Izveštaj projektnog zadatka iz Sistema odlučivanja u medicini (13053SOM)

Božidar Obradović, 2017/0113
Septembar 2020.

Sadržaj

1) Osnovne informacije o obrađenoj bazi – Thoracic Surgery (20.)

Podaci iz ove baze su posvećeni klasifikaciji obolelih od raka pluća godinu dana nakon operacije. Baza je retroaktivno sastavljena u Centru za torakalnu hirurgiju u Vroclavu u periodu od 2007. do 2011. godine. Podaci su prikupljeni od pacijenata koji su bili podvrgnuti resekciji pluća usled pojave karcinoma. Centar je povezan sa odeljenjem za torakalnu hirurgiju Medicinskog fakulteta u Vroclavu i Centra za pulmonalne bolesti Donjeg Šleskog, dok je baza deo Nacionalnog registra za rak pluća kojim upravlja Institut za tuberkulozu i plućne bolesti u Varšavi.

Atributi

Originalna baza je imala 139 obeležja od kojih su 36 bili pred-operativna, 37 peri-operativniaa 46 (17 patološkog tipa) post-operativna [1]. Redukcija pred-operativnih obeležja je izvršena koristeći IG vrednosti kao kriterijum [1]. Odabrana obeležja (njih 16 + pripadnost klasi) se nalaze u bazi na kojoj je rađeno (atributi kod kojih ne stoje vrednosti u zagradi su tipa true/false):

1. Dijagnoza lekara – specifična kombinacija ICD-10 kodova za primarne, sekundarne i po potrebi višestruke tumore (***DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1***)
2. Forsirani vitalni kapacitet – FVC (numerički)
3. Zapremina izdahnuta nakon prve sekunde prinudnog izdaha – FEV1 (numerički)
4. Procena opšteg telesnog stanja pacijenta na Zubrodovoj skali (***PRZ2, PRZ1, PRZ0***)
5. Bol pre operacije
6. Hemoptizija pre operacije (iskašljavanje krvi / krvnih ispljuvaka)
7. Dispneja pre operacije (nedostatak vazduha praćen subjektivnim osećajem otežanog disanja)
8. Kašalj pre operacije
9. Slabost pre operacije
10. T u TNM oznaci – veličina originalnog tumora (od najmanjeg ka najvećem: ***OC11, OC12, OC13, OC14***
11. Dijabetes melitus tj. tip 2
12. Srčani udar u poslednjih 6 meseci
13. Periferne arterijske bolesti (PAD)
14. Pušenje cigareta
15. Astma
16. Broj godina (numerički)
17. Stanje pacijenta nakon godinu dana (F – živ, T – mrtav)

2) Analiza baze

Baza sadrži 3 nominalna atributa (dijagnoza, stanje na Zubrodovoj skali i veličina tumora), 3 numerička atributa (FVC, FEV1 i broj godina) dok su ostali atributi binarnog tipa (true/false). Atributi binarnog tipa su prevedeni u skup {0,1} po pravilu $T \rightarrow 1, F \rightarrow 0$.

Postoje dva pristupa kodiranju nominalnih atributa:

1. ***Prevođenje uzimajući u obzir ugrađeni poredak atributa gde je to moguće***

   Na Zubrodovoj skali 0 predstavlja smrt, 1 da ima simptome a 2 da je sposoban, samim tim može da se uspostavi poredak. Shodno tome, taj atribut je preveden u numerički po sledećem pravilu:

   $$PRZi \rightarrow i, i \in \{0, 1, 2\}$$

   Atribut koji govori o veličini tumora je definisan tako da OC11 predstavlja slučaj sa najmanjom veličinom tumora i da sa porastom tumora poslednji broj oznake raste, tj. OC14 predstavlja slučaj sa najvećim tumorom. S obzirom na poredak definisan atributom, on se prevodi u numerički na sledeći način:

   $$OC1i \rightarrow i, i \in \{1, 2, 3, 4\}$$

   Kod atributa dijagnoze lekara, za razliku od prethodna dva, ne može da se uspostavi poredak, tj. da se na osnovu poslednjeg slova (i u DGNi) vrednosti poređaju od „najbolje“ do najgore. U ovom slučaju prevođenje u numerički domen je urađeno na sledeći način:

   - Vrednosti atributa su podeljene u klase ekvivalencije (u svakoj klasi se nalaze atributi iste vrednosti. broj klasa je jednak broju različitih vrednosti koje atribut može da ima, klase su međusobno disjunktne) [2]
   - Za svaku klasu ekvivalencije, njenim članovima je dodeljena numerička vrednost:

     $$R = \frac{n+1}{2}$$

     gde n predstavlja broj elemenata klase ekvivalencije u kojoj se vrednost atributa nalazi (tj, broj ponavljanja te vrednosti atributa u celom skupu)

   Time se skup vrednosti ovog atributa {DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1} preslikava u skup brojeva {175, 26.5, 24, 2.5, 8, 1.5, 1} na kom može da se vrši dalja obrada.

2. ***Konverzija svih atributa zanemarujući interni poredak***

   U ovom slučaju svi atributi se kodiraju kao atribut dijagnoze, na osnovu klasa ekvivalencije. Identičnim postupkom skup vrednosti atributa Zubrodove skale {PRZ0, PRZ1, PRZ2} se preslikava u {14, 157. 65.5}, dok se vrednosti atributa veličine tumora {OC11, OC14, OC12, OC13} preslikavaju u skup {89. 9. 129. 10}. Skup vrednosti atributa dijagnoze lekara je isti kao i u slučaju pod 1.

3) IG i korelisanost obeležja

Za računanje IG odabrana su sledeća obeležja: dijagnoza lekara, stanje na Zubrodovoj skali, bol, hemoptizija, dispneja, kašalj, slabost, veličina tumora, dijabetes, srčani udar. Dobijene IG vrednosti su identične za oba slučaja kodovanja. One su prikazane u Tabeli 1:

Atribut	***IG vrednost***
Dijagnoza lekara	0.0243
Stanje na Zubrodovoj skali	0.0065
Bol	0.0021
Hemoptizija	0.0029
Dispneja	0.0067
Kašalj	0.0060
Slabost	0.0050
Veličina tumora	0.0210
Dijabetes	0.0072
Srčani udar	0.0009

Tabela 1: IG vrednosti odabranih atributa

Prilikom računanja korelisanosti između atributa dolazi do razilaženja predloženih modela kodovanje, što je prikazano u Tabeli 2:

$Obeležja\ na\ kojima\ je\ računat\ CFS \,/\, Model\ kodovanja$	2	1
Pun skup obeležja	0.0702	0.1448
Dijagnoza i veličina tumora (kao najinformativnija)	-0.1602	0.0356

Tabela 2: Prikaz korelisanosti obeležja skupa

Na osnovu ove dve tabele mogu se doneti sledeći zaključci:

1. Po modelu 1, najinformativnija obeležja su manje korelisana nego pun skup obeležja. što po definiciji CFS metode nije moguće. **Zbog toga je u daljem računanju korišćen model 2, tj. kodiranje obeležja na osnovu klase ekvivalencije**.
2. Od odabranih obeležja najviše informacije pružaju dijagnoza lekara i veličina tumora, što je za očekivati, međutim, korelisanost ovih obeležja nije velika.

4) Smanjenje dimenzija i parametarska klasifikacija

Za redukciju dimenzije podataka korišćena je metoda na bazi mere rasipanja. Ona se zasniva na ekstremizaciji pogodnih kriterijuma koji imaju cilj da održe razdvojivost klasa. U ovom slučaju minimiziran je kriterijum:

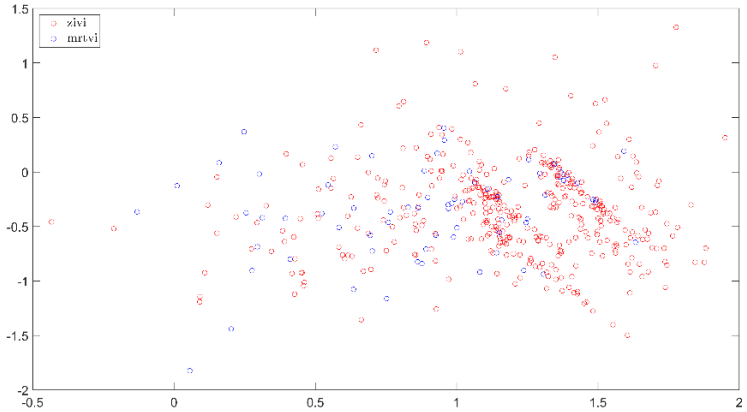
$$J = tr(S_w^{-1} S_b)$$

gde su S_w i S_b matrice unutarklasnog i međuklasnog rasejanja. Za ovaj kriterijum ekstremizacija se svodi na traženje m sopstvenih vektora matrice $S_w^{-1} S_b$ kojima odgovara m najvećih sopstvenih vrednosti; m predstavlja broj dimenzija novonastalog (smanjenog) vektora. Dobijeni vektori $\Psi_1, \Psi_2, \ldots, \Psi_m$ formiraju transformacionu matricu kojom se množi početni vektor podataka.

Primenom navedenog metoda na date podatke dobijene su sledeće sopstvene vrednosti:

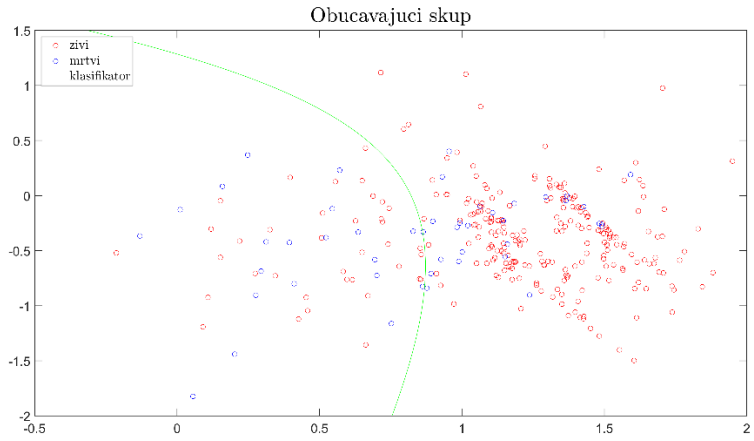
$$\lambda_1 = 0.1181$$
$$\lambda_2 = 1.1342 * 10^{-17}$$
$$\lambda_3 = \lambda_4 = 4.8457 * 10^{-18}$$
$$\lambda_5 = \lambda_6 = 4.5651 * 10^{-17}$$

Shodno rezultatima, transformacionu matricu A čine vektori koji odgovaraju sopstvenim vektorima λ_1 i λ_2. Rezultati dobijeni ovom metodom su prikazani na Slici 1:

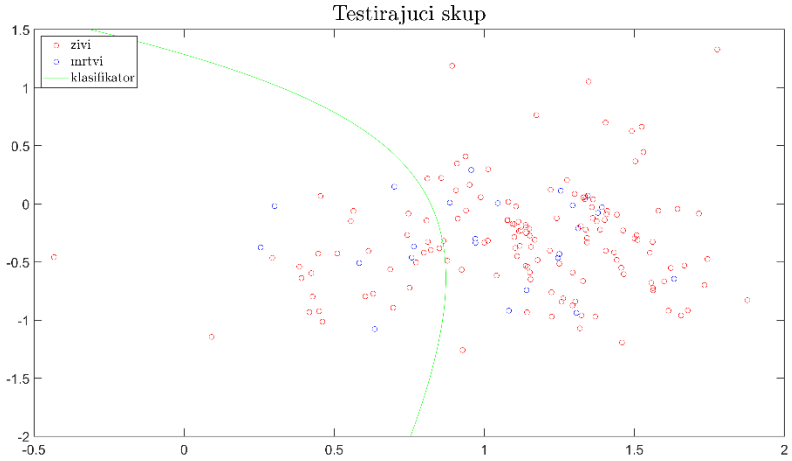


Slika 1: Podaci nakon redukcije dimenzija

Dodavanje treće dimenzije nije poboljšalo razdvojivost klasa, pa je zadržano smanjenje na dve dimenzije. Dobijeni podaci su klasifikovani kvadratnim klasifikatorom sa željenim izlazom. Podaci su podeljeni na obučavajući i testirajući skup u odnosu 65:35. Rezultati klasifikacije na obučavajućem i testirajućem skupu su prikazani na slikama 2 i 3:

