Napovedni model za napovedovanje deleža maščob v moških telesih

Rok Rajher[1] in Jan Kuhta[1]

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

Uvod

Natančno poznavanje telesne maščobe je pomembno, saj je prekomerna količina maščobe v telesu povezana s številnimi zdravstvenimi težavami, kot so srčno-žilne bolezni, diabetes, visok krvni tlak, debelost in drugo, vendar so natančne meritve deleža maščob v človeških telesih so zamudne in drage. Cilj raziskovalne naloge je ustvariti napovedni model na podlagi metod strojnega učenja, ki bo znal dobro napovedati delež maščob na podlagi enostavnih meritev.

Metodologija

Analiza in priprava podatkov

V nalogi je bila uporabljena množica podatkov, ki vsebuje 252 vrstic in 13 atributov ter ciljno spremenljivko ki opisuje delež telesne maščobe. Atributi vključujejo meritve različnih delov telesa, kot so vrat, prsi, trebuh, kolki, stegna, gležnji in zapestja. V podatkih ni bilo manjkajočih vrednosti, kar je omogočilo gladko obdelavo in analizo podatkov.

Pred nadaljnim procesiranjem podatkov sva med atribute dodala atribute "BMI", ki predstavlja razmerje med višino in kvadratom telesne teže, atribut "abdomen-to-hip ratio", ki predstavlja razmerje obsega trebuha in bokov, "abdomen-to-height ratio" ki predstavlja razmerje obsega trebuha in telesne višine, ter atribut "mean joints (cm)", ki predstavlja povprečje seštevka normaliziranih meritev kolena, zapestja, gležnjev ter vratu.

Zaradi visoke korelacije med atributi sva naknadno iz podatkov odstranila atribute starosti, višine, teže, vratu, prsi, kolen, zapestja, gležnjev, obsega trebuha ter boke, saj so bili ti atributi preveč korelirani in bi negativno vplivali na interpretacijo modelov.

Na podlagi vrednosti BMI sva iz podatkov odstranila dva posamezika, pri enem je najverjetneno šlo za napako pri vnosu podatkov, saj je imel vrednost BMI nad 160, pri drugem primeru pa je šlo za osamelca, ki je imel vrednost BMI 48, kar ni reprezentativno za slovensko populacijo. Na koncu sva odstranila atribut BMI, saj je imel izjemno visoko korelacijo. Končna korelacija med atributi je prikazana z heatmapom na sliki 2.

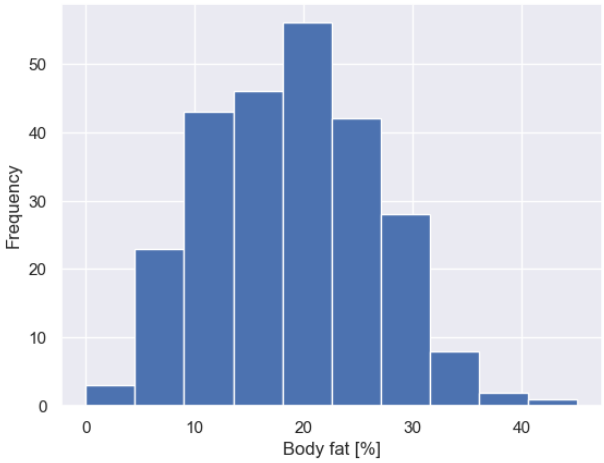

Figure 1: *Histogram porazdelitve deleža telesne maščobe*

Pred nadaljnjo obdelavo podatkov sva izvedla še standardizacijo atributov, saj imajo tako atributi kot ciljna spremenljivka normalno porazdelitev, kar je vidno iz grafa o ciljni spremenljivki na sliki 1.

Na koncu sva tako za strojno učenje uporabila množico podatkov, ki vsebuje 250 vrstic in 6 atributov. Podatke sva razdelila po principu 80-20 na učno (200 primerov) in testno (50 primerov) množico. Učna množica je bila namenjena treniranju modelov in izvedbi prečnega preverjanja, testna množica pa končnemu testiranju najboljših dveh modelov.

