REKAPITULACIJA

DARKO SVILAR,
SV50/2021

Statistička analiza podataka

Analizom raspodela podataka zasebnih obeležja u skupu podataka, pokazalo se da je oko 14,5% unetih vrednosti za holesterol pacijenta neispravno, jer je nemoguće da je njihov holesterol bio 0 mg/dl.

Dalje, vrednosti oldpeak obeležja sadrže oko 1% negativnih vrednosti, što je nemoguće, obzirom da ovo obeležje opisuje prethodnu amplitudu ST talasa pacijenta, koja je nenegativna.

Isto kao i za holesterol, procenat pacijenata sa unetom vrednošću 0 za šećer u krvi je oko 0.09%.

Nasuprot pretpostavljenom, što je pokazala matrica korelacije na osnovu koje je određena korelacija svakog obeležja sa ciljnim obeležjem, najznačajnija obeležja nakon pretprocesiranja podataka su ST slope, excercise angina i chest pain type, tim poretkom, prema nerastućim vrednostima korelacije ovih obeležja sa ciljnim obeležjem.

Pretprocesiranje podataka

Kako skup podataka nije sadržao nula vrednosti, vrednosti koje nisu numeričkog tipa, niti duple redove, nije bilo potrebe za njihovim uklanjanjem.

Sve pogrešne vrednosti u poglavlju iznad, zamenjene su srednjim vrednostima za njihova obeležja, respektivno.

Podaci su, takođe, pre treniranja modela nad njima, skalirani, što nije eksplicitno navedeno u predlogu projekta.

Rezultati evaluacije modela

Nakon treniranja svakog modela i vršenja klasifikacije pomoću njih, ispostavilo se da je najkvalitetniji model, prema pomenutoj najbitnijoj metrici, tačnosti, random forest model sa 100 stabala.

Kako je gorepomenuti model bio kvalitetniji od ostalih i po svakoj drugoj metrici, pored tačnosti, nema potrebe za pominjanjem rezultata tih ostalih modela.

Random forest model sa 100 stabala je ostvario u proseku oko 94% tačnosti, preciznosti, odziva i F1 metrike.Ovo je vrlo dobar rezultat, jer se po vrednostima preciznosti i odziva vidi da model u 94% slučajeva tačno identifikuje prave pozitivne slučajeve, u odnosu na lažne pozitivne slučajeve i u 94% slučajeva tačno identifikuje prave pozitivne slučajeve, u odnosu na lažne negativne slučajeve. F1 metrika pokazuje da će model u vrlo velikom procentu slučajeva ostvariti visoku preciznost i odziv, što se može videti i na matricama konfuzije priloženim uz rešenje problema.

Rekapitulacija najznačajnijih obeležja

Nakon treniranja i testiranja najkvalitetnijeg modela, izvučene su važnosti obeležja tokom čitavog procesa klasifikacije nad datim skupom podataka.

Rezultati nalažu da je za bolju klasifikaciju potrebno obeležje koje će mnogo češće ukazivati na postojanje srčanog oboljenja, jer nijedno od postojećih obeležja nema dovoljno veliku važnost (ST slope sa samo ~0.16 je na vrhu liste važnosti).

Takođe se ispostavilo da obeležje excercise angina nema toliko veliku važnost kao što je pokazala matrica korelacije, jer se nalazi tek na 8. mestu na listi važnosti, dok max heart rate ima mnogo veću važnost, jer se našla čak na 2. mestu na listi važnosti.