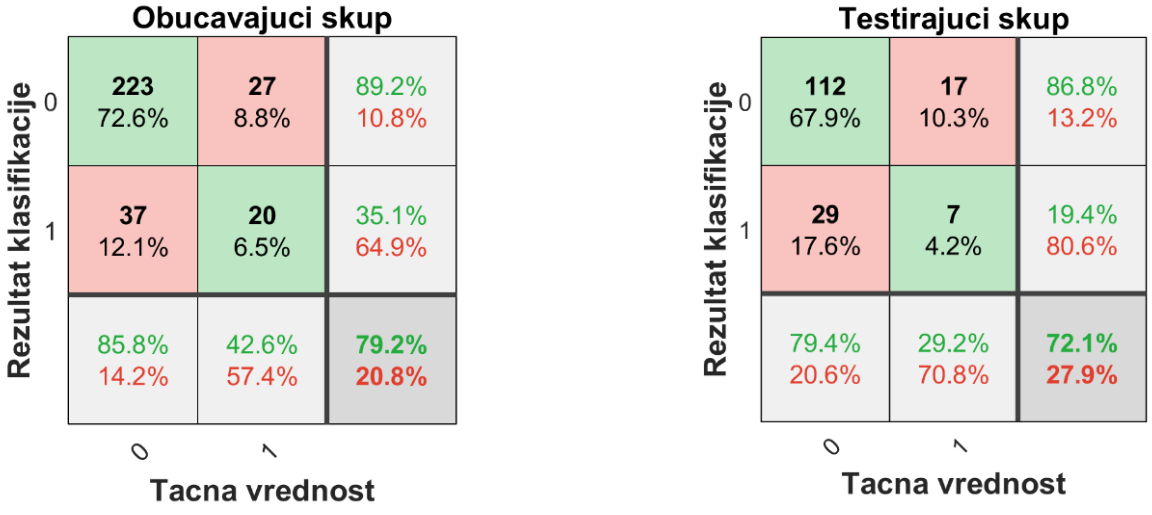


Slika 2: Rezultati klasifikacije na obučavajućem skupu



Slika 3: Rezultati klasifikacije na testirajućem skupu

Konfuzione matrice za oba slučaja su prikazane na Slici 4:



Slika 4: Konfuzione matrice za obučavajući i testirajući skup

Željeni izlaz je biran tako da se da prioritet smanjenju verovatnoće propuštene detekcije (misdetection - procenat bolesnih koji su klasifikovani kao zdravi). Za obučavajući skup ova verovatnoća iznosi 0.892, a za testirajući skup 0.868. Kao posledica ovoga povećana je verovatnoća lažnih alarma (false alarms – procenat zdravih koji su klasifikovani kao bolesni) - ona iznosi 0.351 i 0.194 za obučavajući i testirani skup, respektivno.

5) Neuralne mreže

Izvršeno je obučavanje i testiranje neuralnih mreža nad datim podacima različitih struktura sa jednim ili više skrivenih slojeva. Zajedničko za sve neuralne mreže jeste broj ulaznih neurona - 16 i broj izlaznih neurona – jedan.

