Procjena rizika od kardiovaskularnih bolesti

Matej Petrinović

1 Uvod

Kardiovaskularne su bolesti (KVB) većinom uzrokovane aterosklerozom odnosno promjenama, oštećenjima i naslagama na stijenci arterija.

Svaka bolest srca, vaskularne bolesti mozga odnosno bolesti krvnih žila mozga spadaju u kardiovaskularne bolesti. Najčešće kardiovaskularne bolesti uključuju koronarnu srčanu bolest (npr. srčani udar) i cerebrovaskularnu bolesti (npr. moždani udar). Kontrola čimbenika rizika kao što su prehrana, tjelesna aktivnost, upotreba duhanskih proizvoda i kontrola krvnog tlaka mogu smanjiti rizik od navedenih bolesti.

KVB su vodeći uzrok smrti i invaliditeta u svijetu: preko 17,5 milijuna ljudi svake godine umre od kardiovaskularnih bolesti. Ishemijska srčana bolest (npr. srčani udar) je odgovorna za 7,3 milijuna od ukupno broja kardiovaskularnih smrti, a cerebrovaskularne bolesti (npr. moždani udar) su odgovorne za 6,2 milijuna smrti.

Cilj ovog malog projekta je procjeniti vjerojatnost oboljenja od srčanih bolesti na temelju određenih faktora. Za modeliranje koristit ćemo Heart Disease UCI podatke dostupne na `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`, te logističku regresiju.

2 Analiza podataka

Najprije ćemo pogledati tabličan pregled podataka. Nakon pregleda slijedi opis i analiza podataka pojedine varijable gdje ćemo dati opis, opisnu statistiku, te provesti testiranja koliko ta varijabla utječe na rizičnost od oboljenja.

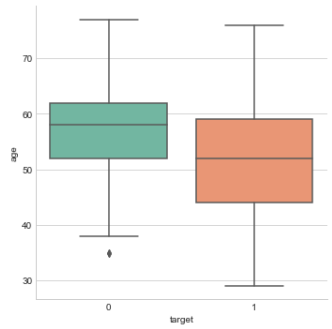
2.1 Dob

Varijabla **Dob** (Age) je numeričkog tipa i prikazuje dob pacijenta u godinama. U idućoj tablici su prikazani osnovni pojmovi varijable dob.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age	29.00	47.50	55.00	54.37	61.00	77.00

Tablica 1: Osnovna statistika varijable dob

Iz tablice vidimo kako je prosječna dob pacijenata 54 godine, dok najmlađi odnosno najstariji pacijent imaju 29 i 77 godina redom. Na idućoj slici pogledat ćemo usporedni kutijati dijagram dobi po rizičnosti.



Slika 1: Usporedni kutijasti dijagram dobi

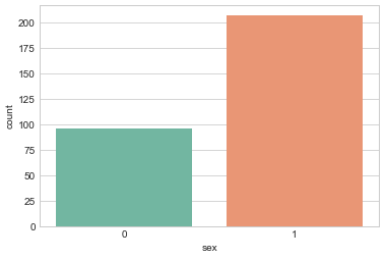
Čini se da pacijenti s odsutnošću i prisutnošću srčanih bolesti imaju nešto drugačiju raspodjelu, jer se čini da pacijenti sa srčanim bolestima pokazuju tendenciju da budu mlađe dobi. Testirat ćemo hipotezu o prosječnoj dobi obzirom na rizik. Na razini značajnosti $\alpha = 0.05$ dobivamo T-testom p-vrijednost 0.00000281. Možemo zaključiti kako postoje statistički značajne razlike dobi u odnosu na rizičnost.

2.2 Spol

Varijabla **Spol** (Sex) je kategorijalna varijabla koja označuje dob pacijenta, pri čemu oznaka 1 označava muškarce, 0 žene. Radi uporedbe pogledajmo tablicu i grafički prikaz distribucija po spolu.

	F	M
Count	96	207

Tablica 2: Raspodjela po spolu



Slika 2: Distribucija po spolu

Pogledat ćemo sada usporednu tablicu spola ovisno o riziku.

sex/ risk	0	1	Total
0	24	72	96
1	114	93	207
Total	138	165	303

Testiranjem zavisnosti χ^2 testom dobivena je p-vrijednost 0.00000187 pa možemo tvrditi postojanje zavisnosti. Ako izračunamo odds ration dobivano vrijednost 3.7, što govori da žene imaju 3.7 puta veću šansu za oboljenje od kardivaskularnih bolesti.

