Primena neuronskih mreža na Network Intrusion Data

Projekat u okviru kursa Mašinsko učenje

Vladana Đorđević 1092/2019
Aleksandra Jovičić 1088/2019

July 7, 2020

Uvod
Opis baza podataka

- Primena veštačkih neuronskih mreža na skup podataka koji se odnosi na napadanje računarskih mreža u cilju utvrđivanja da li je data konekcija normalan saobraćaj ili napad.
- Ovaj skup je korišćen na KDD (Knowledge Discovery and Data Mining) kupu održanom 1999. godine.
- Celokupan skup sadrži oko 5 000 000 instanci, a u ovom projektu je korišćeno 10% skupa što iznosi 494020 instanci.
- Svaka instanca predstavlja jedan zapis o konekciji i opisana je pomoću 41 atributa i označena kao napad na mrežu određenog tipa ili kao normalan sadržaj.

Uvod
Opis baza podataka

- Svi atributi se mogu podeliti u tri grupe, za svaku grupu smo izdvojile najvažnije atribute i prikazale ih tabelarno:

1. osnovni atributi: u ovoj kategoriji se nalaze svi atributi koji mogu biti izdvojeni iz TCP/IP konekcije.

ime atributa	**opis**	**tip**
protocol_type	tip protokola	diskretan
service	internet servis na destinaciji	diskretan
src_bytes	broj bajtova podataka od izvora do destinacije	neprekidan
flag	status konekcije (normalan ili greška)	diskretan

1. atributi saobraćaja: ova kategorija uključuje atribute koji su izračunati uzimajući u obzir vremenski interval i podeljena je u dve grupe ("same host" i "same service").

ime atributa	**opis**	**tip**
count	br. kon. ka istom hostu kao i trenutna kon. u poslednje 2 sekunde	neprekidan
serror_rate	procenat konekcija koje imaju SYN greške	neprekidan
rerror_rate	procenat konekcija koje imaju REJ greške	neprekidan
same_srv_rate	procenat konekcije ka istom servisu	neprekidan
diff_srv_rate	procenat konekcije ka različitim servisima	neprekidan

1. atributi sadržaja: uključuje atribute koji pretražuju sumnjiva ponašanja u podacima.

ime atributa	**opis**	**tip**
num_failed_logins	broj neuspelih pokušaja prijavljivanja	neprekidan
logged_in	1 - za uspešno prijavljivanje, 0 - inače	diskretan
su_attempted	1 - ako je pokušana komanda "su root", 0 - inače	diskretan
num_file_creations	broj operacija kreiranja fajlova	neprekidan

Uvod
Problem

- Dinamički mehanizmi zaštite, sistemi za otkrivanje upada (SZOU), su obično ili zasnovani na hostu ili zasnovani na mreži. SZOU zasnovani na hostu nadgledaju resurse kao što su sistemski logovi, fajl sistemi i resursi sa diska, dok SZOU zasnovani na mreži nadgledaju podatke koji prolaze kroz mrežu.
- Zadatak je otkriti direktno iz (trening) podataka odgovarajuće modele koji su u stanju da razlikuju normalno ponašanje od napada. Dobijeni model se potom koristi za klasifikaciju nad nepoznatim podacima.

Uvod
Problem

- KDD99 baza podataka je zasnovana na inicijativi 1998 DARPA da obezbedi dizajnerima sistema za otkrivanje upada benchmark pomoću koga će oceniti različite metodologije. Da bi se to postiglo napravljena je simulacija veštačke vojne mreže koja se sastoji od tri "ciljane" mašine na kojima se izvršavaju različiti operativni sistemi i servisi. Dodatne tri mašine su korišćene da "zamaskiraju" različite IP adrese i na taj način generišući saobraćaj između različitih IP adresa. Na kraju, korišćeno je "njuškalo" koje beleži sav mrežni saobraćaj. Ukupan simulirani period je sedam nedelja.

