Univerzitet u Beogradu - Elektrotehnički fakultet

Katedra za signale i sisteme

Projekat iz sistema odlučivanja u medicini

Vučenović Luka 2019/0631

Beograd, jul *mesec* 2022. godine

SADRŽAJ

1 UVOD

Prema CDC-u, srčane bolesti su jedan od vodećih uzroka smrti za ljude većine rasa u SAD (Afroamerikanci, Američki Indijanci i starosedeoci Aljaske, i belci). Oko polovine svih Amerikanaca (47 odsto) ima najmanje 1 od 3 ključna faktora rizika za srčane bolesti: visok krvni pritisak, visok holesterol i pušenje. Drugi ključni indikatori uključuju dijabetičarski status, gojaznost (visok BMI), nedovoljno fizičke aktivnosti ili ispijanje previše alkohola. Otkrivanje i sprečavanje faktora koji imaju najveći uticaj na srčane bolesti veoma je važno u zdravstvu. Kompjuterska tehnologija, zauzvrat, omogućava primenu metoda mašinskog učenja da detektuje "obrasce" iz podataka koji mogu da predvide stanje pacijenta.

2 ANALIZA SKUPA PODATAKA

Posmatrana baza podataka poseduje podatke o 319 400 pacijenata .Baza sadrži 18 varijabli(9 varijabli tipa „bool“,5 stringova i 4 decimalna tipa).Podaci o pacijentima su podeljeni u dve klase i 18 obelezja.

Pri analizi skupa podataka,prvenstveno su kategoričkim atributima dodeljene numeričke vrednosti ,gde smo obeležjima koji imaju samo 2 moguća ishoda dodeljivali vrednosti „1“ i „0“,dok smo obeležjima koja imaju više mogućih ishoda dodeljivali vrednosti od 0-n ,gde je n broj ishoda.

Zatim smo za svako obeležje za koje nije izvršena konverzija u numeričko ispitali da li je broj “0” vrednosti veći od jedne trećine prave veličine baze(broja pacijenata),pa ukoliko jeste takvo obeležje bi smo odbacili.

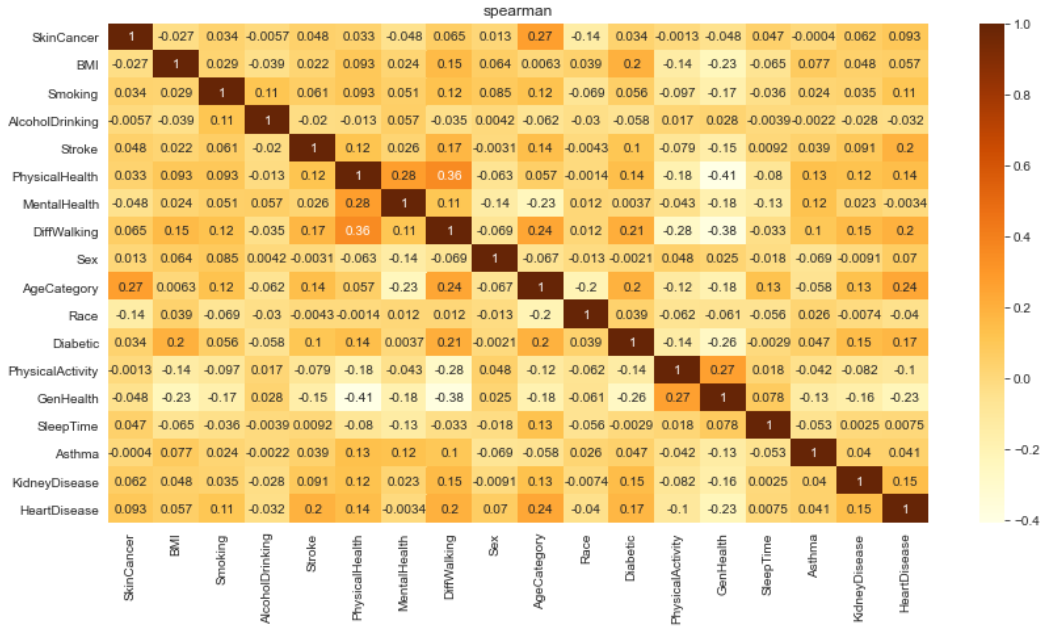
Klasu “ostalo” (iz smernica za izradu projekta) nismo oformili jer baza sadrži 2 klase ,pa je to ne moguće uraditi.

