Projektni zadatak iz Sistema odlučivanja u medicini

Nemanja Saveski 2019/0056

26. avgust 2022.

Transplantacija koštane srži kod dece

1 Uvod i analiza skupa podataka

U ovom zadatku, primarni cilj naučnika je bila komparacija rezultata unrelated donor transplantacije[1] ćelija koštane srži u odnosu na to kako je obavljena transplantacija[2], kao i ispitivanje poboljšanih rezultata transplantacije zbog povećanih doza CD34+ i CD3+ matičnih ćelija u odnosu na one pacijente koji su dobili normalne doze. Takođe je utvrđeno da li dolazi do akutnih ili ekstenzivnih hroničnih Graft versus Host oboljenja[3](u nastavku GvHD). Pacijenti su deca (njih 187), starosne dobi do 15 godina, sa malignim ili benignim oboljenjima. U bazi[4] postoji 38 atributa i jos jedno poslednje obeležje koje predstavlja pripadnost jednoj od dve klase (pacijent preživeo/nije preživeo). Primećuje se da postoje neke grupe atributa u kojima atributi mnogo zavise jedni od drugih, tako da je pri selekciji obeležja, u stvari, prvo ceo skup obeležja podeljen na više manjih podgrupa iz kojih su intuitivno izvučeni oni koji najviše informacija mogu da donesu kasnije u projektovanju klasifikatora. Tako imamo:

ANC i PLT recovery [5] ANCrecovery - izbačeno 5 pacijenata; PLTrecovery prebačen iz kvantitativne u ordinalnu kategoričku promenljivu (1 - oporavljeno, 0 - nije)

GvHD time_to_aGvHD_III_IV - vreme nastanka akutnog GvHD, po meni nepotrebno, izbačeno; IIIV - sadržano u sledeća dva atributa, izbačeno; aGvHDIIIIV - indikacija akutnog GvHD (0 - nema, 1 - ima; u startu je bilo obrnuto); extcGvHD - indikacija ekstezivnog hroničnog GvHD, sadržalo je upitnike, zamenjeno na sledeći način:

- Ako je došlo do akutnog GvHD, onda sigurno nije do ekstenzivnog hroničnog $\rightarrow 0$
- Ako nije došlo do akutnog GvHD, pacijent nije preživeo, trombociti se nisu oporavili, opet $\rightarrow 0$
- Ako nije došlo do akutnog GvHD, pacijent nije preživeo, ali trombociti su se oporavili, onda $\rightarrow 1$

Obeležja vezana za starost Recipientageint - podela pacijenata u 3 intervala po godinama (0:5, 5:10, 10:15), izbačeno; Recipientage10 - podela pacijenata (<10god/>10god), izbačeno; Donorage35 - podela donora (<30god/>30god); Recipientage - godine pacijenta; Donorage - godine donora

[1] transplantacija u kojoj donor i pacijent nisu u srodstvu

[2] 2 načina - 1. matične ćelije iz koštane srži; 2. matične ćelije iz perifernog krvotoka

[3] oboljenja gde ćelije primaoca vide transplantovane ćelije kao opasnost i pokušavaju da se odbrane tako što ih napadaju

[4] Bone marrow transplant: children Data Set

[5] ANC - neutrofili, PLT - trombociti

HLA [6] HLAmismatch - neuklapanje antigena na belim krvnim zrncima, izbačeno; HLAgrI - poklapanje antigena na nerazuman način (brojevima 0-7, iako u specifikacijama baze piše 0-5), izbačeno; Antigen i Alel - potpuno poklapanje bilo jednako -1, nameštено da potpuno poklapanje bude jednako 0 ($+=1$ sve vrednosti)

CMV [7] cela grupa obeležja (CMVstatus, DonorCMV, RecipientCMV) je izbačena jer je nedostajalo dosta vrednosti (oko 10%), a same promenljive su bile kategoričke tako da nije bilo moguće zameniti nekim drugim vrednostima