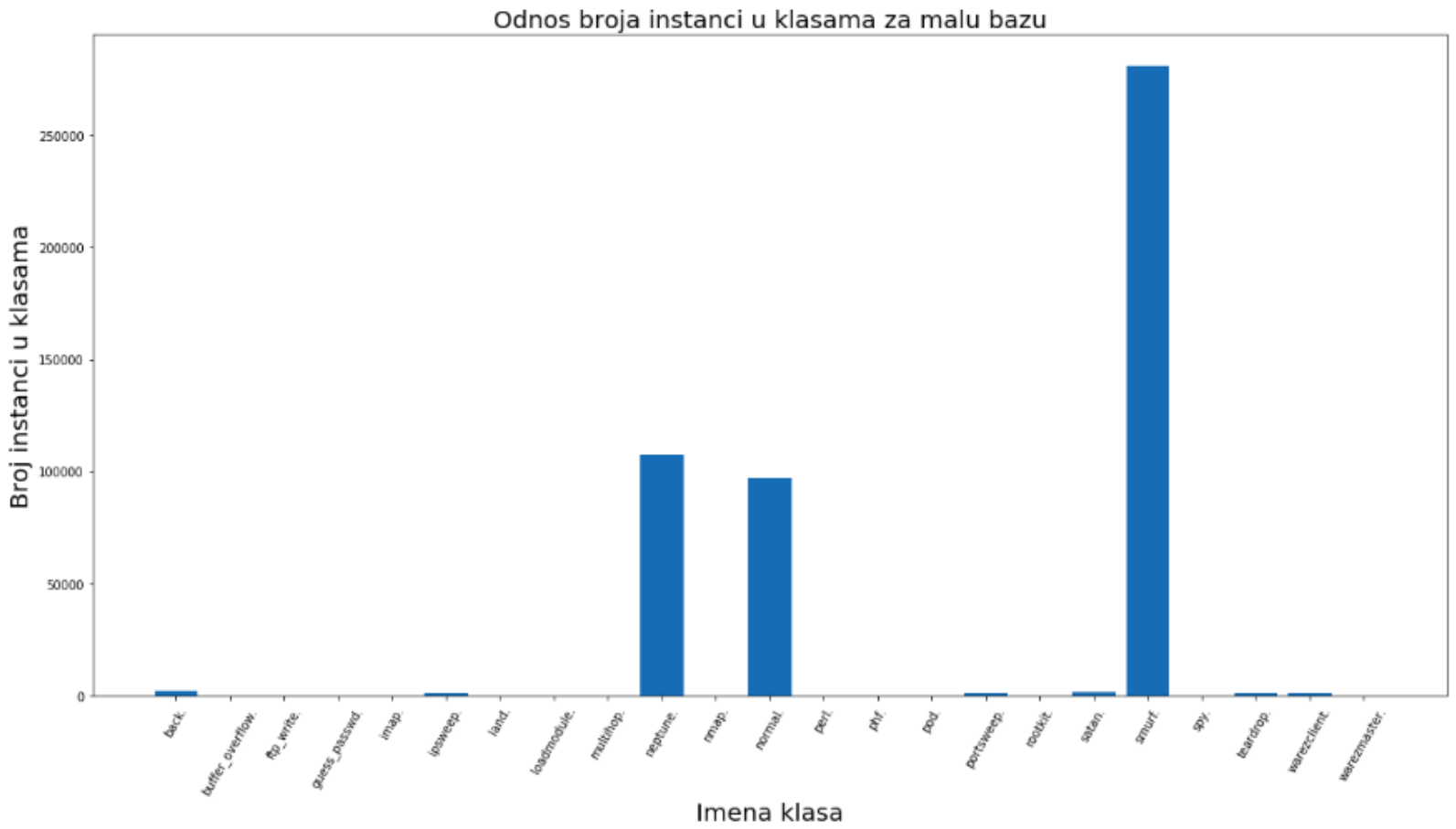
Uvod
Problem

- Normalne konekcije su napravljene da bi se prikazalo šta je očekivano u vojnoj mreži, a napadi potpadaju u jednu od četiri kategorije: User to Root; Remote to Local; Denial of Service; Probe.
  - **Denial of Service (dos):** Napadač pokušava da spreči legitimne korisnike da koriste servis.
  - **Remote to Local (r2l):** Napadač nema nalog na ciljanoj mašini, stoga pokušava da dobije pristup.
  - **User to Root (u2r):** Napadač ima lokalni pristup ciljanoj mašini i pokušava da dobije root privilegije.