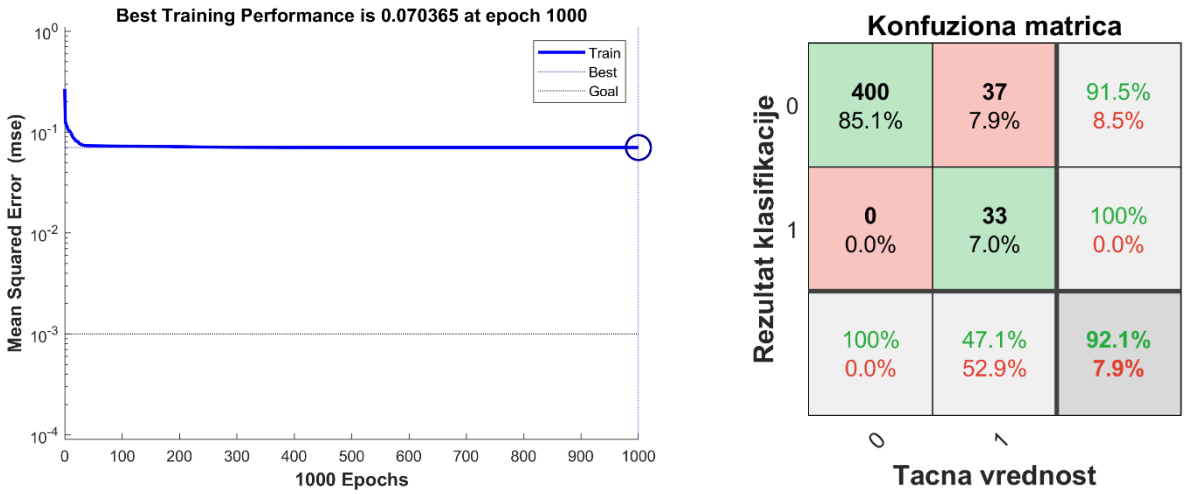
Jedan skriveni sloj

Optimalnan broj neurona u sloju je dobijen tako što je urađeno obučavanje i testiranje mreže za više vrednosti broja neurona u sloju. Svaka mreža je obučavana na isti način, sa istim vrednostima parametara (maksimalan broj epoha 1000, ciljna greška 0.001). Vrednosti koji su ključne za ovo razmatranje su tačnost mreže (suma elemenata na glavnoj dijagonali konfuzione matrice / suma svih elemenata konfuzione matrice) i vreme potrebno za obučavanje i simulaciju (tj. vreme izvršavanja tih linija koda). Dobijeni rezultati su prikazani u Tabeli 2:

Broj neurona u sloju	Tačnost	Vreme izvršavanja (s)
5	0.9404	4.4923
10	0.9468	8.0944
15	0.9723	19.2121
20	0.9851	25.1009
25	0.9915	38.3708
30	0.9979	51.9736
35	1	50.7869
40	1	36.8373
45	1	28.0391
50	1	37.0243
55	1	28.9177
60	1	32.5572
65	1	39.1114
70	1	40.4035
75	1	38.9050
80	1	64.8872
85	1	69.1044
90	1	62.6299
95	1	73.2746
100	1	104.7149

Tabela 2: Performanse neuralne mreže za različit broj neurona u sloju

Na osnovu dobijenih podataka optimalan broj neurona je postavljen na 5. Grafik performanse i konfuziona matrica su prikazani na Slici 5:



Slika 5: Performansa i konfuziona matrica mreže sa optimalnim brojem neurona u skrivenom sloju

Za broj neurona manji od optimalnog postiže se neznatno povećanje brzine na uštrb pada performanse; za 2 sloja vreme izvršavanja padne na 2 sekunde, ali tačnost pada na 85%. Preveliki broj neurona neznatno povećava tačnost na uštrb pada brzine, dobar primer ovoga predstavlja performansa mreže za 25 i 30 slojeva.

Više skrivenih slojeva

Uvođenjem dodatnih slojeva povećava se tačnost mreže uz neznatno povećanje (ili čak smanjenje) vremena izvršavanja. Za računanje optimalnog broja u drugom i trećem sloju korišćena je ista metoda kao i za prvi sloj. U prvom (za drugi i treći) i drugom (za treći) sloju je postavljen prethodno određen optimalan broj neurona. Rezultati ovih analiza se nalaze u tabelama 3 i 4:

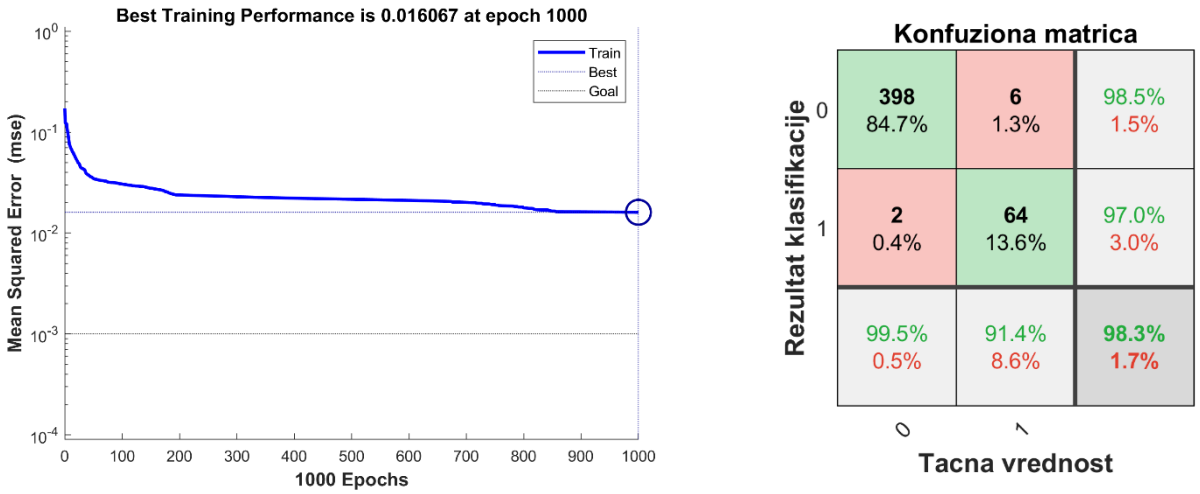
Broj neurona u sloju	Tačnost	Vreme izvršavanja (s)
2	0.9255	44.0426
4	0.9511	40.9217
6	0.9426	42.8919
8	0.9468	48.3210
10	0.9596	45.6747
12	0.9574	53.6237
14	0.9702	75.6885
16	0.9638	94.1756
18	0.9617	20.2383
20	0.9574	11.1229

Tabela 3: Performanse neuralne mreže za različit broj neurona u drugom sloju

Broj neurona u sloju	Tačnost	Vreme izvršavanja (s)
2	0.9766	9.1489
4	0.9702	12.7227
6	0.9702	13.8685
8	0.9617	21.1678
10	0.9702	25.0092

Tabela 4: Performanse neuralne mreže za različit broj neurona u trećem sloju

Na osnovu izvršene analize optimalna struktura mreže sa tri sloja sadrži 5, 20 i 2 neurona. redom. u slojevima. Grafik performanse i konfuziona matrica za ovu mrežu su prikazani na Slici 6:



Slika 5: Performansa i konfuziona matrica mreže sa optimalnim brojem neurona u skrivenim slojevima

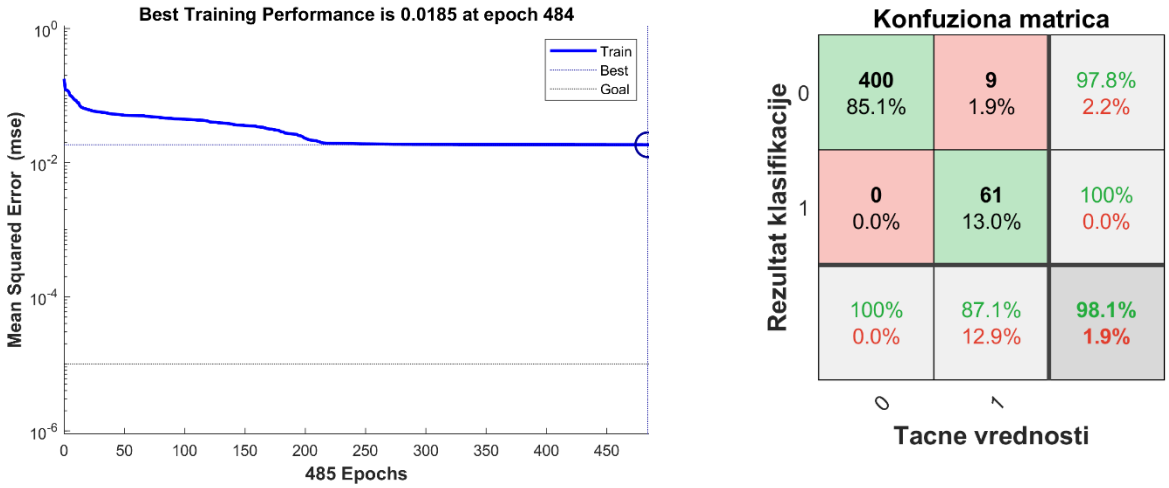
Na osnovu priloženih podataka za drugi i treći sloj dolazi se do sledećih zaključaka:

- Premali broj neurona u drugom sloju je dva, u tom slučaju i tačnost i brzina su znatno ispod optimalnih vrednosti
- U drugom sloju dolazilo je do neznatnog rasta precizosti na uštrb vremena računanja, što je kulminiralo za slučaj 14 neurona u skrivenom sloju. Nakon toga, sa porastom broja neurona u sloju vreme izvršavanja znatno opada uz mali gubitak tačnosti. Ovaj trend se nastavlja dok sloj ne dostigne oko 70 slojeva, kada tačnost opada na oko 85%
- Treći sloj ne unosi veće promene u tačnosti, sa porastom broja neurona ona se ne menja značajno a vreme izvršavanja raste (npr. za 20 neurona u sloju). Optimalnije konfiguracije se mogu dobiti bez trećeg sloja, sa povećanim brojem neurona u drugom sloju

Zaštita od preobučavanja neuralnih mreža sa više slojeva

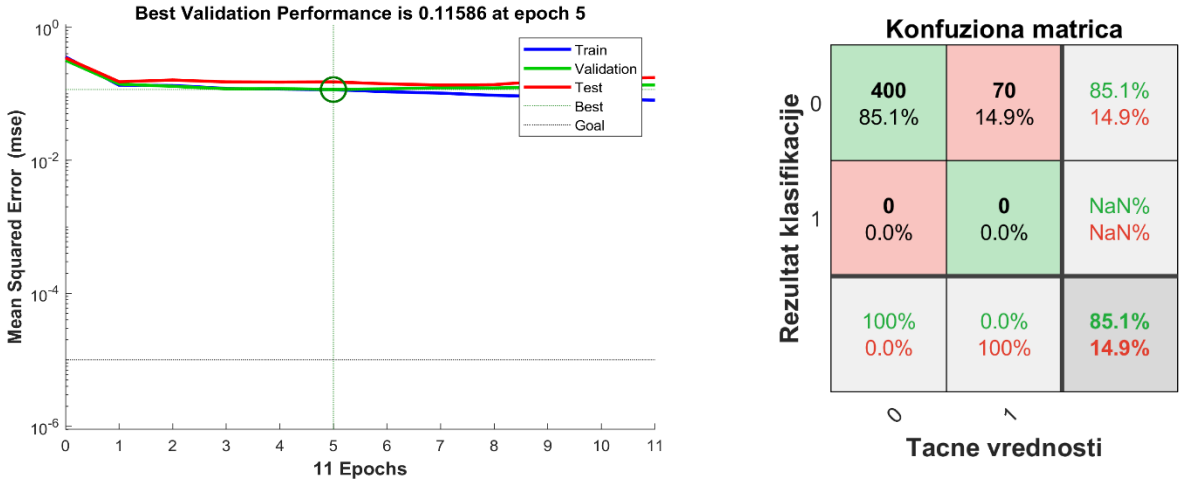
Preobučavanje kao posledicu ostavlja mrežu sa manjkom fleksibilnosti na nove podatke.. U ovom projektu su izložena dve metode zaštite od preobučavanja: regularizacija i rano zaustavljanje. Za obe metode korišćena je neuralna mreža sa strukturom iz prethodne analize (tri skrivena sloja sa 5, 20 i 2 neurona u njima, respektivno):

1. ***Regularizacija*** – uvođenje regularizacionog člana u kriterijumsku funkciju radi sprečavanja zasićenja (posledica prevelikih težinskih koeficijenata), Slika 7:



Slika 7: Performansa i konfuziona matrica mreže sa optimalnim brojem neurona u slučaju zaštite regularizacijom

2. ***Rano zaustavljanje*** – provera vrednosti kriterijumske funkcije (tj. greške) na testirajućem i validacionom skupu; u nekom trenutku na tim skupovima ona počinje da raste iako na obučavajućem skupu ona i dalje opada, Slika 8:



Slika 7: Performansa i konfuziona matrica mreže sa optimalnim brojem neurona u slučaju zaštite regularizacijom

Ovakvi rezultati dobijeni metodama za zaštitu od preobučavanja mogu biti objašnjeni samom bazom podataka. Naime, prethodni rezultati svedoče o činjenici da sa ovim podacima, mreža uopšte nije u stanju da dođe do stadijuma preobučavanja, pa će ranije zaustavljanje mreže samo dovesti do povećanja greške.