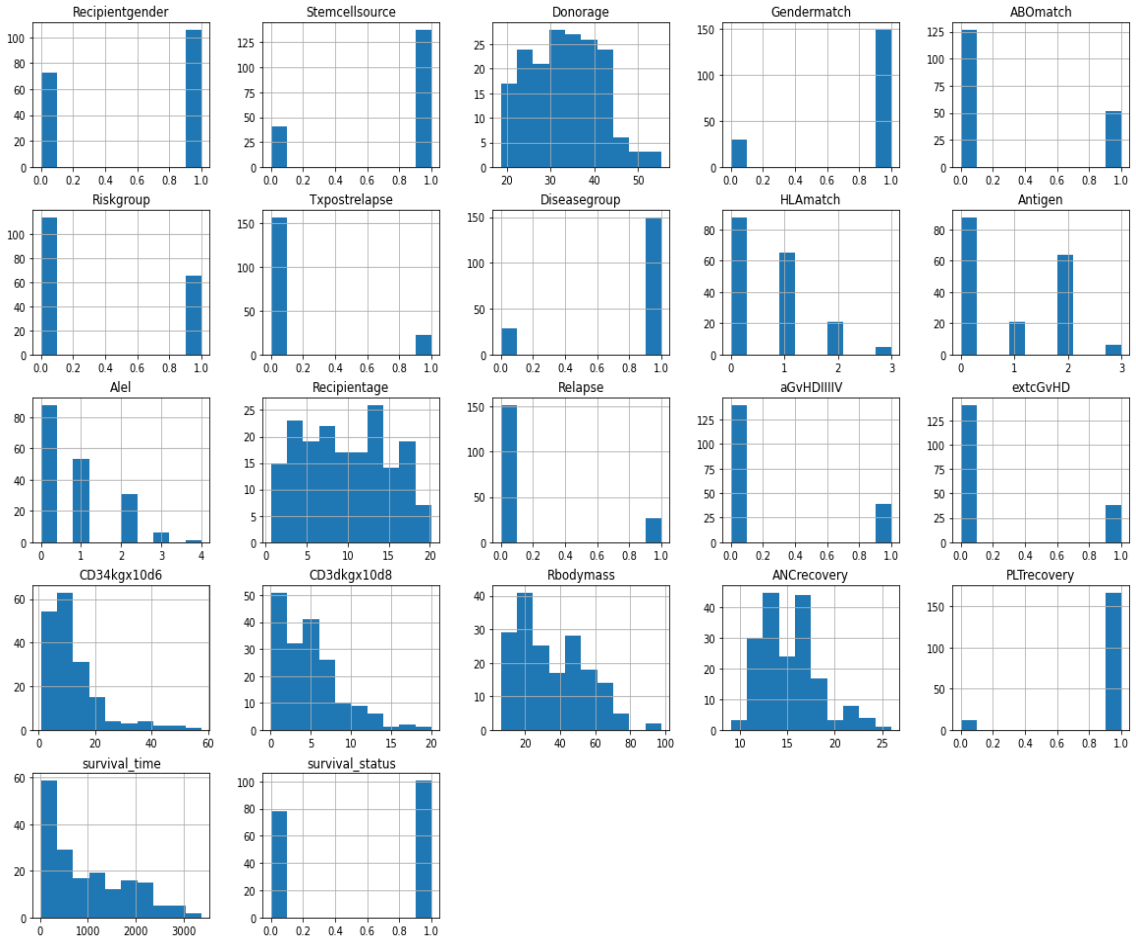
Krvna grupa Iskorišćen samo atribut poklapanja krvne grupe - ABOmatch, ostali su izbačeni (DonorABO, RecipientABO, RecipientRh)

Doze CD34+ i CD3+ CD3dCD34 - odnos datih ćelija, izbačen; CD34kgx10d6 - broj CD34+ ćelija u 10^6 po kilogramu; CD3dkgx10d8 - broj CD3+ ćelija u 10^8 po kilogramu, nepoznate vrednosti zamenjene medijanom zbog outliera

Ostalo Recipientgender - pol pacijenta (0 - žensko, 1 - muško); Stemcellsource - poreklo matičnih ćelija; Gendermatch - poklapanje polova donora i pacijenta; Riskgroup - visoko rizična grupa (1 - da, 0 - ne); Txpostrelapse - transplantacija nakon povraćaja bolesti; Diseasegroup - tip oboljenja (1 - maligni, 0 - benigni); Rbodymass - BMI, nepoznate vrednosti zamenjene medijanom jer je bilo outliera; Relapse- povraćaj bolesti; survival_time - vreme preživljavanja