  - **Probe:** Napadač pokušava da dobije informacije o ciljanom hostu.

Provera reprezentativnosti uzorka

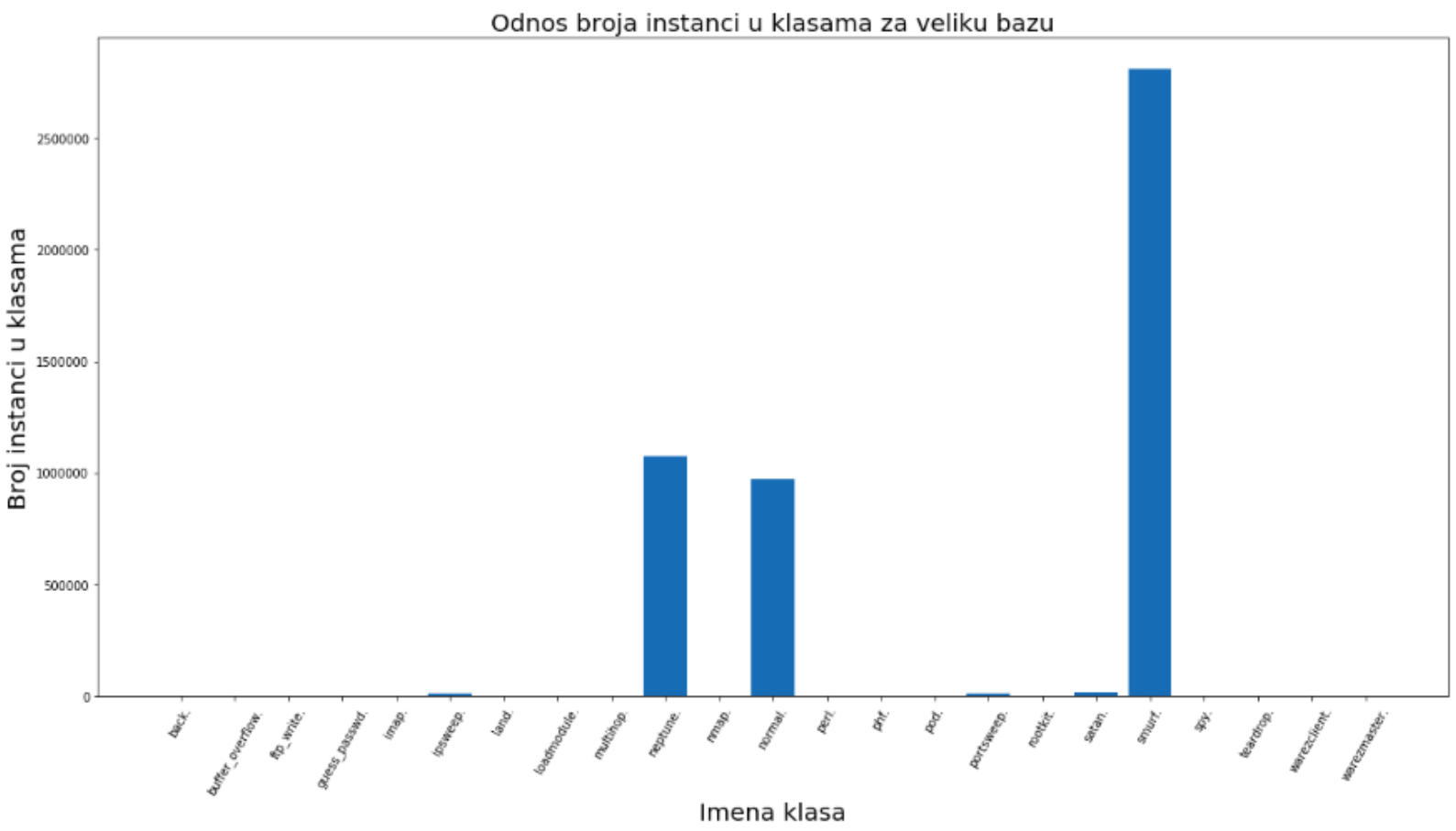
- Kao što je rečeno u navođenju problema, simulacija rada mreže trajala je 7 nedelja i pritom je skupljeno skoro 5 000 000 instanci. KDD99 skup podataka nudi i 10% tog skupa, što iznosi skoro 500 000 instanci. Može se pretpostaviti da je dati umanjeni skup reprezentativan uzorak podataka. Međutim, odlučile smo da proverimo da li je odnos broj instanci po klasama u maloj bazi sličan odnosu broja instanci po klasama u velikoj bazi.

