimo vplih atributov na ciljno spremenljivko v linearni regresiji razvrščeni po absolutni vrednosti padajoče. Naključni gozd sva polega napovedovanja uporabila tudi za določanje pomembnosti atributov, rezultate lahko vidimo na Tabeli 3, ki poleg prikazuje tudi korelacijski koeficijent antributov z ciljno spremenljivko ter koeficiente linearne regresije. Zanimivo je, vrednosti koeficientov in iz naključnih gozdov precej sovpadata, vendar ITM (indeks telesne mase) naključni gozd obravnava kot pomemben atribut, medtem ko je koeficient linearne regresije približno nič.

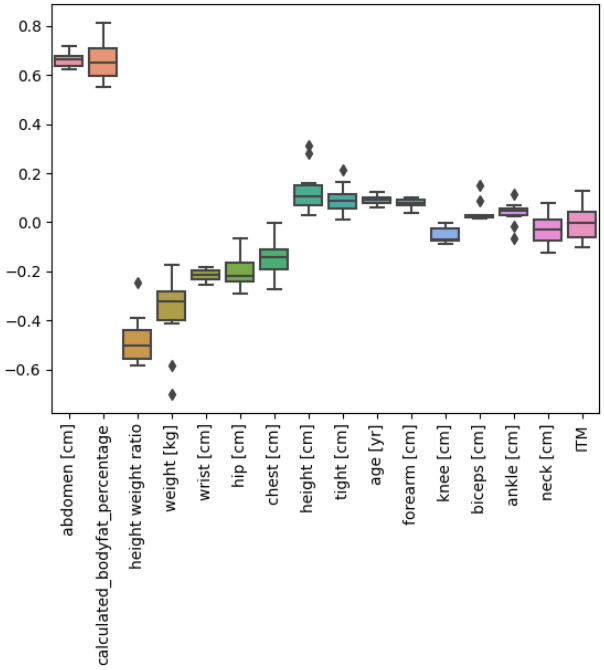

Figure 2: *Primerjava koeficientov pri linearni regresiji*

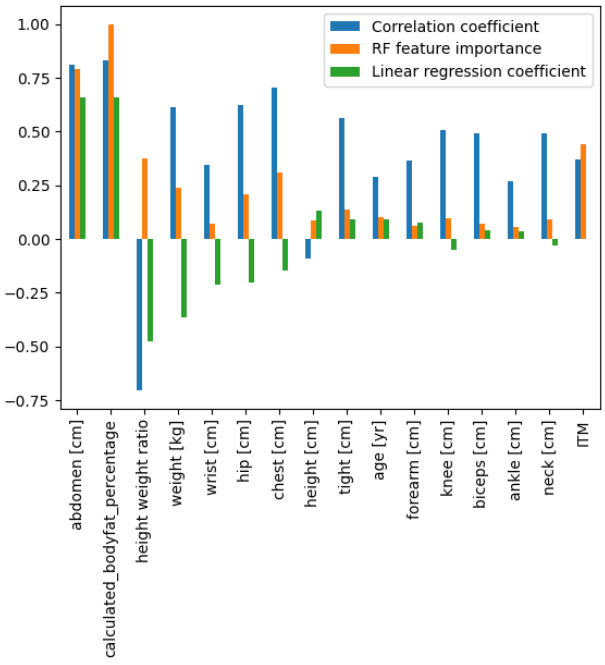

Figure 3: *pomembnost značilk dobljena s korelacijskim koeficientom, naključnimi gozdovi in linearno regresijo. (pomembnost rnačilk z naključnimi gozdovi je normalizirana)*