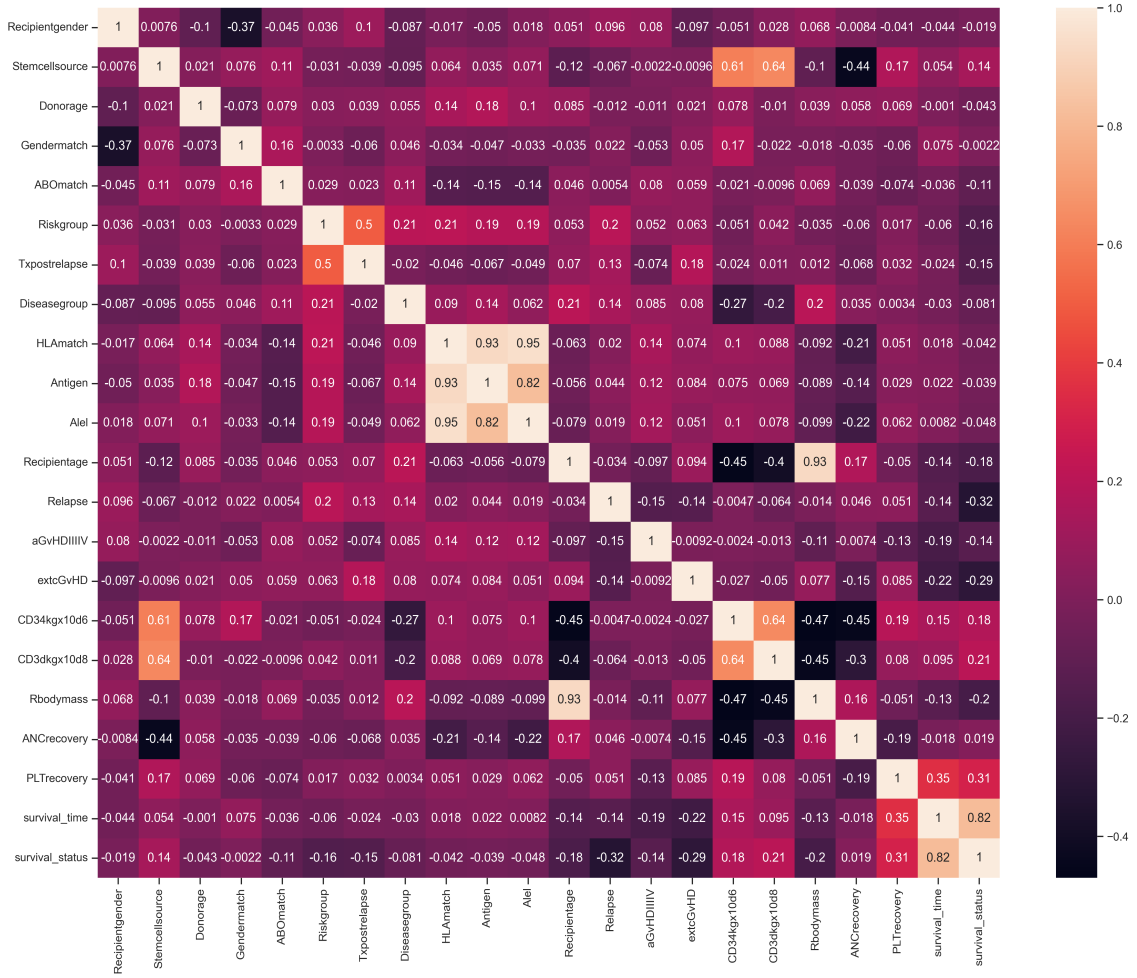
[6] HLA - Human leukocyte antigen
[7] cytomegalovirus, zaraza

Nakon odgovarajućih izmena (i izbacivanja jos 3 pacijenta kojima mnogo vrednosti fali) histogram atributa izgleda ovako:



2 Korelacija i Information gain

Prvo se može pokazati korelacija između preostalih obeležja u bazi (21). Izabrani metod je Spermanov (iako je pokrenut i Pearsonov i daje veoma slične rezultate) jer daje ekstremnije vrednosti tj. veće korelacije.



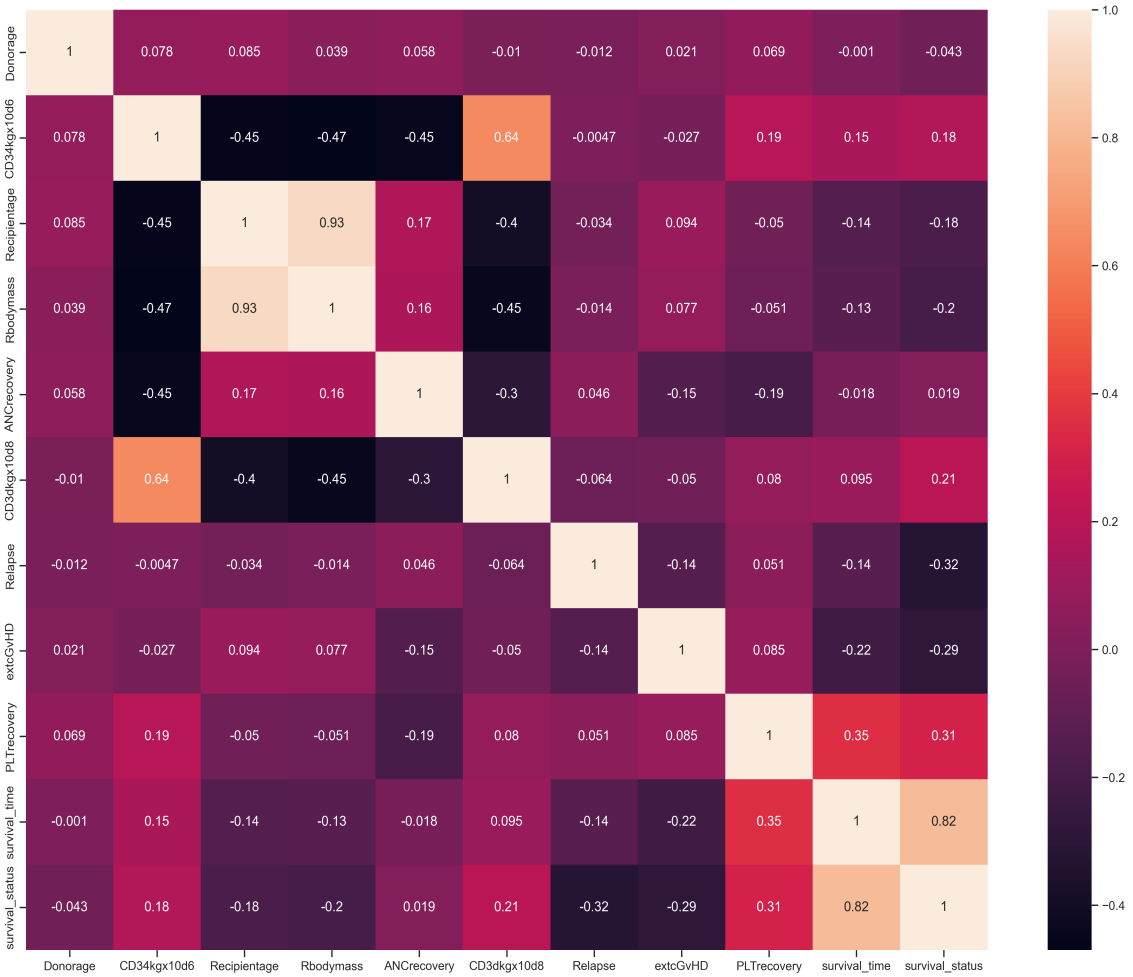
Postoje neke korelacije između atributa i obeležja klase, kao i korelacije med ˙u atributima iz iste grupe atributa, ali jako je teško i dalje spoznati koji atributi su nam bitni. Zato ćemo odmah preći na Information gain. Optimalna vrednost za broj binova na histogramima je 25 (u okviru kojih će se posmatrati podaci pri računu).

