Univerzitet u Beogradu -Elektrotehnički fakultet
Katedra za Signale i sisteme

Sistemi odlučivanja u medicini

- Projektni zadatak -

Detekcija Parkinsonove bolesti iz glasa

STUDENTI

Marko Macura 2018/0261,
Kristina Nikolić 2018/0245

Septembar 2021.

Uvod i baza podataka

U ovom projektu korišćena je baza podataka: "Parkinsons Disease Data Set". Ovaj set podataka sadrži obeležja koja predstavljaju biološke parametre dobijene procesiranjem snimka govora potencijalnih obolelih od Parkinsonove bolesti. Glas je sniman 31 osobi, od kojih je 23 obolelo od Parkinsonove bolesti. Za svaku osobu snimanje je ponovljeno 5-6 puta, što je u zbiru dalo 195 uzoraka. Iz snimaka je izvučeno 22 atributa koji uglavnom opisuju parametre glasa kao što su frekvencija i njena varijacija.

Obeležja data u bazi podataka su:

- MDVP:Fo(Hz) - Prosečna govorna frekvencija
- MDVP:Fhi(Hz) - Prosečna govorna frekvencija
- MDVP:Flo(Hz) - Minimalna govorna frekvencija
- MDVP:Jitter, MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP - Nekoliko mera za varijaciju govorne frekvencije
- MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA - Nekoliko mera za varijaciju u amplitudi
- NHR, HNR - Dve mere odnosa šuma prema tonskim komponentama u glasu
- status - Zdravstveni status osove 1 - Parkinson, 0 - healthy
- RPDE,D2 - Dve linearne mere dinamičke složenosti
- DFA - Eksponent fraktalnog sklairanja signala
- spread1, spread2, PPE - Tri nelinearne mere varijacije fundamentalne frekvencije

Biranje obeležja

Obeležja smo birali tako što smo gledali da budu što manje međusobno korelisana, ali da što više budu korelisana sa izlazom (da li osoba boluje od Parkingsonove bolesti). Određena je korelaciona matrica svih obeležja i izabran je onaj podskup koji najviše zadovoljava željeni kriterijum.
Kao što je očekivano, iz grupe metrika koje opisuju isti parametar dovoljno je uzeti u obzir jedan od njih, jer su oni jako koreliani medju sobom. Takođe, odabrani su i minimalna i prosečna govorna frekvencija, jer zajedno daju informaciju o odstupanju ovog parametra od njegove srednje vrednosti, što se možda može dovesti u vezu sa karakteristikama Parkinsonove bolesti. Na kraju, parametri koji odredjuju odnos šuma prema tonskim komponentama u glasu (NHR i HNR) nisu odabrani jer nisu imali značajnu korelaciju sa statusom ispitanika.

Korelaciona analiza i information gain

Kako su skoro svi atributi kontinualni, da ne bi njihova prosečna veličina uticala na dalji radi, normalizovali smo ih (podelili sa maksimalnom vrednošću u okviru atributa) i zaokružili na 3 značajne cifre. Na taj način smo dobili sledeće informacione dobiti:

$MDVP:Fo(Hz)$	$MDVP:Flo(Hz)$	MDVP:Jitter(Abs)	MDVP:Shimmer	MDVP:APQ
0.6859	0.7358	0.2687	0.5948	0.6987
RPDE	DFA	spread1	spread2	D2
0.6884	0.6217	0.6050	0.7538	0.6602

Za korelacionu analizu korišćena je sledeća formula:

$$r = \frac{k \cdot r_{zi}}{\sqrt{k + k(1+k)r_{ii}}}$$

gde je k ukupan broj atributa (k=10), r_{ii} je srednja međukorelacija između atributa i r_{zi} je srednja korelacija između atributa i klase. Kao rezultat dobijeno je $r = 0.4176$.

Redukcija dimenzija

Podelili smo na testirajući i obučavajući skup holdout metodom tako što smo 80% podataka iskoristili za obučavanje, a ostatak za testiranje.