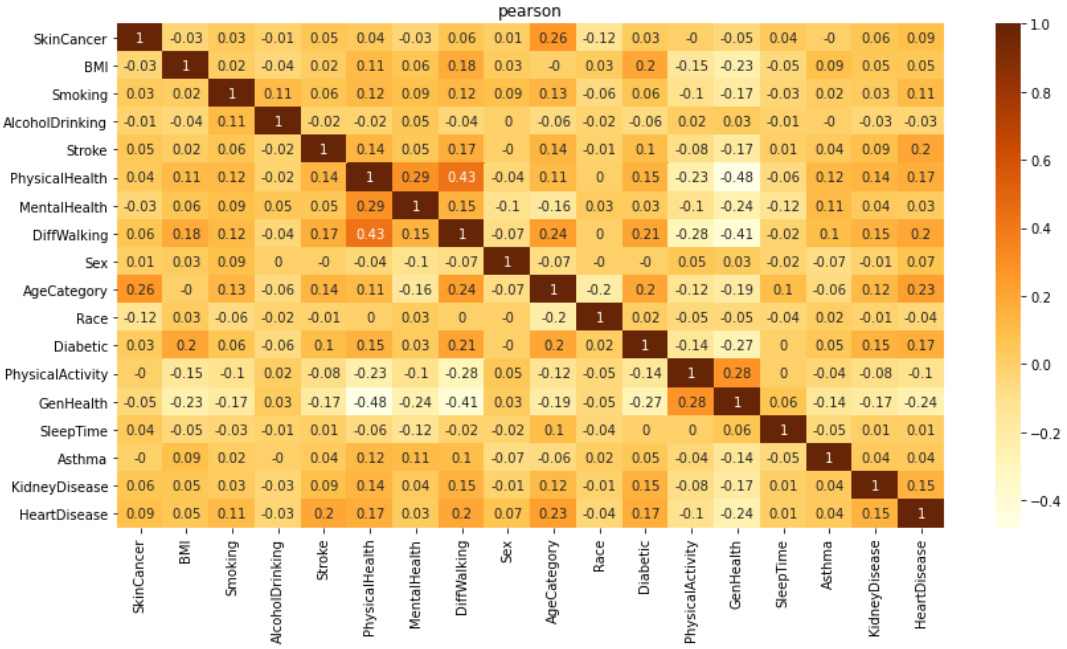
Nedostajuće attribute nismo menjali sa očekivanim vrednostima ,jer nedostajućih atributa nije bilo.

3 KORELACIJA IZMEĐU OBELEŽJA i INFORMATION GAIN

Nakon analize skupa podataka ,korelaciju između obeležja dobili smo korišćenjem dve metode .Prvu metodu koju smo koristili je pirsonovu,dok je druga korišćena metoda bila spirmanova,gde se spirmanova uzima za relevantnu ako pirsonova ne daje najbolji rezultat.

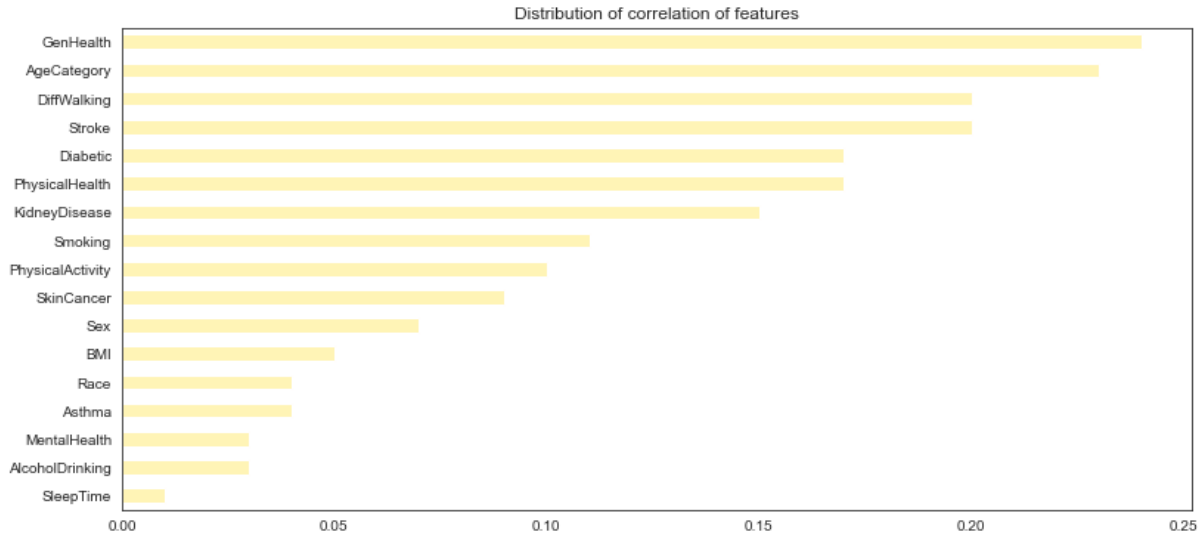
Na slikama ispod prikazane su korelacije za svaku metodu .