10 atributa sa najvećim information gainom su:

1. $survival\_time \rightarrow \text{IG} = 0.75434$
2. $Donorage \rightarrow \text{IG} = 0.16100388$
3. $CD34kgx10d8 \rightarrow \text{IG} = 0.117488$
4. $Recipientage \rightarrow \text{IG} = 0.1157216$

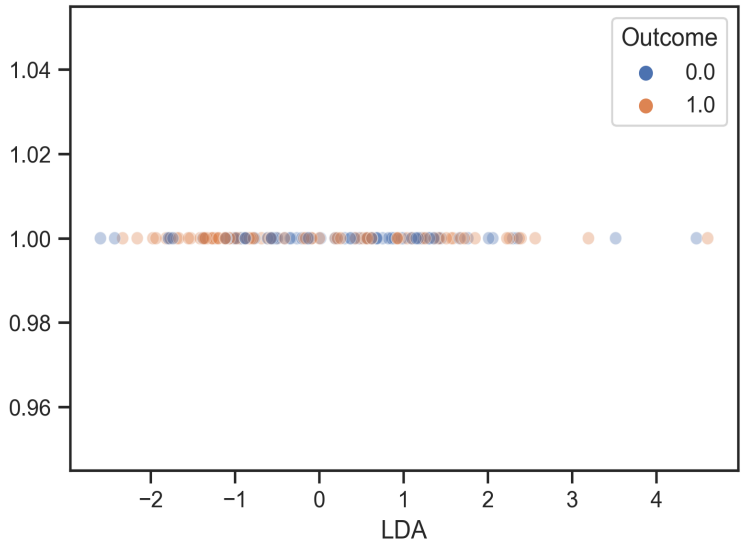
5. $Rbodymass \rightarrow \text{IG} = 0.11469766$
6. $ANCrecovery \rightarrow \text{IG} = 0.10642599$
7. $CD3dkgx10d8 \rightarrow \text{IG} = 0.08971392$
8. $PLTrecovery \rightarrow \text{IG} = 0.08487735$
9. $Relapse \rightarrow \text{IG} = 0.07754614$
10. $extcGvHD \rightarrow \text{IG} = 0.06011445$

Nijedan od ovih atributa nema preterano veliku vrednost information gaina, ali vidi se da je bitan broj godina kako donora tako i pacijenta. Takođe obe povećane doze matičnih ćelija imaju veliki uticaj na preživljavanje pacijenta (to je stvar koju smo mogli da izvučemo i sa korelacione tablice). Veliki deo ovih atributa će zatrebati za kasnije. Korelaciona tablica sa 10 atributa najvećeg information gaina sada izgleda (i dalje Spearman):