Dimenzije smo redukovali LDA metodom. Odlučili smo se za dve dimenzije jer su sve sopstvene vrednosti osim jedne približno jednake nuli.Da bismo redukovali dimenzije LDA metodom morali smo da definišemo sledeće matrice:

$$S_w = \sum_{i=1}^{L} P_i \cdot \Sigma_i$$

$$S_B = \sum_{i=1}^{L} P_i \cdot (M_i - M_0) \cdot (M_i - M_0)^T$$

$$M_0 = \sum_{i=1}^{L} P_i \cdot M_i$$

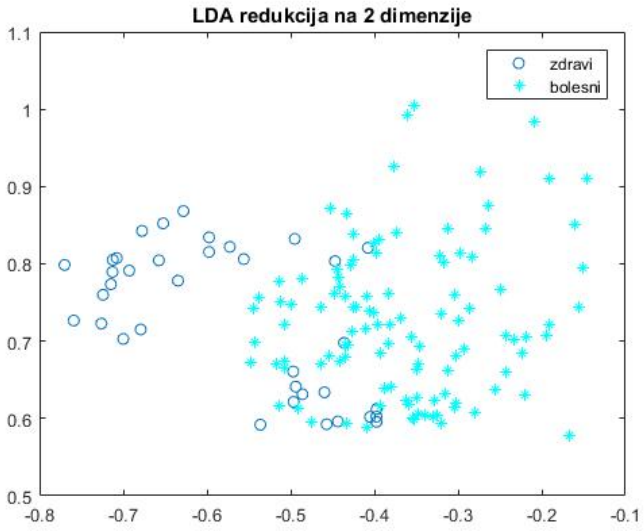
Gde je L broj klasa (u našem slučaju 2), P_i je apriorna verovatnoća klase i M_i i Σ_i su matrica matematičkog očekivanja i kovarijaciona matrica. Matrica S_w se zove matrica unutarklasnog rasejanja, a matrica S_B međuklasnog rasejanja. M_0 je združeni vektor matematičkog očekivanja.
Kod ove metode kriterijumska funkcija je $J = tr(S_M^{-1} S_B)$. Taj kriterijum se svodi na traženje maksimalnih sopstvenih vrednosti matrice $S_M^{-1} S_B$ i smeštanje sopstvenih vektora koji odgovaraju tim sopstvenim vrednostima u matricu Φ. Biramo onoliko sopstvenih vrednosti i vektora na koliko dimenzija želimo da redukujemo obeležja.
Formula kojom redukujemo obeležja:

$$Y = \Phi^T \cdot X$$

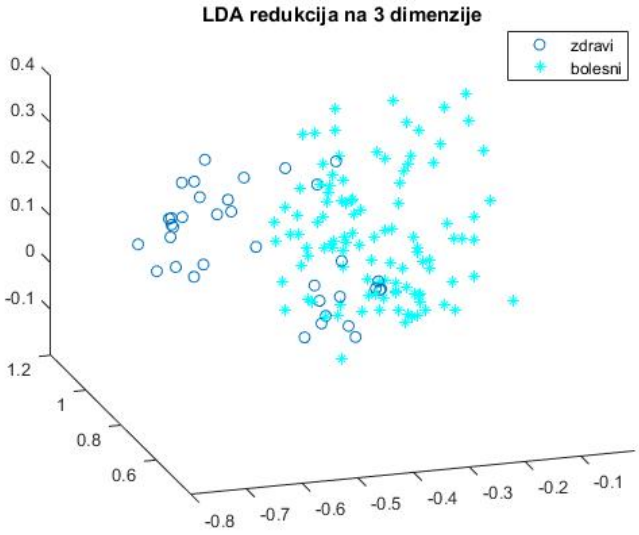
$$\Phi = [\Phi_1 \Phi_2 ... \Phi_m]$$

m - broj dimenzija, Φ_k - sopstveni vektor koji odgovara k-toj najvećoj sopstvenoj vrednosti, X - originalan skup podataka, a Y rezultat redukcije dimenzija.