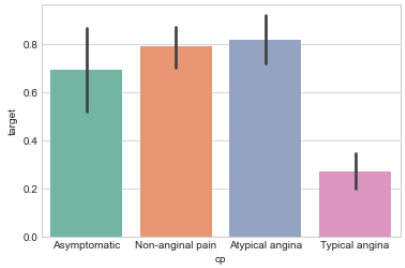
2.3 Tip boli

Tip boli je kategorijalna varijabla koja opisuje tip boli u prsima pacijenta. Testiranjem zavisnosti χ^2 testom

	0	1	Total
Asymptomatic	7	16	23
Atypical angina	9	41	50
Non-anginal pain	18	69	87
Typical angina	104	39	143
Total	138	165	303

Tablica 3: Tip boli povezan s rizikom

dobivena je p-vrijednost reda 10^{-16} pa možemo tvrditi postojanje zavisnosti.



Slika 3: Distribucija rizičnosti obzirom na bol

Čini se da je tip tipične angine najmanje ozbiljan. također izgleda da ne postoji jasan obrazac povećanja rizika od srčanog udara u usporedbi između tipične i atipične angine i asimptomatske.

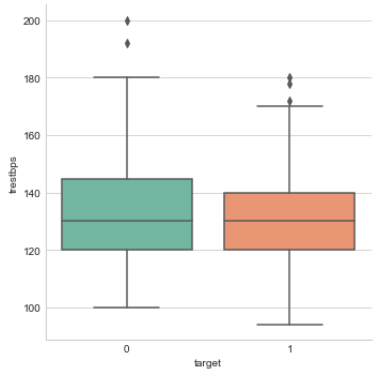
2.4 Krvni tlak

Varijabla **Krvni tlak** (trestbps) je numerička varijabla koja opisuje pacijentov krvni tlak u mirovanju (u mm Hg pri prijemu u bolnicu).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
KT	94.00	120.00	130.00	131.62	140.00	200.00

Tablica 4: Osnovna statistika varijable krvi tlak

Na idućoj slicipogledat ćemo usporedni kutijati dijagram dobipo rizičnosti.



Slika 4: Usporedni kutijasti dijagram krvnog tlaka

Mala je razlika u krvnom tlaku u mirovanju između 2 skupine. Čini se da je ova varijabla beznačajna u predviđanju bolesti srca. No testiranjem T-testom dobivena je p-vrijednost 0.01, pa možemo tvrditi postojanje razlike.

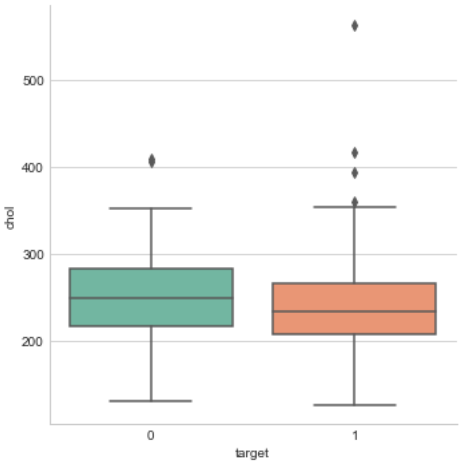
2.5 Kolesterol u serumu

Kolesterol u serumu je numerička varijabla koja opisuje serumski kolesterol u mg / dl.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
KS	126.00	211.00	240.00	246.26	274.50	564.00

Tablica 5: Osnovna statistika varijable serumski kolestrol

Pogledajmo nadalje usporedni grafički prikaz.



Slika 5: Usporedni prikaz varijable

Zanimljivo je da je prosjek bolesnika bez srčanih bolesti veći od onih sa srčanim bolestima. To ukazuje na to da visoka razina serumskog kolestola nije osobito korisna ako ne uzmemo u obzir razinu različitih vrsta kolesterola. Također testiranje T-testom dobivamo p-vrijednost 0.136 pa nema razloga sumjanjati u različitost kolesterola.