Pretpostavkom da je pirsonov metod dao očekijuće rezultate rangiramo

Koeficijente korelacije od najvećeg do najmanjeg (slika ispod).



Na osnovu ovog grafa iznad vidimo koji faktori najviše utiču na to da li osoba ima bolest srca ili ne.

Racunanje information gaina ,smo uradili tako što smo izračunali meru entropije klase pa zatim meru entropije klase u odnosu na atribut ,pa zatim oduzeli te dve i dobili podatak o informativnosti,tj. IG.

Zatim smo rangirali dobijene rezultate i dobili sledeće.

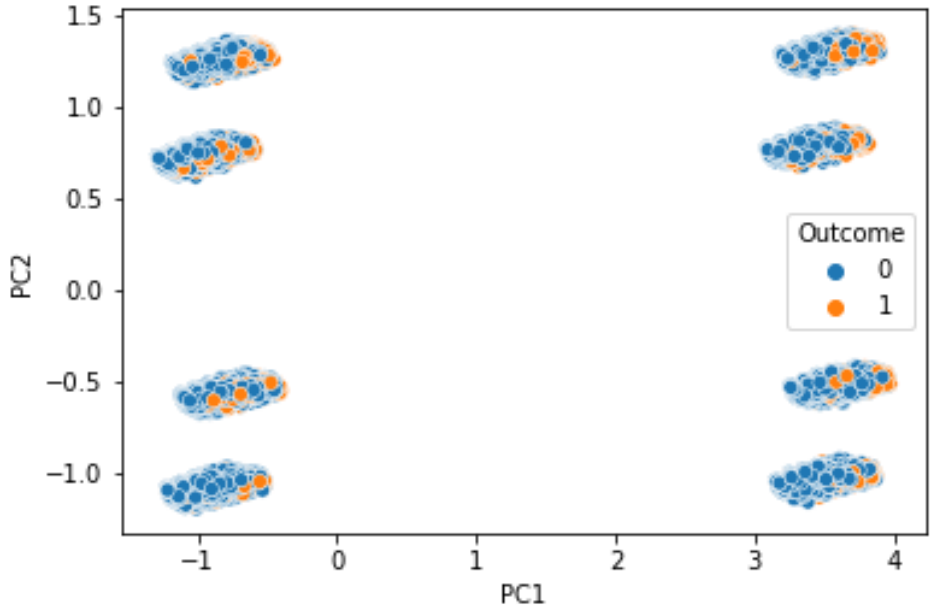
AlcoholDrinking 00 8.43886754e-04
Asthma 01 1.14628285e-03
Race 1.58938278e-03
MentalHealth 2.14337504e-03
BMI 2.48612342e-03
Sex 3.54284268e-03
SleepTime 01 4.59414726e-03
SkinCancer 5.17148279e-03
PhysicalActivity 6.51833911e-03
Smoking 8.22260752e-03
KidneyDisease 01 1.00965503e-02
Stroke 1.71799420e-02
PhysicalHealth 1.71879659e-02
Diabetic 1.72957661e-02
DiffWalking 2.24992992e-02
GenHealth 4.05516256e-02
AgeCategory 4.59991631e-02