Rezultat redukcije dimenzija je:



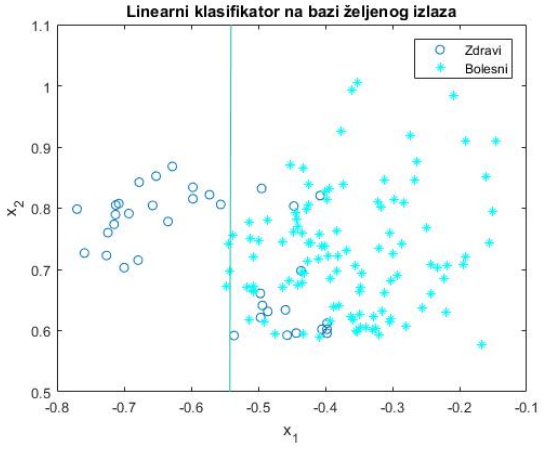
Kako su sve vrednosti sopstvenih vrednosti jako male (jedna je jednaka 0.7379, a ostale su reda 10^{-16}) jasno je da su podaci najviše rasejani u samo jednoj dimenziji. Kako bismo ilustrovali da je separabilnost na 2 dimenzije slična kao na 3 dimenzije prilažemo i grafik LDA redukcije dimenzija na 3 dimenzije:

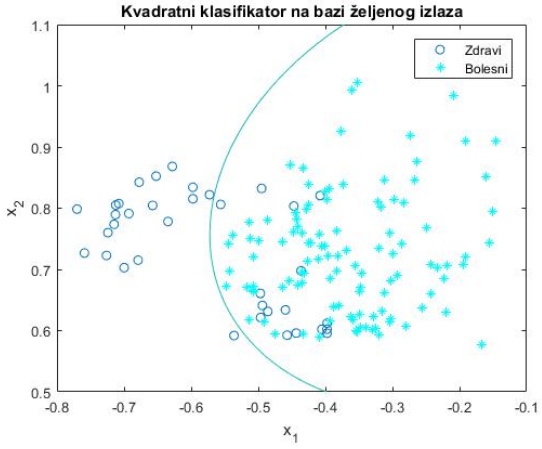


Klasifikacija

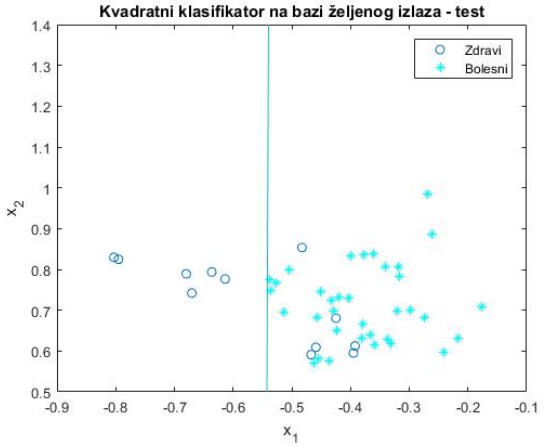
Parametarska klasifikacija

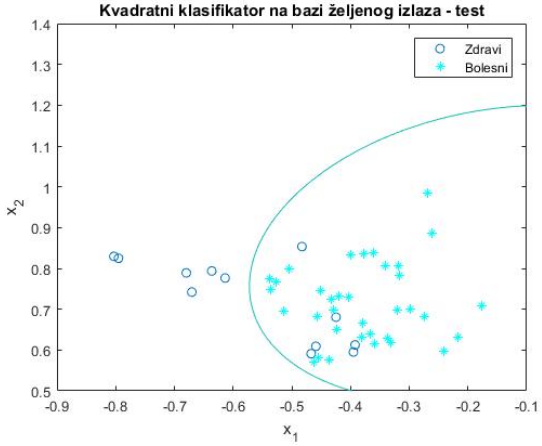
Napravili smo i linearni i kvadratni klasifikator kako bismo uporedili rezultate. Diskriminacione krive za ta dva klasifikatora su: $h(x) = v^T x + v_0$ i $h(x) = xQx^T + v^t x + v_0$. Nakon treniranja naših klasifikatora dobili smo:



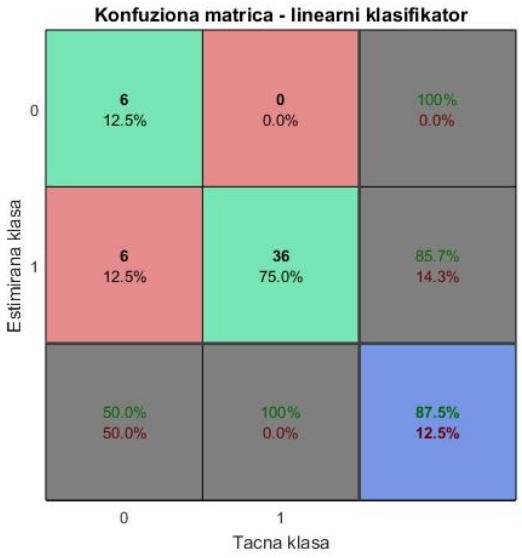


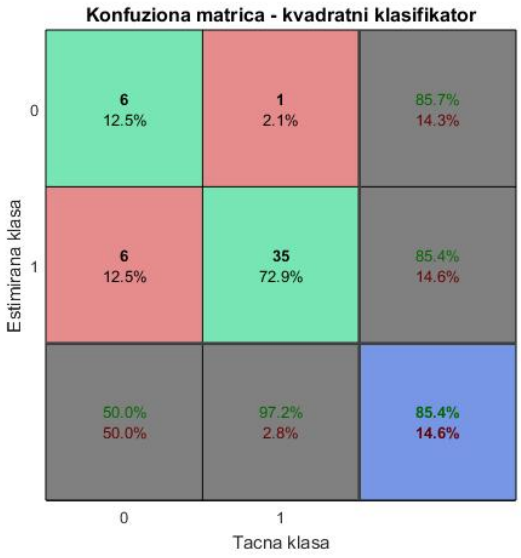
A rezultati na testirajućem skupu podataka su:



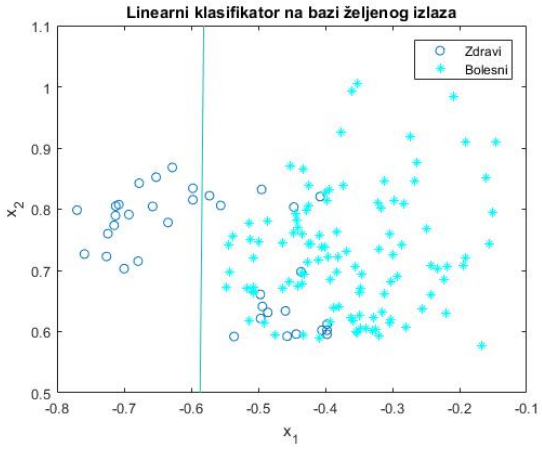


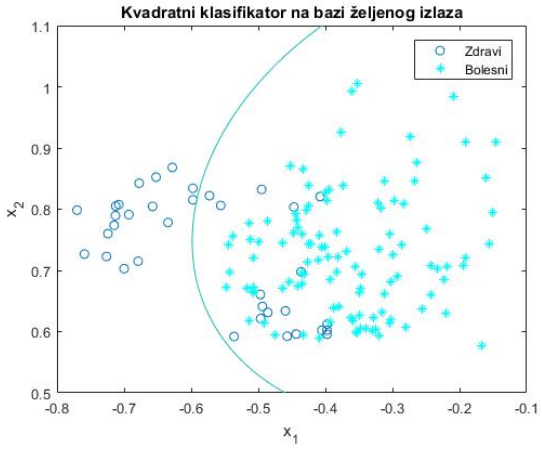
Konfuzione matrice:



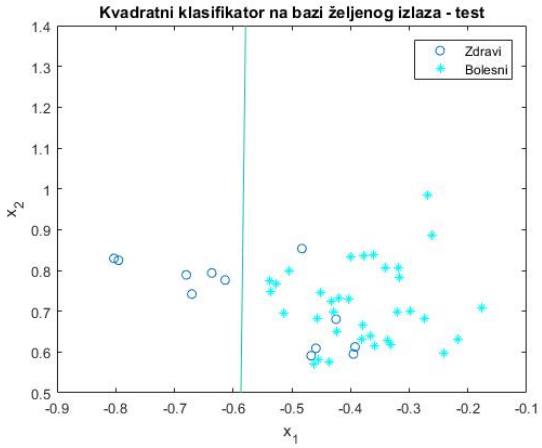


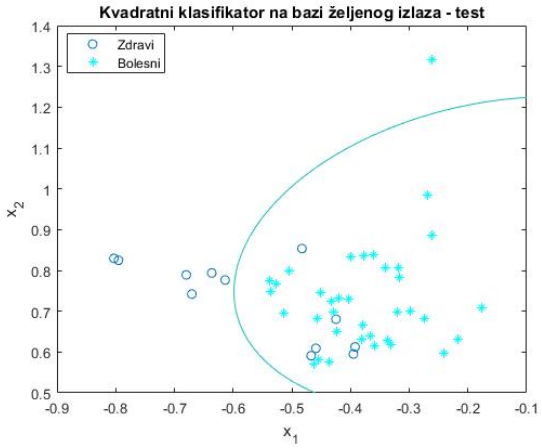
Kao što se vidi iz kofuzionih matrica, a i sa slika kvadratni klasifikator je malo bolji od linearnog. Ako uzmemo u obzir da je kvadratni klasifikator složeniji od linearnog i zahteva više vremena da bi se napravio dovodi se u pitanje koji klasifikator je bolje koristiti. Ukoliko nam treba brži odgovor tada je bolje koristiti linearni, ali u medicini češće nam trebaju tačniji podaci nego brzina, pa je po našem mišljenju bolje koristiti kvadratni.
U želji da smanjimo broj false negative pacijenata u matricu Γ umesto svih jedinica (sve odbirke uzimamo sa jednakom težinom), smo stavili za zdrave 0.7, a za bolesne 1. Kao rezltat dobili smo da su nam diskriminacione krive translirane ulevo, što se može videti na sledećim slikama.



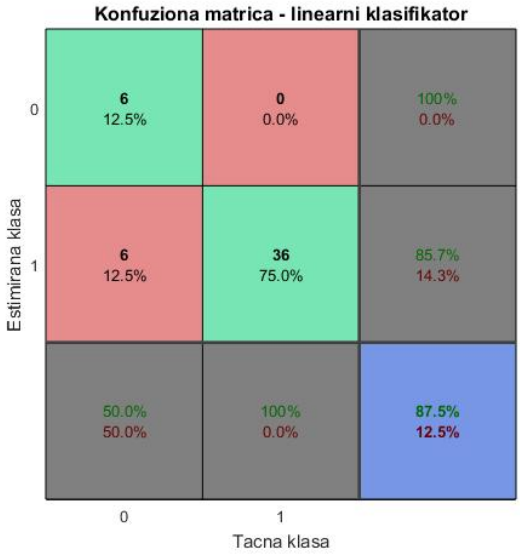


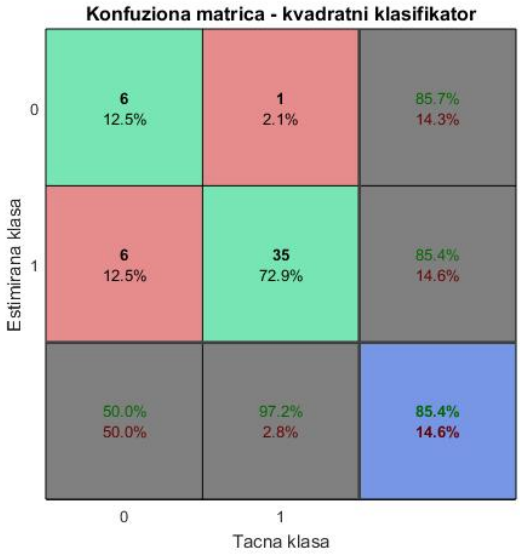
Na testirajućem skupu podataka:





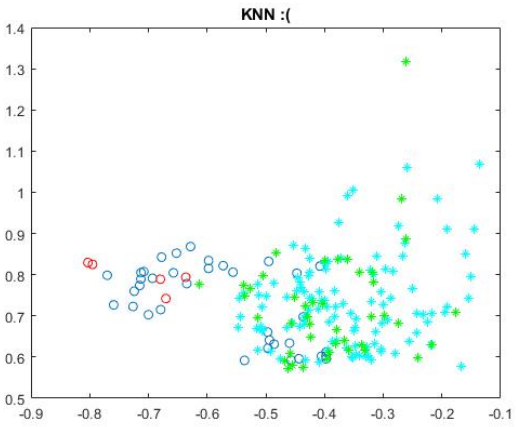
Konfuzione matrice:

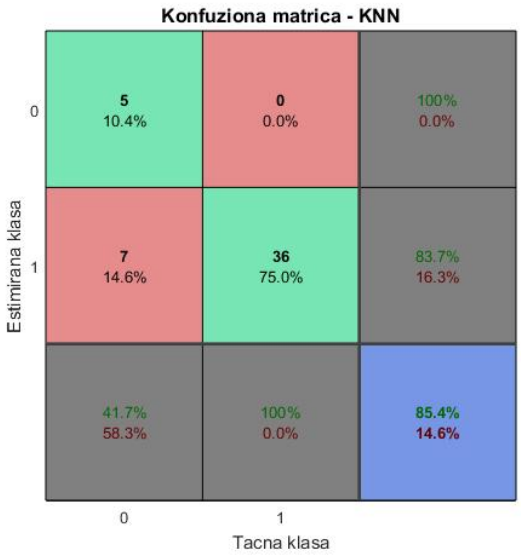




KNN klasifikator

Kako bismo doveli dobili širu sliku o klasifikacijama napravili smo i jedan neparametarski klasifikator. Kao funkciju $dist(A, B)$ koristili smo Euklidsko rastojanje.
Rezultati su sledeći:





Plave zvezdice na grafiku su bolesni iz obučavajućeg skupa, a zelene su klasifikovani kao bolesni. Plavi krugovi su zdravi iz obučavajućeg skupa, a crveni krugovi su klasifikovani kao bolesni. Ukoliko ima isti broj bolesnih i zdravih u Euklidskom rastojanju koje smo zadali, taj test smatramo pozitivnim kako bismo smanjili false negative slučajeve.

Neuralne mreže

Obučavane su neuralne mreže sa potpuno povezanim slojevima. Variran je broj skrivenih slojeva kao i broj čvorova po sloju. Ove mreže su obučavane na početnom skupu od 22 atributa.

Balansiranje skupa podataka

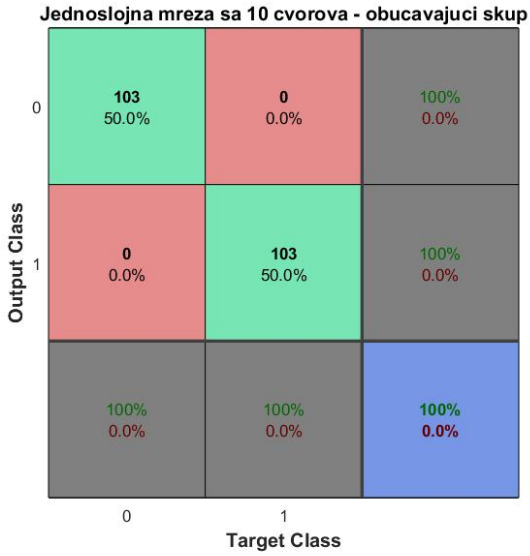
Početni odnos obolelih i zdravih u skupu je 70/30. Kako bi se posmatrao balansiran skup i nadomestio nedostatak podataka u slučaju zdravih osoba, skup je balansiran. Dodat su zašumljeni podaci generisani na osnovu varijanse i očekivanja svakog od atributa. Ovo nam omogućava da pokrijemo slučaj u kome bi se mreže koristile za obradu skupova koji su relatvno balansirani. Međutim, u većini primena, detekcija bolesti se vrši na skupovima koji imaju više pozivinih uzoraka, jer uglavnom postoji početna sumnja na datu bolest. Da bi se pokrila oba slučaja, mreže su trenirane i na balansiranom i na početnom skupu podataka.