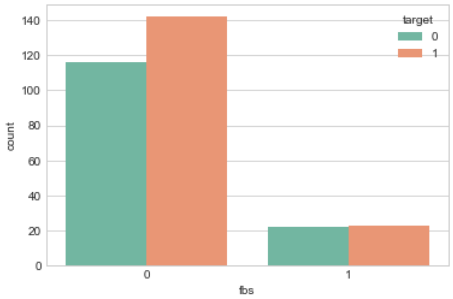
2.6 Šećer u krvi

Ova varijabla je kategorijalna i pokazuje visinu šećera u krvi i to tako da 1 označuje $>$ 120 mg/dl, dok 0 manje.

visina/risk	0	1	Total
0	116	142	258
1	22	23	45
Total	138	165	303

Tablica 6: Krostablica visine šećera i rizika

Testiranjem zavisnosti χ^2 testom dobivena je p-vrijednost 0.7444 pa ne možemo tvrditi postojanje zavisnosti. Samim ne postojanem zavisnosti nije potrebno računati ni odds ratio, jer bi njegov pouzdani interval sadržavao 1, te stoga se nebi donjeo kvalitetan zaključak.



Slika 6: Odnos šećera u krvi po rizičnosti

Nema značajnih razlika između zdravih i bolesnika sa srčanim bolestima u razini šećera u krvi, bilo više ili manje od 120 mg / dl. Ali lagani porast vjerojatnosti može se primijetiti kod pacijenata s visokom razinom šećera.

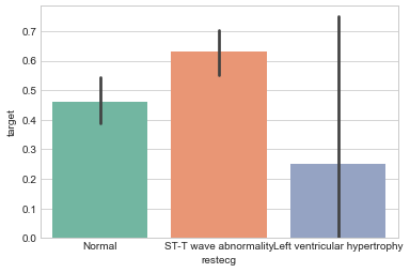
2.7 Elektrokardiografski rezultati

Kategorijalna varijabla koja opisuje rezultate elektrokardiografskog testiranja. Sastoji se od 3 kategorije i usporednu tablicu s rizikom se može vidjeti u nastavku:

Rez/ risk	0	1	Total
Left ventricular hypertrophy	3	1	4
Normal	79	68	147
ST-T wave abnormality	56	96	152
Total	138	165	303

Tablica 7: Krostablica rezulata

Testiranjem zavisnosti χ^2 testom dobivena je p-vrijednost 0.006661 pa možemo tvrditi postojanje zavisnosti.



Slika 7: Distribucija obzirom na rizičnost

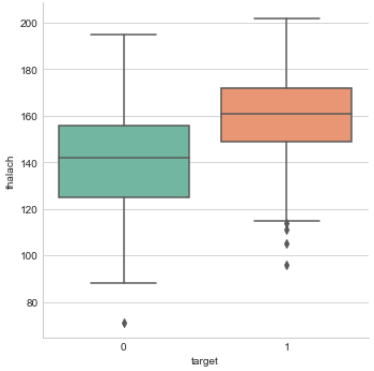
Čini se da pacijenti s restecg-om abnormalnosti ST_T vala imaju najveći rizik od oboljenja srca. Međutim, budući da postoje velike razlike u restecg-u, ne možemo zaključiti da pacijenti s restecgom hipertrofije lijevog ventrikula imaju manje bolesti srca.

2.8 Maksimalni puls

Maksimalni puls je numerička varijabla koja opisuje postignut maksimalni puls pacijenta. Osnove podatke se mogu vidjeti u tablici:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MP	71.00	133.50	153.00	149.65	166.00	202.00

Tablica 8: Osnovna statistika varijable puls



Slika 8: Usporedni kutijasti dijagram

Čini se da pacijenti sa srčanim bolestima imaju veći maksimalni puls postignuti od onih koji nemaju. To može biti dobar pokazatelj bolesti. T-test daje p-vrijesnost reda 10^{-14} te stoga se može tvrditi postojanje stat. značajne razlike.

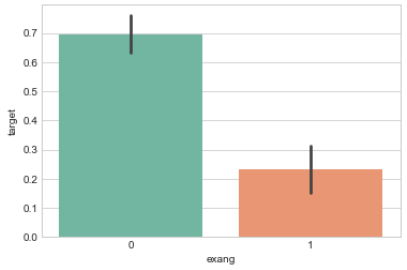
2.9 Angina

Varijabla Angina je kategorijalna varijabla i opisuje je li bol u prsima izazvana vježbom, i to na način da je 1 je da, 0 ne.

Angina/risk	0	1	Total
0	62	142	204
1	76	23	99
Total	138	165	303

Tablica 9: Krostablica rizičnosti

Testiranjem zavisnosti χ^2 testom dobivena je p-vrijednost 10^{-14} pa možemo tvrditi postojanje zavisnosti. Odds ratio iznosi 7.36, što znači da osobe kod kojih se nije javila angina bol imaju 7.36 posto veću šansu za razvijanje rizika.