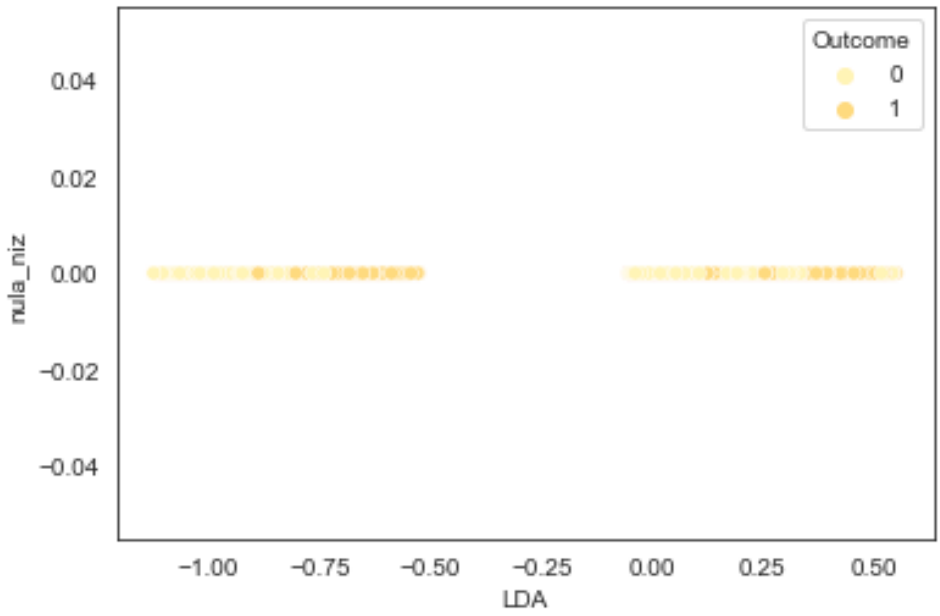
Odavde vidimo da se dobijeni rezultati u velikoj meri poklapaju sa rezultatima dobijenim pirsnovom metodom korelacije .

4 PCA I LDA METODA ZA REDUKCIJU DIMENZIJA

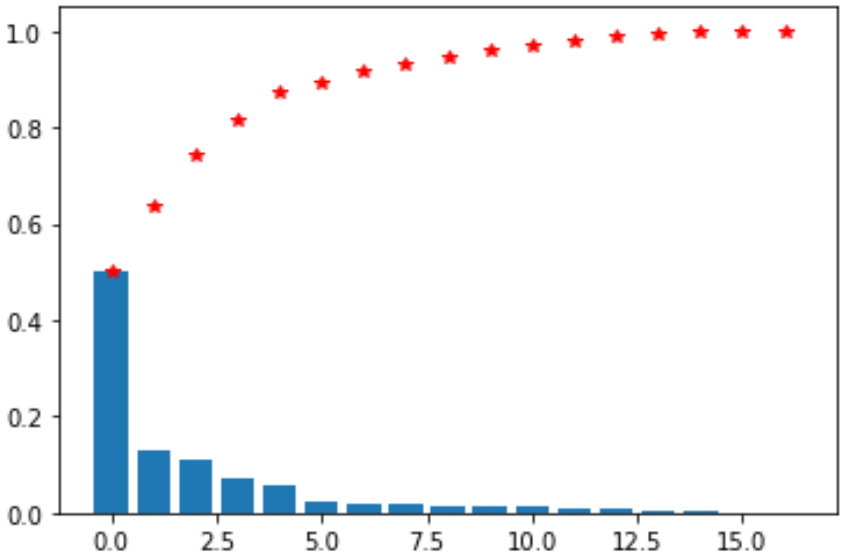
Pošto je skup podeljen na dve klase,koristili smo PCA metodu redukcijom na dve dimenzije i LDA metodu redukcijom na jednu dimenziju,zatim smo uporedili ove dve metode i uvideli koja daje bolji rezultat.

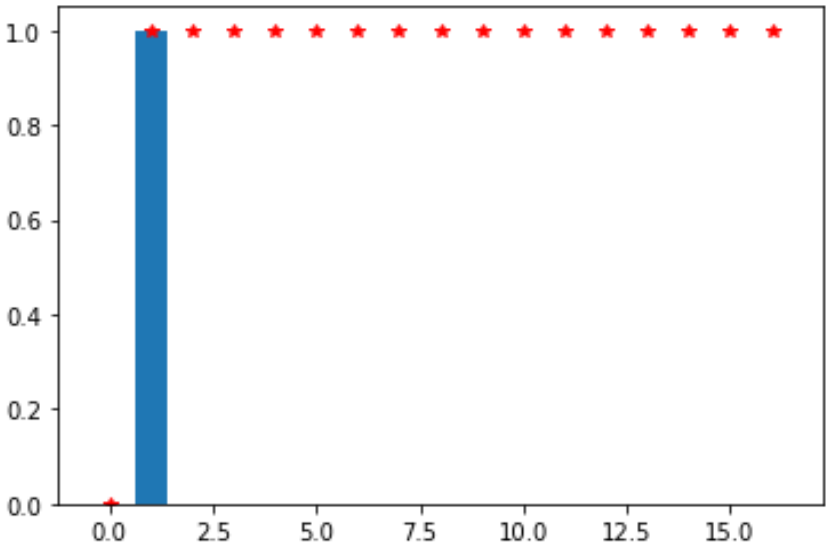
Dobijeni rezultati za obe metode su prikazani ispod





Linearnu separbilinost nije moguće uočiti ni kod PCA a ni kod LDA analize ,što se jasno može videti sa grafova,dok se ovakva raspodela prikazana na graficima može pripisati tome što je dosta obeležja tipa bool ,tj. ima samo dve moguće vrednosti.