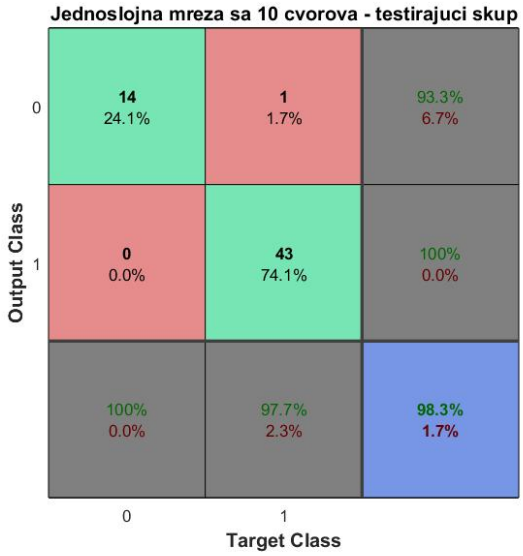
Jednoslojna neuralna mreža

Najpre je formirana neuralna mreža sa jednim skirvenim slojem. Broj čvorova u sloju je variran kako bi se uočio opseg koji daje najbolje rezultate. Trenirane su mreže sa $1, 3, 7, 10, 15$ i 25 čvorova u skrivenom sloju. Pri obučavanju neuralne mreže sa jednim skrivenim slojem, za aktivacionu funkciju skrivenog sloja korišćena je logsig funkcija, dok je za izlazni sloj korišćena bipolarna linearna aktivaciona funkcija sa zasićenjem (satlins).

10 čvorova

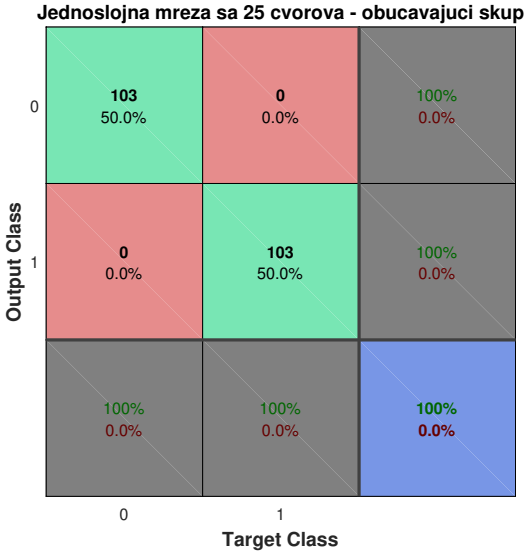
Neuralna mreža sa 10 čvorova u skrivenom sloju dala je najbolje rezultate za balansirane podatke. Rezultati su prikazani konfuzionim matricama. Možemo primetiti da je mreža dostigla tačnost od 100% na trenirajućem skupu, dok je na testirajućem tačnost 98.3%. Bitno je napomenuti da je false negative!!!!!! rate 93.3. U medicinskim problemima, ovo je jedna od važnijih metrika, jer propuštanje pozitivnog ishoda dovodi do negativnih posledica.