Slika 9: Distribucija obzirom na rizik

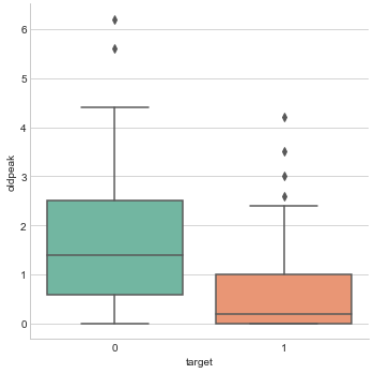
Vidimo jasnu korelaciju između angine izazvane vježbom i cilja; Pacijenti bez angine uzrokovane tjelovježbom imaju visok rizik od oboljenja.

2.10 ST Segment

ST-Segmet je numerička varijabla koja opisuje ST vrijednost potaknute vježbanjem u odnosu na odmor.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ST	71.00	133.50	153.00	149.65	166.00	202.00

Tablica 10: Osnovna statistika varijable ST



Slika 10: Usporedni kutijasti dijagram

Čini se da postoji tendencija da oni s nižim oldpeakom imaju veći rizik od oboljenja srca. Čini se da je ovo još jedan dobar pokazatelj. T-test daje vrlo malu p-vrijesnost stoga možemo tvdriti postojanje razlika.

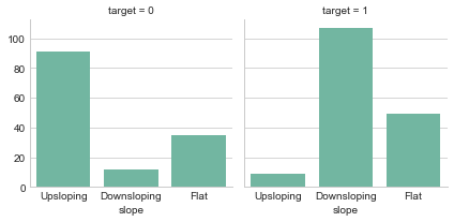
2.11 Nagib

Varijabla Nagib opisuje nagib vršnog vježbanja ST segmenta. Krostablica je dana u nastavku.

	Downsloping	Flat	Upsloping	Total
0	35	91	12	138
1	107	49	9	165
Total	142	140	21	303

Tablica 11: Kros tablica

Testiranjem zavisnosti χ^2 testom dobivena je p-vrijednost 10^{-11} pa možemo tvrditi postojanje zavisnosti.



Slika 11: Distribucija o riziku

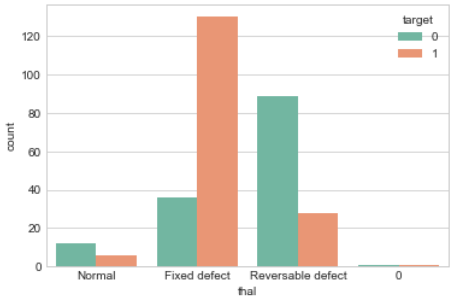
Pacijenti koji nemaju bolest imaju tendenciju da imaju nagib blizu Upslopinga dok pacijenti sa srčanim bolestima imaju nagib blizu Downslopinga.

2.12 Thalassemia

Varijabla Thalassemia je kategorijana varijabla koja indicira izloženost thalassemiji.

	0	Fixed defect	Normal	Reversable defect	Total
0	1	36	12	89	138
1	1	130	6	28	165
Total	2	166	18	117	303

Testiranjem zavisnosti χ^2 testom dobivena je p-vrijednost 10^{-14} pa možemo tvrditi postojanje zavisnosti.

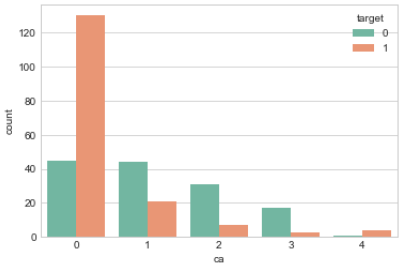


Slika 12: Razlika izloženosti po rizku

Možemo utvrditi da većina bolesnika s otkrivenom srčanom bolešću pokazuje reverzibilni nedostatak, a većina zdravih bolesnika pokazuje abnormalne rezultate.

2.13 Krvne žile

varijabla **Krve žile** pokazuje broj glavnih krvnih žila.



Slika 13: Odnos na rizičnost

Budući da je za ca$>=1$ dostupno manje podataka, ne možemo dobro razumjeti kako ti brojevi pridonose predviđanju. Ali gledajući ca 0, više je nego dvostruki broj pacijenata koji imaju srčane bolesti u odnosu na one koji nemaju. Ovaj se rezultat može recionalizirati da oni koji imaju manje glavne žile imaju veću vjerojatnost da imaju srčane bolesti. Testiranjem zavisnosti χ^2 testom dobivena je p-vrijednost 10^{-15} pa možemo tvrditi postojanje zavisnosti.