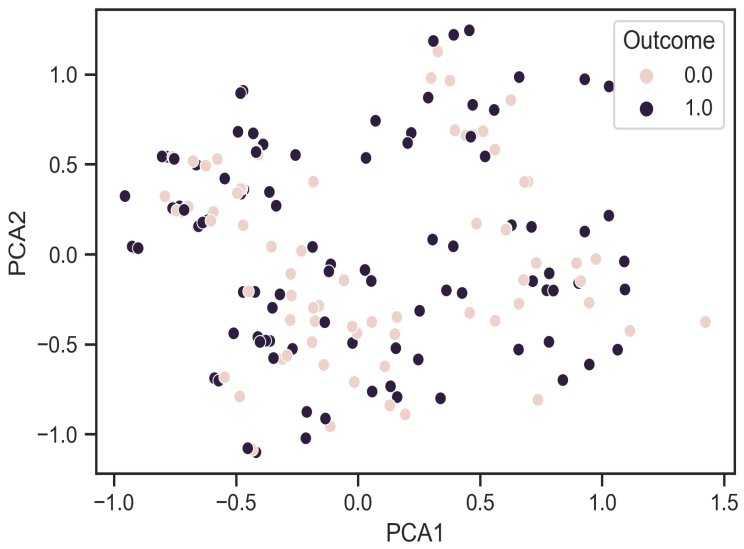


3 LDA i PCA metode

Pošto je ovo problem dve klase, nesmisleno je raditi LDA redukcije na dve ili tri dimenzije (može se redukovati u najviše $\min(\texttt{br\_klasa} - 1, \texttt{br\_obelezja})$). LDA urađen na dvoklasnom problemu stvara jednu osu na koju se nagomilavaju odbirci iz p-prostora (gde je p broj obeležja) tako da se napravi što veće rastojanje izmed ̄u centara dve klase ali da se minimizuje unutarklasna varijansa. Tako da se LDA može u ovom problemu iskoristiti isključivo da problem postane 1D:

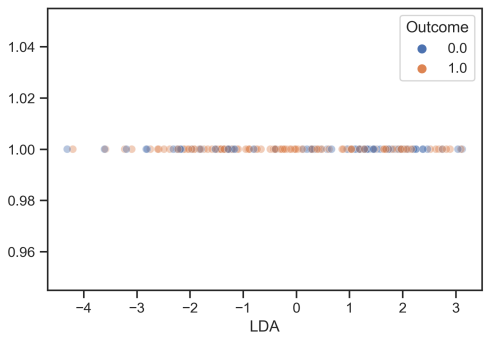


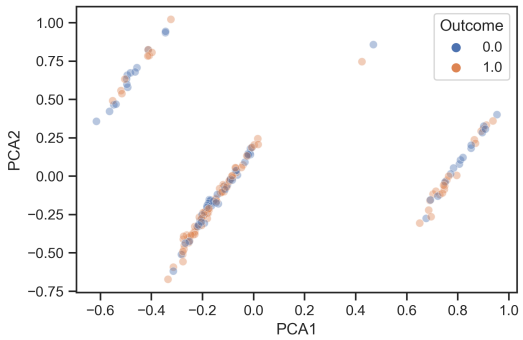
Ovde ne možemo napraviti nikakvu seprabilnost, klasifikator bi imao veoma lošu tačnost. Rezultati za 2D PCA metodu su takođe loši, pa i ovde nemamo separabilnost:



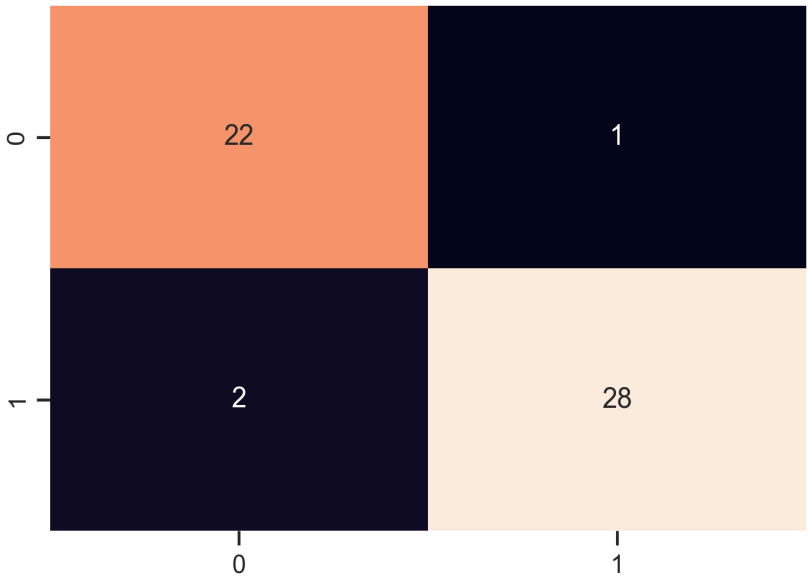
4 Klasifikator

Zbog veoma loših rezultata LDA i PCA redukcije ne može se odmah iskoristiti parametarski klasifikator nego se prvo moraju izbaciti neka obeležja. Može se pokušati izbaciti neki broj atributa pa onda uraditi LDA i PCA. Ali ni u tim slučajevima se ne dobija ništa smisleno:





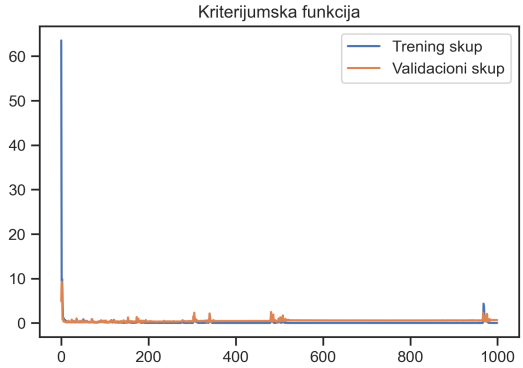
Korisitćemo deo kombinacije atributa sa najvećim information gainom od ranije (Donorage, CD34kgx10d8, Recipientage, Rbodymass, CD3dkgx10d8, Relapse, extcGvHD, survival_time) i napravićemo **Bajesov klasifikator** (verovatno bi prošao i sa minimalnim cenama ali ovde je rađen običan jer ima malo instanci (pacijenata)). Odnos obučavajućeg i test skupa je **70/30**. Dobija se konfuziona matrica:

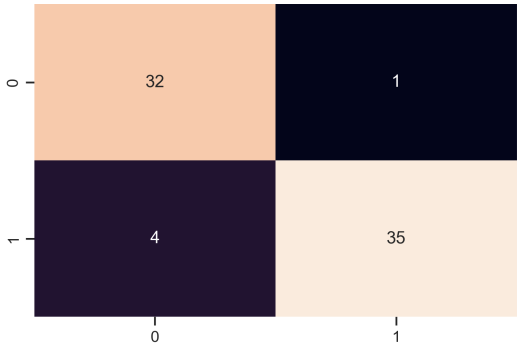


Tačnost ovog klasifikatora je **94.34%**. Treba napomenuti da je velika tačnost posledica dobre korelacije atributa survival_time sa obeležjem klase. Ako se survival_time izbaci iz baze, dobija se tačnost klasifikatora **77%**, što nije toliko loše računajući da smo imali samo 187 pacijenata, a 38 atributa.

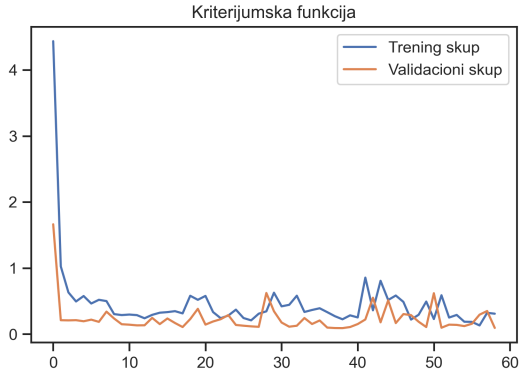
5 Neuralna mreža

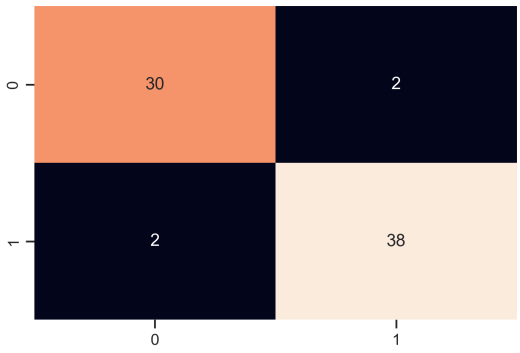
Na kraju analizirana je klasifikacija klasičnim *feedforward* neuralnim mrežama, sa različitim brojem skrivenih slojeva, kao i različitim brojem neurona u slojevima (negde je broj neurona u skrivenim slojevima isti, a negde različit). Za početak imamo neuralnu mrežu sa 1 skrivenim slojem od 200 neurona, na kojoj će se takođe prikazati i zaštita od preobučavanja ranim zaustavljanjem, kao i L2 regularizacijom. Treba napomenuti i da je za svaku neuralnu mrežu odnos odbiraka u obučavajućem i testirajućem skupu **60/40**. Kriterijumska funkcija i konfuziona matrica date neuralne mreže:



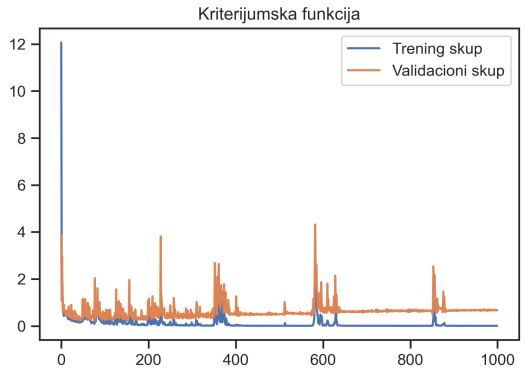


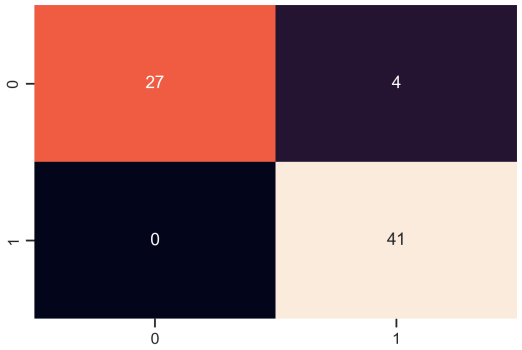
Dobija se tačnost **93.05%** što nije toliko loše u odnosu na to koliki je skup podataka. Uvođenjem mehanizma **ranog zaustavljanja** neuralna mreža se zaustavlja u **62.** epohi i dobija se tačnost **94.44%**, što je bolje nego ranije:



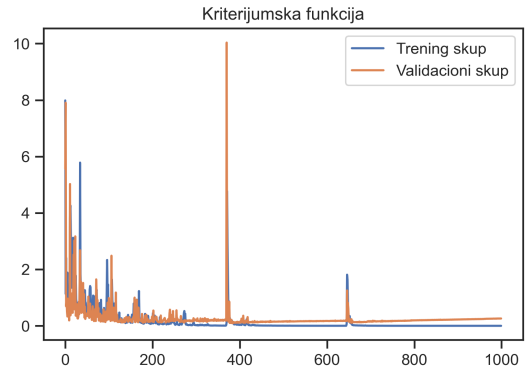


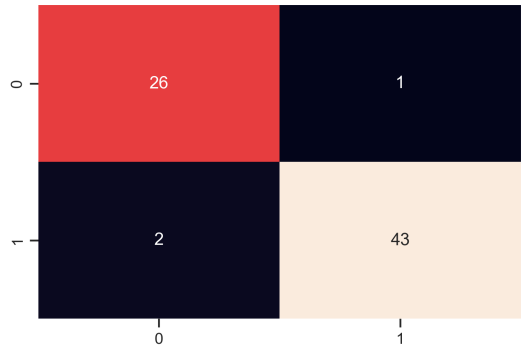
Korišćenjem **regularizacije** sprečava se da velike težine dođu do predela zasićenja, pa se i ovde dobija bolja tačnost od **94.44%**, ali problem je u tome što ima više propuštenih detekcija. Kriterijumska funkcija i konfuziona matrica:



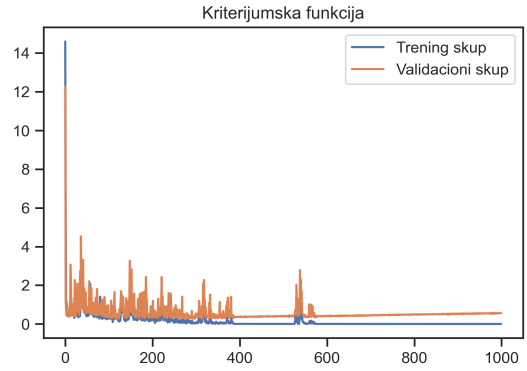


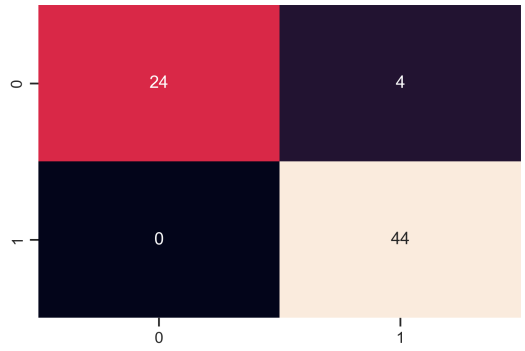
Sada će biti prikazane konfuzione matrice i kriterijumske funkcije različitih kombinacija skrivenih slojeva i broja neurona u njima:



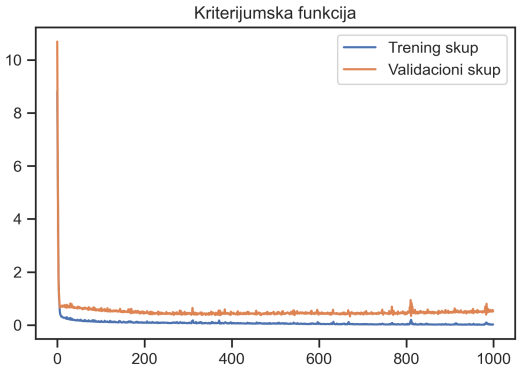


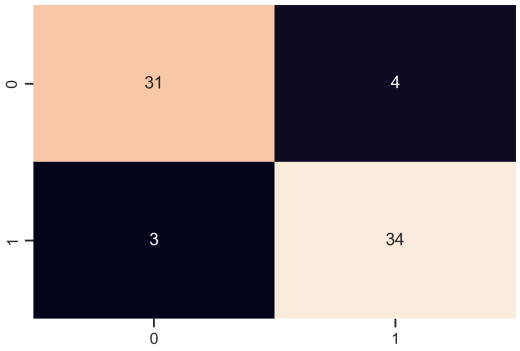
200x200 neurona (tačnost **95.83%**)