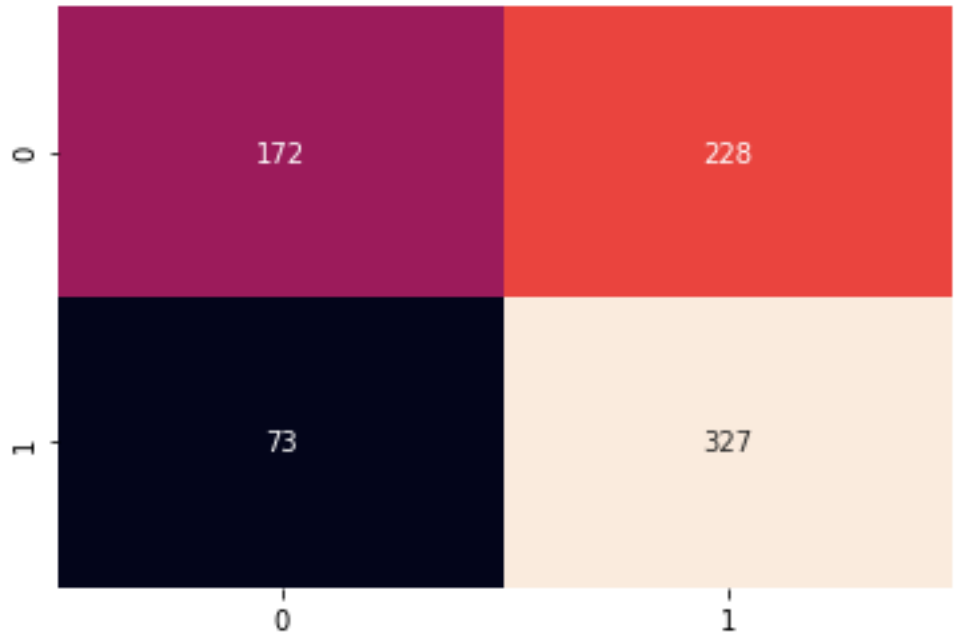


Sortirane sopstvene vrenosti za redom PCA I LDA analizu su prikazane iznad .

5 PARAMETARSKI KLASIFIKACIJA

Pošto smo na osnovu prošle tačke zaključili da ni LDA a ni PCA metoda ne daju odgovarajuću linearnu separabilnost,odlučili smo da uzmemo šest obeležja sa najvećim information gain-om,pa zatim prokektujemo klasifikator.

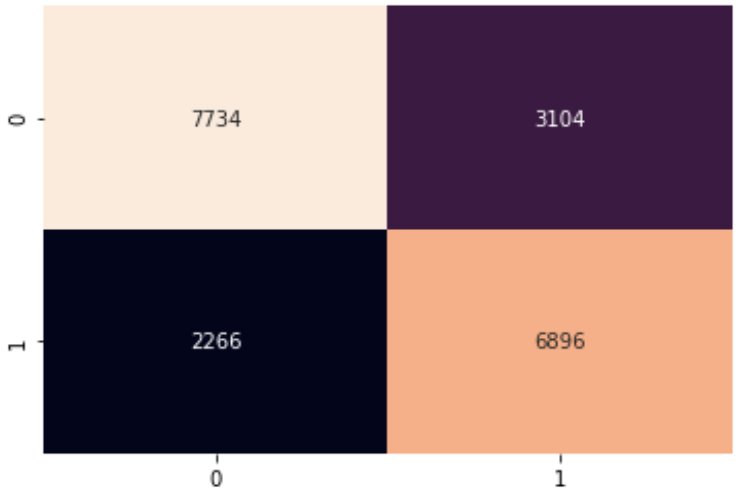
Kontuziona matrica dobijena korišćenjem Bajasovog klasifikatora prikazana je ispod.