Uporabljene metode strojnega učenja

V okviru seminarske naloge sva poskušala zgraditi čim več regresijskih modelov in oceniti njihovo delovanje. Za oceno deleža telesne maščobe sva V nalogi uporabila modele SVM, KNN, nevronsko mrežo, XGBoost, Gradient Boost, Linearno regresijo in Naključni gozd. Vsi modeli so bili zgrajeni na enakih standardiziranih podatkih. Modele sva tudi optimizirala s prilagajanjem hiperparametrov (Grid Search).

Mere uspešnosti

Uspešnost posameznih modelov strojnega učenja sva merila z različnimi mejami uspešnosti. Uporabila sva srednjo kvadratno napako (MSE), srednjo absolutno napako (MAE), koren srednje kvadratne na-

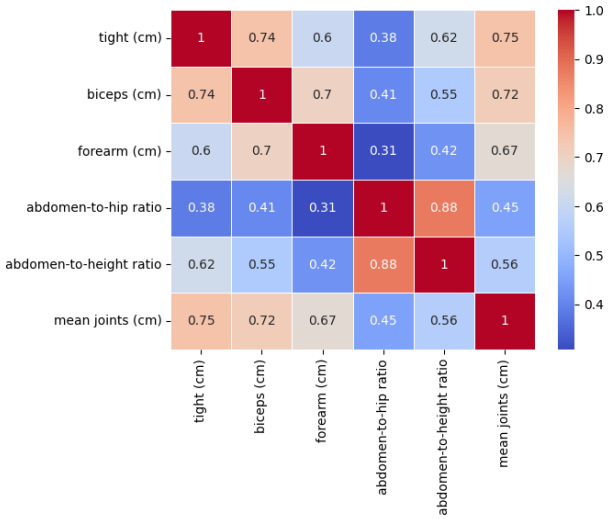

Figure 2: *Heatmap*

Table 2: *Mere uspešnosti modelov na testni množici*

Model	MSE	MAE	RMSE	R2
SVM	16.58	3.26	4.07	0.66
Ridge	16.84	3.30	4.10	0.63

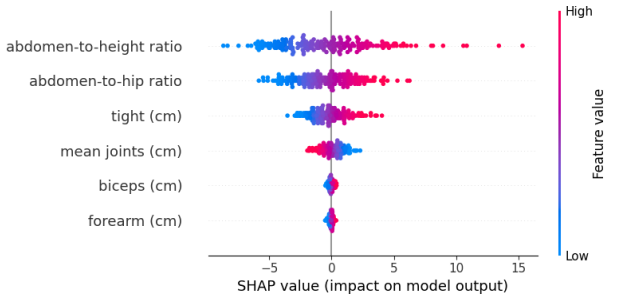

Figure 3: *SHAP vrednosti za model SVM*

pake (RMSE) in R-kvadrat (R2). Slednja dva sva uporabila le pri testiranju na testni množici.

Modele sva evalvirala z uporabo prečnega preverjanja tipa "Leave one out", v vsaki iteraciji izmerila zgoraj naštete metrike in izračunala povprečje, standarni odklon ter 95 odstotni interval zaupanja.

Rezultati

Vsakega izmed zgoraj naštetih modelov strojnega učenja sva s prečnim preverjanjem tipa "Leave one out" naučila in evalvirala z uporabo metrik MSE in MAE. Za vsako mero sva izračunala povprečje in standardni odklon. Dobljeni rezultati so prikazani v tabeli 1 naraščajoče po MSE.

Table 1: *Mere uspešnosti modelov na učni množici*

Model	MSE μ	MSE σ	MAE μ	MAE σ
SVM	17.05	21.31	3.34	2.42
Ridge	17.25	20.20	3.42	2.35
Lasso	17.26	20.26	3.42	2.36
LR	17.35	20.76	3.42	2.38
XGBoost	19.07	26.93	3.58	2.50
GB	19.36	25.38	3.65	2.46
RF	19.72	27.46	3.64	2.54
KNN	20.66	26.83	3.72	2.61
NN	56.03	88.67	5.89	4.62
Dummy	60.98	82.20	6.35	4.55

Modela, ki sta dosegla najboljše rezultate pri prečnem preverjanju, torej SVM in Ridge sva nato naučila na celotni učni množici in ju testirala na do tega trenutka še ne videni testni množici. Izračunala sva metrike MSE, MAE, RMSE in R2. Rezultati so prikazani v tabeli 2 naraščajoče po MSE.