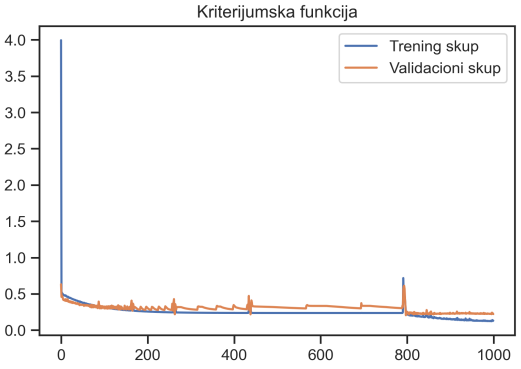


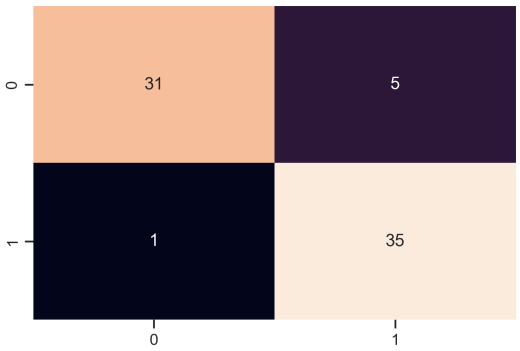
50x200 neurona (tačnost **94.44%**)



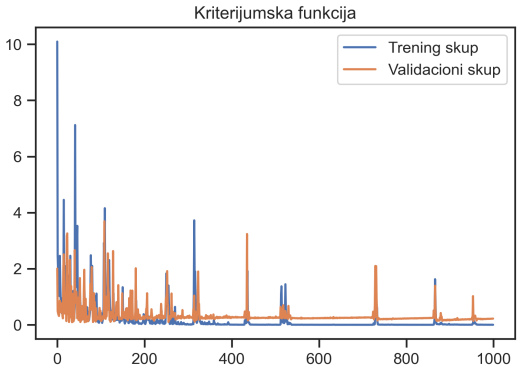


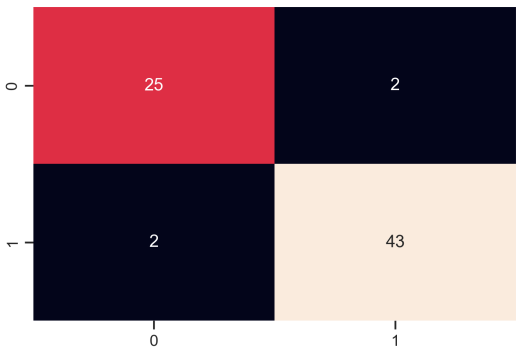
10 neurona 1 sloj (tačnost **90.28%**)



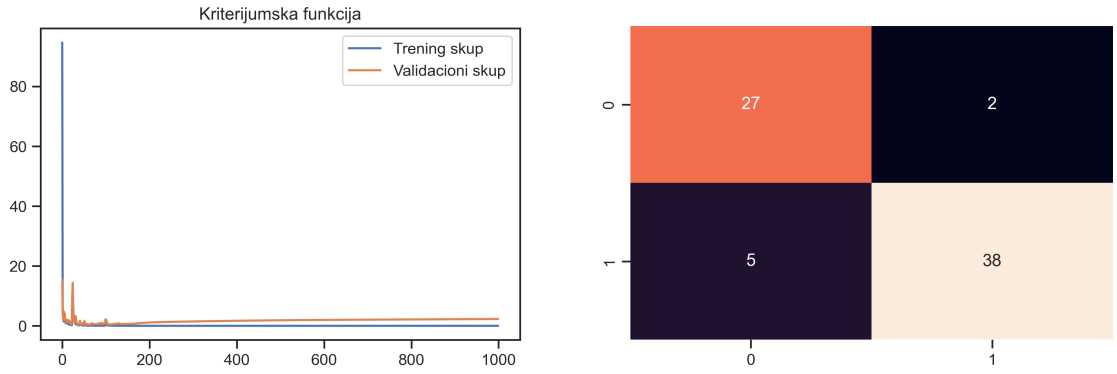


10x10x10x10 neurona (tačnost **91.67%**)





2000 neurona 1 sloj (tačnost **94.44%**)



2000x2000 neurona (tačnost **90.28%**)

Sve neuralne mreže su pokrenute u **1000** epoha. Kao što se može primetiti najveću tačnost (**95.83%**) postiže neuralna mreža sa 2 skrivena sloja sa 200 neurona u oba. Kombinacije 10 neurona u jednom sloju i 2 sloja od 2000 neurona postžu najgore rezultate, tačnost **90.28%**, prva mreža od dve zbog proste arhitekture, a druga od dve zbog mogućeg preobučavanja (a i mnogo vremena zahteva da se realizuje). Stvaranjem mreže sa 4 sloja i u svakom sloju 10 neurona, minimalno se poboljšava tačnost (**91.67%**) mreže sa jednim slojem od 10 neurona. Iz datih rezultata, dobija se da su optimalne mreže sa dva skrivena sloja, i srednjim brojem neurona (dobre rezultate daje i neuralna mreža sa 50 i 200 neurona u skrivenim slojevima redom - tačnost **94.44%**), kao i jednoslojna mreža sa mnogo neurona (**94.44%**).