Provera reprezentativnosti uzorka



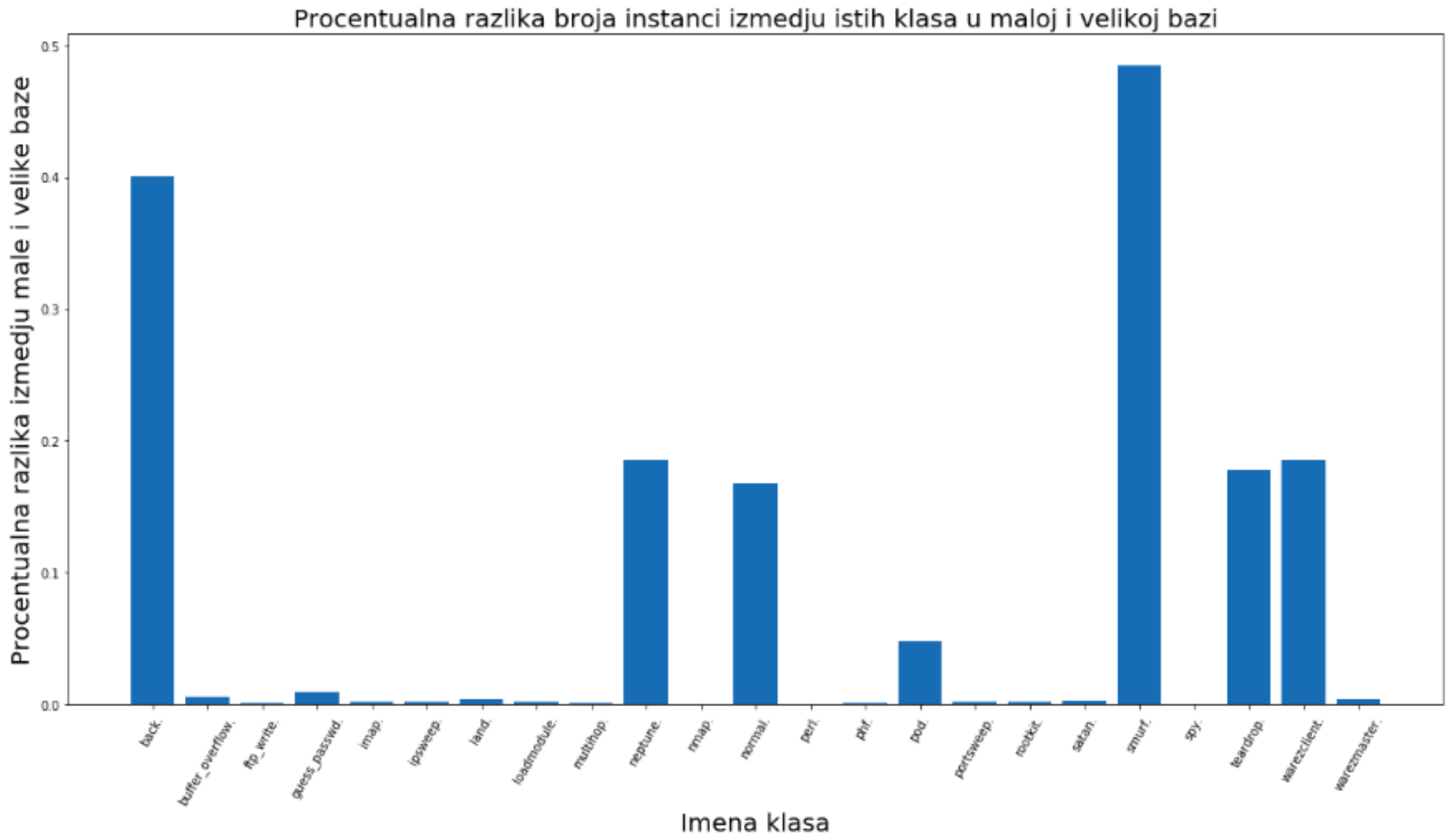
МАТЕМАТИЧКИ ФАКУЛТЕТ

Provera reprezentativnosti uzorka



МАТЕМАТИЧКИ ФАКУЛТЕТ

Provera reprezentativnosti uzorka

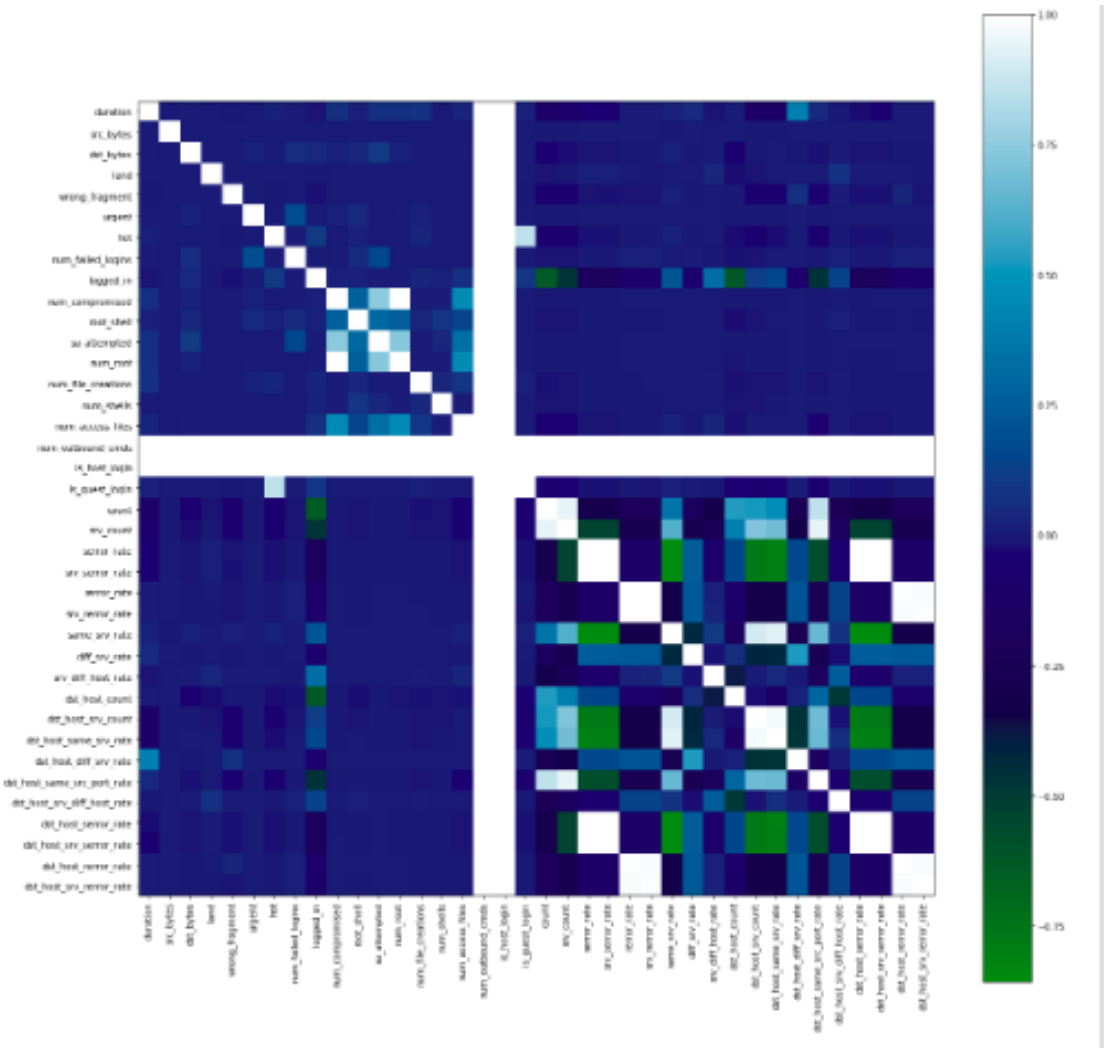


МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Binarna klasifikacija
Pretprocesiranje

- Želimo da klasifikujemo konekcije na: normalna i napad. Ciljnu promenljivo prebacujemo u 0 - normalna i 1 - napad.
- Primećujemo da je skup nebalansiran - ima više instanci koje su napad u odnosu na normalne. 97278 : 396743
- Skup podataka je podeljen na trening i test skup u odnosu 2:1, pri čemu je ciljna promenljiva stratifikovana.
- Posmatramo matricu korelacije na trening podacima:

Rešenje - Binarna klasifikacija
Pretprocesiranje