Razprava

Iz rezultatov (mer uspešnosti posameznih modelov) je razvidno, da je večina uporabljenih modelov občutno izboljša napako pri napovedovanju v primerjavi z regresorjem z ničelnim koeficientom. Najmanjše napake je dosegel model SVM, za njim pa so sledili linearni modeli, torej Ridge, Lasso in linearna regresija. Model nevronske mreže se ni pretirano izboljšal v primerjavi z regresorjem z ničelnim koeficientom, verjetno zaradi enostavnosti problema, malega števila primerov in nizke globine nevronske mreže.

Model SVM

Iz tabele 1 je razvidno, da je model SVM dosegel povprečno kvadratno napako 17.05 s standardnim odklonom 21.31 in 95 odstotnim intervalom zaupanja [14.08, 20.02] ter povprečno absolutno napako 3.35 z odklonom 2.42. Iz slike 3 je razvidno, da je model SVM baziral na novo dodanih atributih "abdomen-to-hip ratio" in "abdomen-to-height ratio", kar upravičiči najino izbirno dodatnih atributov. Višji kot sta ti dve razmerji in večji kot je obseg stegen, večji je odstotek maščobe. Zanimiv atribut je "mean joints", torej povprečje normaliziranih obsegov sklepov, ki je z odstotkom maščobe obratno sorazmeren. Razlaga tega je, da imajo natrenirani moški z manjšim odstotkom maščob močnejše oz večje sklepe. V primerjavi z regresorjem z ničelnim koeficientom, ki ima povprečno kvadratno napako 60.98, je prišlo do 73 odstotne izboljšave napovedi.

Model Ridge

Za razlago modela Ridge sva pogledala vrednosti koeficientov in tako ocenila pomembnost atributov.

Model je zgradil naslednjo linearno kombinacijo atributov, kjer atributi po vrsti predstavljajo atribute tight (cm), biceps (cm), forearm (cm), abdomen-to-hip ratio, abdomen-to-height ratio, mean joints (cm).

$$y = 19.00 + 1.29x_1 + 0.25x_2 + 0.10x_3 + 2.55x_4 + 3.54x_5 - 0.59x_6$$

Iz koeficientov je razvidno, da imata največjo težo pri odločitvi abdoment-to-height ratio in abdomen-to-hip ratio, medtem ko imata biceps (cm) ter forearm (cm) zelo mali doprinos. Slednji je tako kot pri modelu SVM obratno sorazmeren s ciljno spremenljivko.

Kljub dobri izboljšavi, so napake še vedno relativno velike, kar je najverjetneje posledica majhne učne množice. Z dodajanjem večjega števila primerov v učno množico bi lahko modele izboljšali.

Ridge in SVM na testni množici

Modela SVM in Ridge sva na koncu naučila na celotni učni množici in testirala na do sedaj še ne videni testni množici. Iz tabele 2 je razvidno, da sta oba modela dosegla primerljive napake s tistimi na učni množici, kar pomeni, da se modela nista pretirano prilagodila učnim podatkom in ostala relativno natančna tudi na testni množici.

Zaključek

V sklopu seminarske naloge sva analizirala, preoblikovala in standardizirala podatke in jih v nadaljevanju uporabila za izgradnjo osmih različnih regresijskih modelov strojnega učenja. Modele sva evalvirala s prečnim preverjanjem. Za najboljšega sta se izkazala modela SVM in Ridge, ki sta tudi temeljila na novo dodanih atributih. Ta dva modela sva na koncu tudi testirala na testni množici in Končna korenjena povprečna kvadratna napaka je bila 4.07, kar je primerljivo s trenutno natančnostjo slabših tehtnic za merjenje odstotka maščob. Meniva, da bi z uporabo večje učne množice dosegla boljše rezultate in zgradila modele, ki bi lahko bili namenjeni javni uporabi.