

Matematičke osnove mašinskog učenja, kol. 3

U fajlu `sport1.csv` su dati podaci 7 atributa izmerenih kod 20 sportista muškog pola slične starosti, visine i mase.

Atributi su:

Naziv kolone	sadržaj
1. starost	starost u godinama
2. visina	visina u centimetrima
3. masa	telesna masa u kilogramima
4. sistolni	sistolni pritisak u mirovanju mmHg
5. dijastolni	dijastolni pritisak u mirovanju mmHg
6. pocetak	početak pika opterećenja
7. kraj	kraj pika opterećenja

1. Napraviti preko eksplanatornih promenljivih "starost", "masa" **general linear model** `sport.lm` multiple regresije za response promenljivu "kraj".
Napisati u vežbanki formulu za dobijeni model verovatnoće i statističku značajnost dobijenih parametara, Adjusted R-squared i da li model ima statističku značajnost.
2. Za dobijeni model prognozirati srednju vrednost kraj pika opterećenja za osobe koje imaju 30 godina i 80 kg.
3. Iz modela dobijenog u prethodnom zadatku izbaciti promenljivu koja ima manju statističku značajnost i proveriti da li je dobijen model `sport.lm1` i dalje ima statističku značajnost.
4. U modelu `sport.lm` dodati eksplanatornu promenljivu "sistolni" i napraviti `sport.lm2`. Napisati novodobijeni Adjusted R-squared, p-vrednost modela, statističku značajnost novododate promenljive. Komentarisati koji od modela `sport.glm`, `sport.glm1` i `sport.glm2` je najbolji po kriterijumu Adjusted R-squared, a koji po p-value.
5. (Teorija) Šta je problem regresije? Šta je problem linearne regresije? Šta je ocena maksimalne verodostojnosti? Izvesti formulu za ocenu maksimalne verodostojnosti parametara linearne regresije. Da li su i zašto ocena maksimalne verodostojnosti i ocena minimuma sume kvadrata greške jednake?