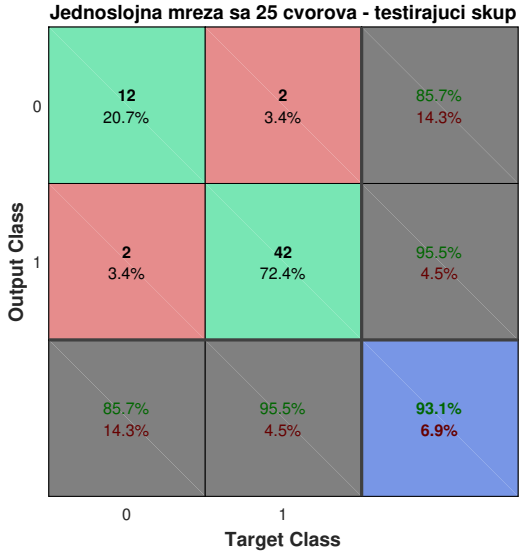




25 čvorova

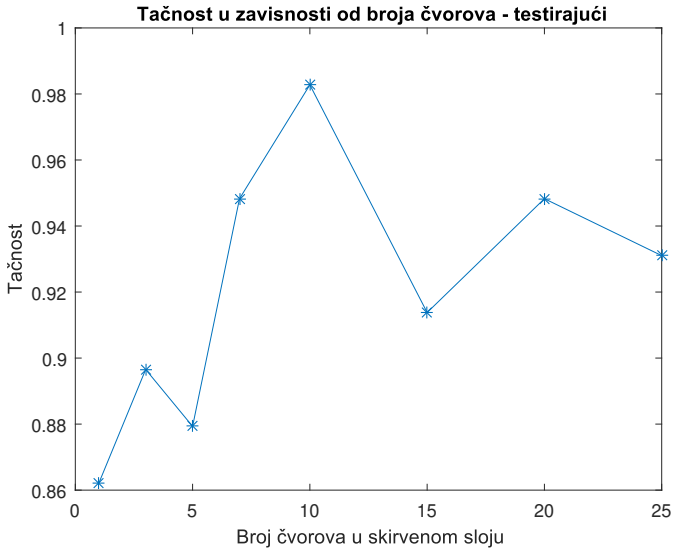
Kod jednoslojne mreže sa 25 čvorova može se primetiti da je tačnost opala, što može da znači da je došlo do preobučavanja mreže. Rezultati su prikazani konfuzionim matricama.





Poređenje broja slojeva

Neuralna mreža je trenirana sa više različitih broejva čvorova u skirvenom sloju. Grafik zavisnosti tačnosti od broja čvorova prikazan je na slici. Možemo primetiti da tačnost raste do 10 čvorova, kada je najveća, a zatim krene da opada jer je došlo do preobučavanja mreže.



Višeslojna mreža sa regularizacijom

Trenirane su mreže sa više od jednog skrivenog sloja. Mreže sa 2 i 3 sloja davale su približno iste rezultate, a mreže sa više od 3 sloja nisu davale poboljšanje. Ove mreže davale su lošije rezultate od mreža sa jednim skirvenim slojem, što je verovatno rezultat preobučavanja.
Pri obučavanju neuralne mreže sa više skrivenih slojeva, za aktivacionu funkciju skrivenih slojeva korišćena je

tansig funkcija, dok je za izlazni sloj korišćena bipolarna linearna aktivaciona funkcija sa zasićenjem (satlins).

Kako bi se sprečilo preobučavanje, dodata je regularizacija. Ovime je u standardnu kriterijumsku funckiju dodat član koji sprečava da težine u čvorovima previše porastu i tako dovedu do preobučavanja. Sada kriterijumska funkcija ima sledeći oblik:

$$E(w) = \gamma mse + (1-\gamma)msw$$

$$msw = \frac{1}{N}\sum_{i=1}^{N} w_j^2$$

Prikazane su konfuzione matrice pre i nakon dodate regularizacije. Možemo primetiti da su rezultati nakon dodate regularizacije bolji i ako su oba modela ostvarila maksimalnu tačnost na trenirajućem skupu. Prikaz je za nebalansirane podatke.

Troslojna mreža pre (levo) i posle regularizacije (desno)

Pošto je skup nebalansiran, određena je i balansirana tačnost po formuli: $\frac{(\frac{TP}{P}+\frac{TN}{N})}{2} = 85\%$

subsectionRano zaustavljanje

Još jedna metoda za zaštitu od obučavanja je rano zaustavljanje. Ukoliko obučavanje traje previše dugo mreža će početi da uviđa detalje koji važe samo u trenirajućem skupu i očekivaće da to važi i za sve skupove na kojima ona bude primenjivana. Pošto to nije slučaj, zaustavićemo mrežu dovoljno rano da bi mogla da generalizuje. Skup podataka delimo na obučavajući, validacioni i trenirajući. Mrežu treniramo na obučavajućem, dok pratimo vrednost karakteristične funcije na validacionom skupu. Kada greška na validacionom počne da raste, znak je da mreža ulazi u period preobučavanja i treba zaustaviti trening.

Rezultati za troslojnu mrežu sa ranim zausavljanjem prikazani konfuzionom matricom. Možemo primetiti da model nije uspeo da završi treniranje i dostigne tačnost od 100% kao u predhodnim slučajevima, ali da mu je rezultat na testirajućem skupu bolji. Ovo važi i za tačnost i za specifičnost. Ovakva mreža će davati bolje rezultate na nepoznatim skupovima podataka. Performanse obučavanja prikazane su na slici.

Troslojna mreža pre (levo) i posle regularizacije (desno)

Troslojna mreža pre (levo) i posle regularizacije (desno)