Klasifikacija pušača i konzumenata alkohola

Una Aleksić, E9 3/2024, unaaleksic235@gmail.com
Marko Mrđa, E9 4/2024, marko.mrdja123@gmail.com

I. Uvod

Uticaj pušenja cigareta na život je značajan, kao i uticaj konzumiranja alkohola. Američki centar za kontrolu i prevenciju bolesti navodi pušenje kao razlog za smrt jedne petine amerikanaca svake godine. Pušenje značajno povećava rizik od srčanih bolesti, moždanog udara i razvijanja raka pluća.

Pored toga, dugotrajno konzumiranje alkohola takođe može da izazove bolesti srca, bolesti jetre, visok krvni pritisak i razne druge zdravstvene probleme.

Značaj analize podataka o pušačima i ljudima koji konzumiraju alkohol nalazi se u mogućnosti ranijeg uočavanja bolesti koje su karakteristične za ove grupe ljudi.

II. Baza podataka

Baza podataka se sastoji od informacija o pacijentima iz Južne Koreje, prikupljenih iz zavoda za nacionalno zdravstveno osiguranje. Među podacima se nalaze osnovne informacije o pacijentima kao što su pol, godine, visina i težina, informacije njihovim telesnim karakteristikama kao što su vid i sluh pacijenta, krvni pritisak i različiti parametri analize krvi i urina.

U bazi se nalazi 991346 uzoraka opisanih sa 22 obeležja, od kojih su 18 numerička i 4 kategorička obeležja. Baza nema nedostajućih vrednosti, ali sadrži nevalidne vrednosti u obeležjima: obim struka, HDL holesterol, LDL holesterol, kreatinin i AST.

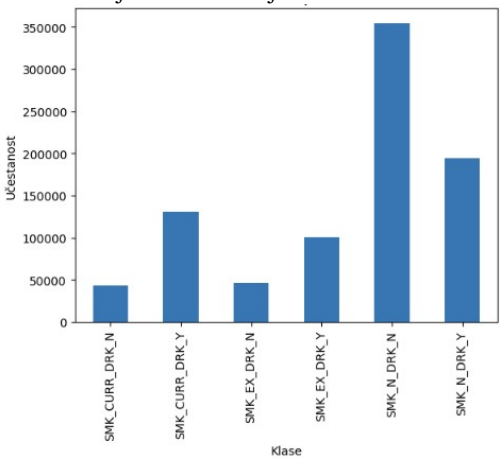
Klasu nepušača čini 60.77% uzoraka, bivših pušača 17.65% te klasu pušača čini 21.58% uzoraka. Klasu konzumenata alkohola čini 49.98% uzoraka koji konzumiraju alkohol i 50.02% koji ne konzumiraju alkohol.

III. Analiza i priprema podataka

Prvi korak analize podataka je provera da li je baza uspešno učitana u sistem proverom prvih nekoliko redova baze. Nakon toga analiziramo opis vrednosti svih obeležja u bazi, da li postoje nedostajuće vrednosti i da li postoje duplikati uzoraka.

Kada su ustanovljene osnovne informacije o bazi podataka, prelazi se na semantičku analizu gde analiziramo raspodelu kategorija pušača i kategorija konzumiranja alkohola po godinama pacijenata. Zaključuje se da broj pacijenata koji konzumira alkohol opada sa godinama, dok se kod pušača primeti da se broj ljudi koji prestaju da konzumiraju cigarete ubrzano povećava od srednjih godina pa nadalje. Kao dodatna analiza se prave „box i whisker“ grafikoni koji na pokazuju raspodelu podataka po obeležjima i pomaže nam da uočimo vrednosti koje se smatraju izuzecima (*engl. outliers)*. Uočava se da izuzetne vrednosti postoje u više obeležja, a to su: obim struka, HDL holesterol, LDL holesterol, trigliceridi, SBP, DBP, kreatinin, BLDS, AST, ALT i ukupni holester. Pravi se i korelaciona matrica kako bi se analizirala korelacija između podataka u bazi. Uočava se da postoji jaka pozitivna korelacija između LDL holesterola i ukupnog holesterola od 0.91. Postoje i jake negativne korelacije kao što je slučaj kod HDL holesterola i triglicerida od -0.41. Ova saznanja nam pomažu u razmevanju odnosa između obeležja u bazi podataka.

Nakon analize se pristupa pripremi podataka. Pošto postoji 26 duplih uzoraka oni se uklanjaju iz baze. Uklanja se i deo podataka koji se smatra izuzecima, tačnije podaci desetog percentila i podaci iznad devedesetog percentila. Sledi transformacija kategoričkih obeležja za pol i informacija o konzumiranju alkohola u brojne vrednosti. Uklanjaju se obeležja o sluhu i vidu iz razloga što za zabeležene vrednosti vida nedostaju jedinice, dok su kategorije koje opisuju sluh nedovoljno precizno definisane. Dalje se pristupa transformaciji klasnih labela. Cilj je da se predvidi da li neko spada u kombinaciju originalnih klasnih labela tj. da li je pacijent pušač i da li konzumira alkohol. Od 3 klase pušača i 2 klase ljudi koji konzumiraju alkohol se pravi 6 klasa kombinujući svaku od njih i analiziraju se nove dobijene izlazne labele.