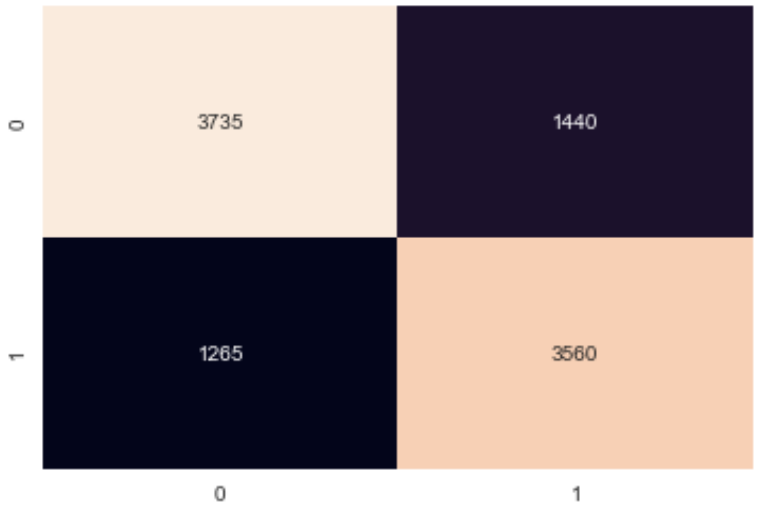


Dobijena kontuziona matrica bi trebala da ima većinu instanci raspoređenu na dijagonali što je donekle ispunjeno ,prostim računom dobili smo da je oko 63% instanci rasporedjeno na dijagonali matrice.

Pošto ova metoda ne daje najbolji rezultat izvršili smo parametarsku klasifikaciju,korišćenjem linearnog klasifikatora na bazi željenog izlaza,i dobili odgovarajuću separabilnost klasa.

Rezultate ove metode prikazali smo na slikama ispod.





Korišćenjem linearnog klasifikatora za N=20 000,uzimanjem četiri obeležja sa najvećim IG ,dobili smo sledeću konfuzionu matricu ,gde se očigledno vidi da je najveći broj instanci rasporedjen na dijagonali ,što ukazuje na uspešno isprojektovan klasifikator.

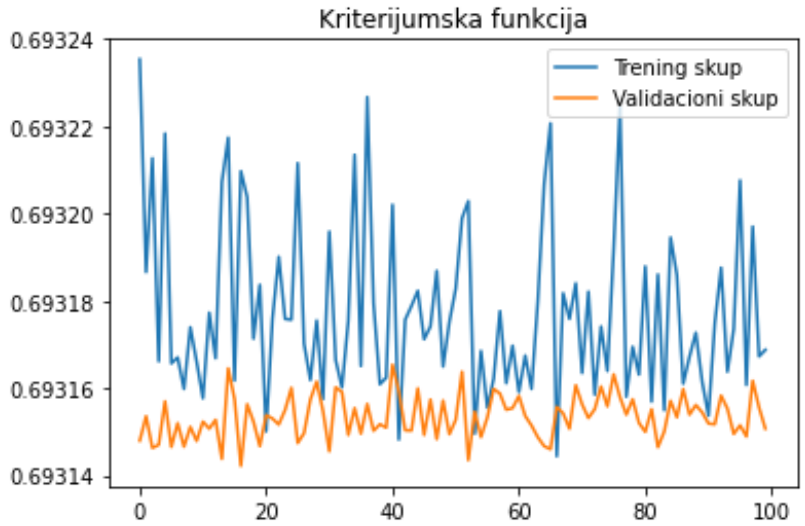
Prva slika je konfuziona matrica trening skupa dok je druga slika konfuziona matrica test skupa,odavde vidimo da vecu tacnost dobijamo na trening skupu sto je i ocekivano.

Takođe smo isprojektovali kvadratni klasifikator,ali nismo dobili odgovarajuću tačnost,pa uzimamo linearni klasifikator za relevantni.

6.NEURALNE MREŽE

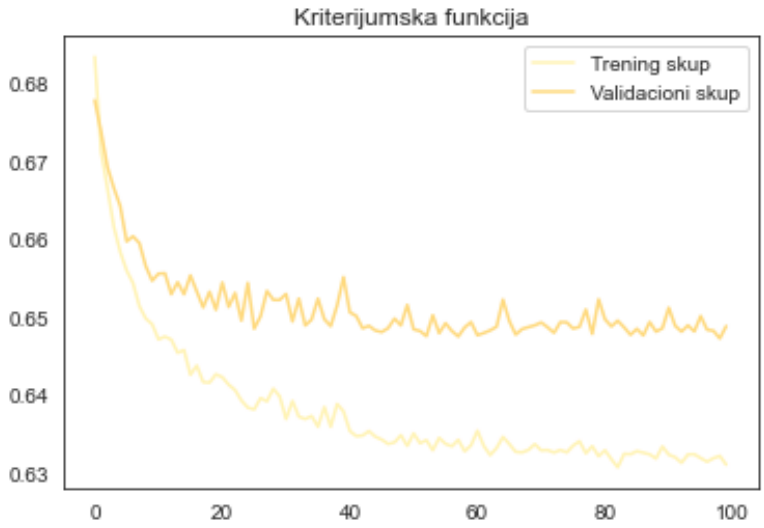
Prvo smo koristili neuralnu mrežu sa jednim skrivenim slojem sa jednim neuronom,i ulaznim slojem sa 10 neurona,bez zaštite od preobučavanja.Dobijeni rezultati su prikazani ispod.Izlazni sloj u svakom modelu imati jedan neuron.

