Navodila za končni projekt - Delež maščob

Uvod v odkrivanje znanja iz podatkov

March 31, 2023

1 Navodila za delo

Pred vami so navodila za pripravo projekta pri predmetu Uvod v odkrivanje znanj iz podatkov. Dani so podatki o deležu maščob in različnih meritvah 252 moških teles. Ker so natančne meritve deleža maščob zamudne in drage, želimo narediti model, ki ga bo napovedal na podlagi enostavnih meritev.

V nalogi boste morali:

- pregledati podatke in narediti **vsaj** dve zanimivi vizualizaciji podatkov,
- iz osnovnih podatkov izpeljati **vsaj** dve novi značilki in jih obrazložiti,
- izbrati najprimernejši napovedni model in svojo izbiro podkrepiti,
- ovrednotiti in razložiti model.

Oblikujte skupine po največ dva študenta. Oba študenta bosta za delo ocenjena z enako oceno, zato partnerja izberite skrbno.

Od vas pričakujemo:

- GitHub repozitorij z vašo kodo za analizo podatkov (eden v skupini ustvari nalogo preko povezave, kolega se pridruži ustvarjeni skupini).
- poročilo vašega dela (največ 2 strani) napisano z IMRaD strukturo (napisano v LaTeX, predlagamo pisanje v Overleafu).
- 3 minutno predstavitev vašega dela (brez ppt, na projektorju bo vaše poročilo), kateremu sledi kratka diskusija.

Do roka oddaje morate na spletno učilnico oddati PDF poročilo. Del ocene predstavlja tudi vaša koda, zato se potrudite, da bo vaš repozitorij lepo urejen. Zagovori bodo potekali v tednu po roku oddaje.

2 Opis podatkov

Podatki vsebujejo meritve deleža maščob (angl. *body fat*) preko meritev gostote telesa in teže pod vodo. Meritev deleža maščob bo vaša ciljna spremenljivka.

Vaše značilke so starost (angl. *age*), teža (angl. *weight*), višina (angl. *height*) ter meritve obega delov telesa (vrat, prsni koš, trebuh, boki, stegna, kolena, gležnji, biceps, podlaket, zapestje). Enote si lahko prosto pretvorite v tiste, s katerimi boste lažje operirali.

3 Navodila za izvedbo analize

Kako se lotite analize podatkov, je prepuščeno vam. Če dobro zadovoljite zgoraj naštete kriterije, boste opravili projekt. Pomembno pa je, da znate upravičiti vse korake analize. Če razlage ne boste napisali v poročilo, vas bomo zagotovo o tem vprašali na zagovoru.

Izbira okolja za analizo je prepuščena vam (npr. Python, Orange, Matlab, Java). Dovoljena je uporaba vseh knjižnic, ki vam bodo olajšale analizo (npr. pandas, seaborn, scikit-learn). Algoritmov za obdelavo in modelov vam ni treba implementirati. Enostavno lahko uporabite katero od knjižnic, ki vam te metode že ponuja (npr. scikit-learn). Priporočamo, da uporabljene metode poznate.

Vaš repozitorij naj bo urejen in naj vsebuje README.md datoteko z opisom vašega dela. Pomembno je tudi, da ustrezno opišete korake, ki so potrebni za ponovitev vaše analize. Če zainteresirani bralec ne more ponoviti vaše analize, je ta neuporabna.

4 Priprava poročila in zagovor

Poročilo pripravite skladno z IMRaD strukturo. Omejitev dveh strani je striktna. Poudarek je na korektno izvedeni analizi, modeliranju in poročanju z informativnimi slikami. Ne pozabite, slika pove več kot 1000 besed, zato veliko časa vložite v dobre vizualizacije.

Poročilo bo služilo kot vaša predstavitev na zagovoru. Pripravite 5 minutno predstavitev, kjer predstavite samo najpomembnejše izsledke. Sledi diskusija o vaši analizi, kjer bomo preverili, če jo razumete.

Ocena vašega dela je sestavljena iz kode, poročila, predstavitve in zagovora.