3 Modeliranje rizika

3.1 Procjena parametara

U ovom djelu rada bavimo se modeliranjem rizika od oboljena. Rizik ćemo u ovom slučaju shvaćati kao vjerojatnost od oboljenja. To ćemo procjenjivati koristeći logističku regresiju. Logistička regresija je funkcija $\sigma : \mathbb{R} \to (0,1)$ zadana formulom:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

U našem modelu ćemo koristiti sve varijable koje su pokazale statističku značajnost. U idućoj tablici možemo vidjeti procjenjene parametre u modelu logističke regresije, kao i p-vrijednosti, te pouzdani intervali.

	Estimate	Std. Error	z value	Pr(>|z|)	2.5 %	97.5 %
(Intercept)	0.67	3.04	0.22	0.83	-4.47	6.55
sex	-1.23	0.47	-2.61	0.01	-2.18	-0.33
cpAtypical angina	-0.86	0.74	-1.16	0.24	-2.35	0.58
cpNon-anginal pain	0.15	0.64	0.23	0.82	-1.14	1.37
cpTypical angina	-1.78	0.62	-2.84	0.00	-3.06	-0.59
thalach	0.01	0.01	1.47	0.14	-0.00	0.03
exang	-0.78	0.41	-1.89	0.06	-1.59	0.03
oldpeak	-0.54	0.22	-2.47	0.01	-0.98	-0.12
slopeFlat	-0.98	0.45	-2.20	0.03	-1.87	-0.12
slopeUpsloping	-0.17	0.88	-0.20	0.84	-1.84	1.61
ca	-0.81	0.20	-4.13	0.00	-1.21	-0.44
thalFixed defect	1.61	2.62	0.61	0.54	-2.88	6.22
thalNormal	1.54	2.70	0.57	0.57	-3.07	6.39
thalReversable defect	0.18	2.63	0.07	0.94	-4.29	4.83

Tablica 12: Procjenjeni parametri modela

Vidmo kako neki koeficijenti u modelu nisu statistički značajni (p-vrijednosti > 0.05 odnosno pouzdani interval sadrži 0). Radi jednostavnosti rada, neće se detaljnije raditi uklanjanje i detaljna analiza prediktora.

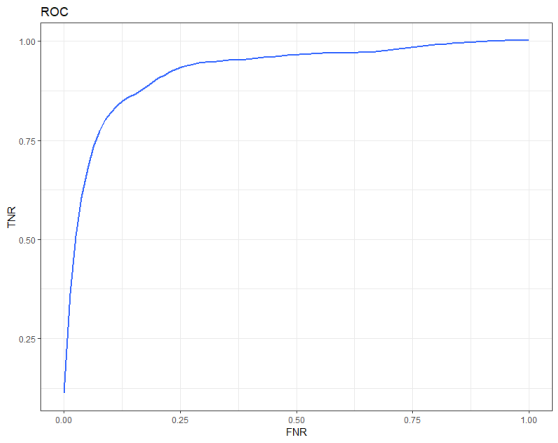
3.2 Validacija modela

Ovaj dio rada posvećen je validiranju modela, odnosno izračunavanja njegovih performansi. Kako logistička regresija vraća vrijednost iz intervala $(0,1)$ odredit ćemo tresh hold $T = 0.5$, stoga ćemo sve vrijednosti veće od T klasificirati tako rizično (1), u suprotnom nerizično. Pogledajmo confusion matricu za model:

Real/Pred	0	1
0	110	28
1	15	150

Tablica 13: Confusion matrica modela

Preciznost model izračunava se kao suma dijagonalnih elemenata kroz ukupna suma elemenata, tako da preciznost iznosi $ACC = 85.8\%$. Ono što je važno pogledati je ROC krivulja, te površina ispod nje.



Slika 14: ROC krivulja

ROC krivulja mora biti što dalje od simetrale kvadrata. što je ovdje i slučaj. Površina ispod ROC krivulje iznosi AUC $= 0.9245$. To nam govori da ovaj model dobro razlikuje rizične od ne rizičnih pacijenata, te se kao takav može koristit dalje.

Model je ilustrativne svrhe i za veće testiranje treba napraviti detaljniju analizu koeficijenata, napraviti CV, te isprobati druge modele.