Slika 1 - ***Balansiranost klasa***

Primećuje se da su transformisane klase izuzento nebalansirane što može da oteža preciznu predikciju. Odlučeno je da se pacijenti koji su pušači i pacijenti koji su bivši pušači grupišu u istu klasu zajedno sa informacijom da li pacijent konzumira alkohol ili ne i na taj način se dobijaju 4 klase za koje će se vršiti predikcija. Pošto postoje podaci o težini i visini pacijenta, može se napraviti novo obeležje „BMI“ koje nam daje informaciju o indeksu telesne mase pacijenta.

IV. TRENIRANJE I EVALUACIJA MODELA

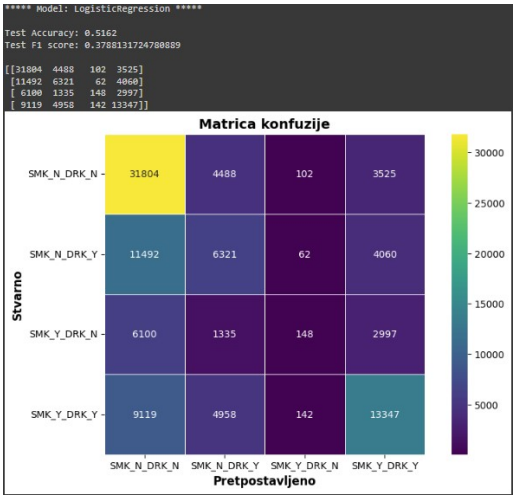
Podaci su podeljeni na obeležja i klasne labele i te se skupovi dodatno dele na trening skup i test skup. Baza podataka sadrži veliki broj podataka i iz tog razloga za test skup se uzima 25% podataka. Dalje se definišu modeli koji će da se koriste za predikciju, a to su: K najbližih suseda, logistička regresija i slučajna šuma i definišu se mogući hiperparametri za te modele.

Prelazi se na biranje parametara i evaluaciju modela korišćenjem unakrsne validacije. Trening podaci se dele na 5 delova (*engl. folds)*. Za svaki od delova se pravi novi trening set i validacioni set čije se vrednosti standardizuju. Na obeležja se primenjuje i LDA redukcija dimenzionalnosti zbog brzine modela. Vrši se obuka i validacija modela sa svakim od zabeleženih hiperparametara. Računa se tačnost modela i F1 rezultat. Računa se prosečna vrednost tačnosti za svaki od 5 delova. Modeli sa najboljom prosečnom vrednošću tačnosti se biraju kao pobednici i njihovi hiperparametri se čuvaju za dalji rad.

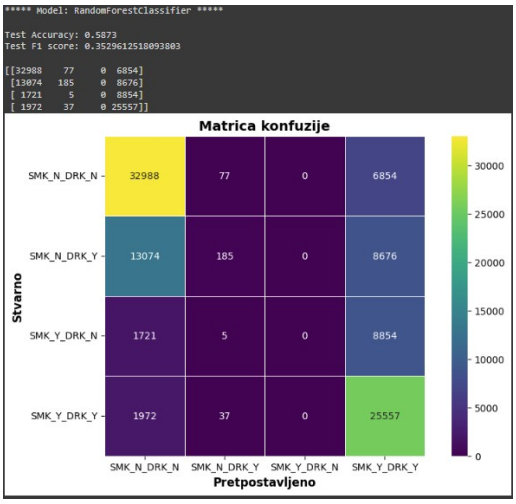
V. PREDIKCIJA TEST UZORAKA I PRIKAZ REZULTATA

Sa odabranim hiperparametrima za svaki model prelazi se na predikciju klasa korištenjem podataka iz test skupa. Pre obuke modela vrednosti obeležja se standardizuju i vrši se redukcija dimenzionalnosti. Obrađeni podaci se koriste za obuku modela, dok se za hiperparametre koriste vrednosti odabrane unakrsnom validacijom.

Za dobijene rezultate se računa tačnost i F1 rezultat. Dobijeni rezultati se ispisuju, a rezultati se prikazuju kao matrica konfuzije gde jasno možemo da vidimo broj uzoraka koji su pravilno i nepravilno klasifikovani i u koje klase su svrstani. Postignuta tačnost modela je između 50% i 60%. Analizirajući matrice konfuzije primećuje se da modeli dobro klasifikuju dominantnu klasu, što govori da su klase i dalje nebalansirane.



Slika 2 - ***Matrica konfuzije modela logističke regresije***



Slika 3 - ***Matrica konfuzije modela slučajne šume***