Tačnost na trening skupu iznosi: 50.062501430511475%.
Tačnost na test skupu iznosi: 49.75000023841858%.

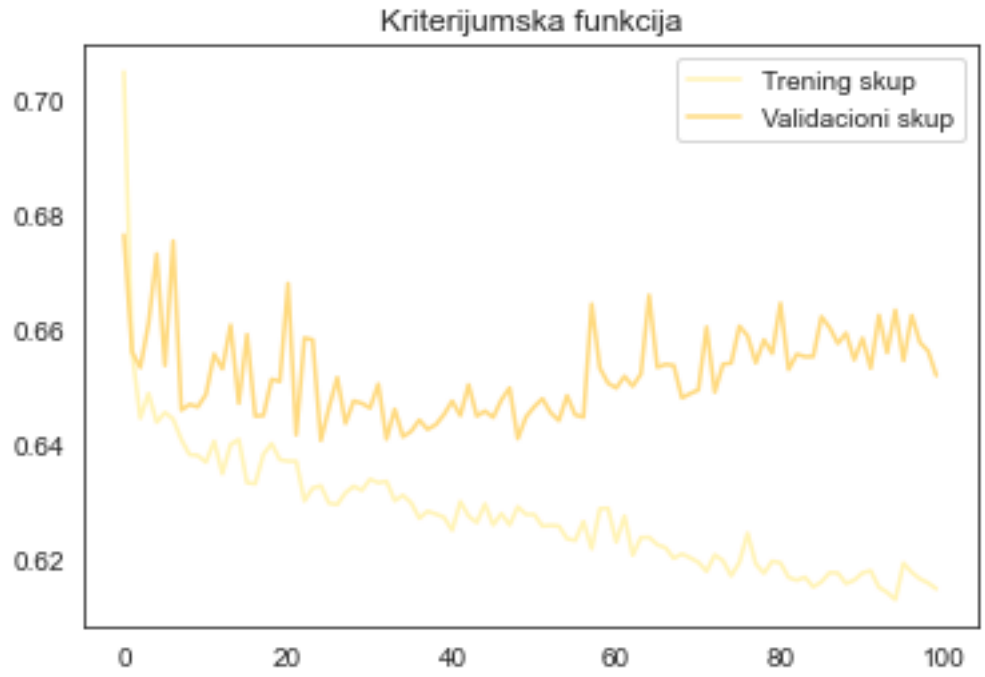


Odavde vidimo da korišćenjem ovakvog modela dobijamo nedovoljnu tačnost,što se pripisuje malom broju neurona u ulaznom i skrivenom sloju.

Zatim je korišćena ista struktura stim što što smo povećali broj neurona u ulaznom sloju na 100.Dobili smo da je tačnost na trening skupu 64.43750262260437%. dok je tačnost na test skupu 61.000001430511475%.Odavde vidimo da povećanjem broja neurona u ulaznom sloju dobijamo veliko povećanje tačnosti.



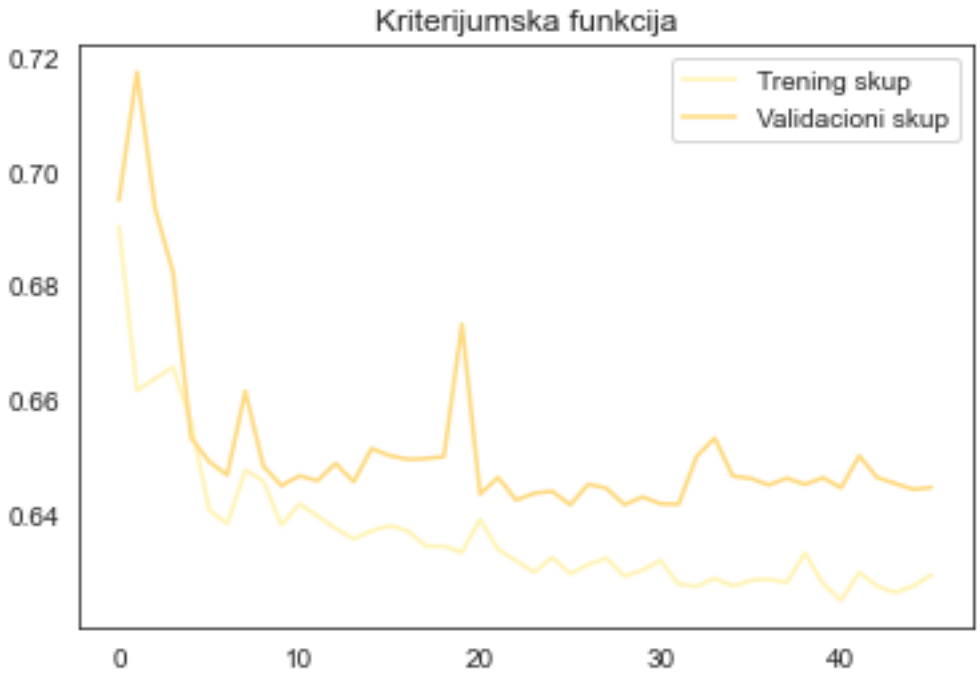
Zatim smo koristili 3 skrivena sloja za strukuru sa 100 neurona u ulaznom sloju i po 1000 neurona u skrivenim slojevima. Dobili smo da je tačnost na trening skupu 65.56249856948853%,dok je tačnost na test skupu 62.00000047683716%.%..



Korišćenjem strukture sa jednim skrivenim slojem sa 1000 neurona,i ulaznim slojem od 100 neurona ali uz korišćenje metode ranog zaustavljanja dobijamo veću tačnost .

Tačnost na trening skupu iznosi: 64.49999809265137%.

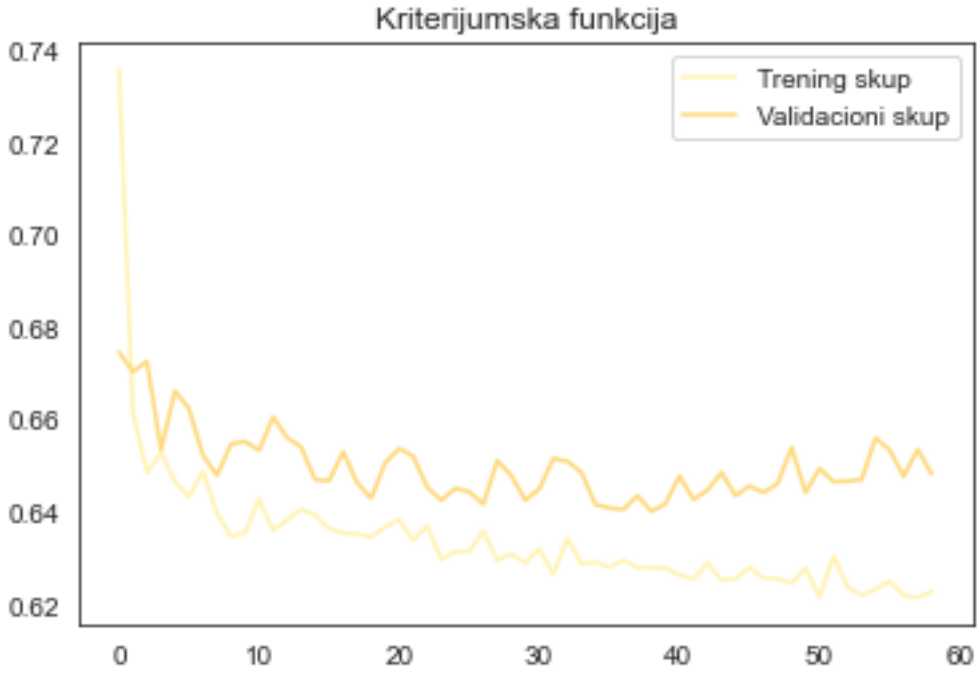
Tačnost na test skupu iznosi: 62.25000023841858%.



Korišćenjem iste strukture samo sa 10 000 neurona u skrivenom sloju dobijamo sličnu tačnost ali je vreme izvršavanja programa dosta veće pa ovakav model nije optimalan.

Tačnost na trening skupu iznosi: 64.81249928474426%.

Tačnost na test skupu iznosi: 61.000001430511475%.



U slučaju korišćenja regularizacije za model sa 100 neurona u ulaznom,1000 neurona u skrivenom i jednim neuronom u izlaznom sloju dobijamo da je tačnost na trening skupu 71.18750214576721%.% dok je tačnost na test skupu 59.75000262260437% što je i očekivan rezultat da tačnost na trening skupu bude veća nego na test skupu.