Rešenje - Binarna klasifikacija
Pretprocesiranje

- Tu primećujemo da su dva atributa potpuno pozitivno korelisana sa ostalim atributima. Što deluje veoma čudno.
- Kada pogledamo vrednosti ta dva atributa u matrici korelacije vidimo da su svuda vrednosti NaN. A preko statistika podataka vidimo da su ta dva atributa zapravo svuda 0.
- S obzirom na to da ne nose nikakvu informaciju, izbačeni su iz trening i test skupa.

Rešenje - Binarna klasifikacija (PCA)
Pretprocesiranje

- Želimo da primenimo PCA dekompoziciju podataka. Stoga, izdvajamo 3 kategorička atributa iz skupa podataka - 'protocol_type', 'service', 'flag' i odvojeno ih enkodiramo koristeći BinaryEncoding iz paketa category encoders.
- Na ostatak atributa primenjujemo standardizaciju i PCA algoritam sa željenim brojem komponenti 20. Na kraju nadovezujemo novodobijenih 20 atributa sa prethodno enkodiranim kategoričkim podacima.
- Primenom PCA algoritma zadržano je 97% informacija.

Rešenje - Binarna klasifikacija (PCA)
Formiranje mreže i treniranje

- Neuronska mreža ima 3 sloja - prvi ima 150 neurona, drugi 50, a treći jedan. Na prva dva sloja se nalazi ispravljena linearna aktivaciona funkcija, a na poslednjem je sigmoidna.
- Kao optimizator je izabran Adam, sa learning rate 0.0001. Funkcija gubitka je binary crossentropy a kao mera kvaliteta je izabrana tačnost.
- Mreža se trenira u 60 epoha, sa veličinom paketića 32, a među trening podacima je uzeto 20% podataka koji će služiti za validaciju rada mreže.

Rešenje - Binarna klasifikacija (PCA)
Grafik funkcije gubitka

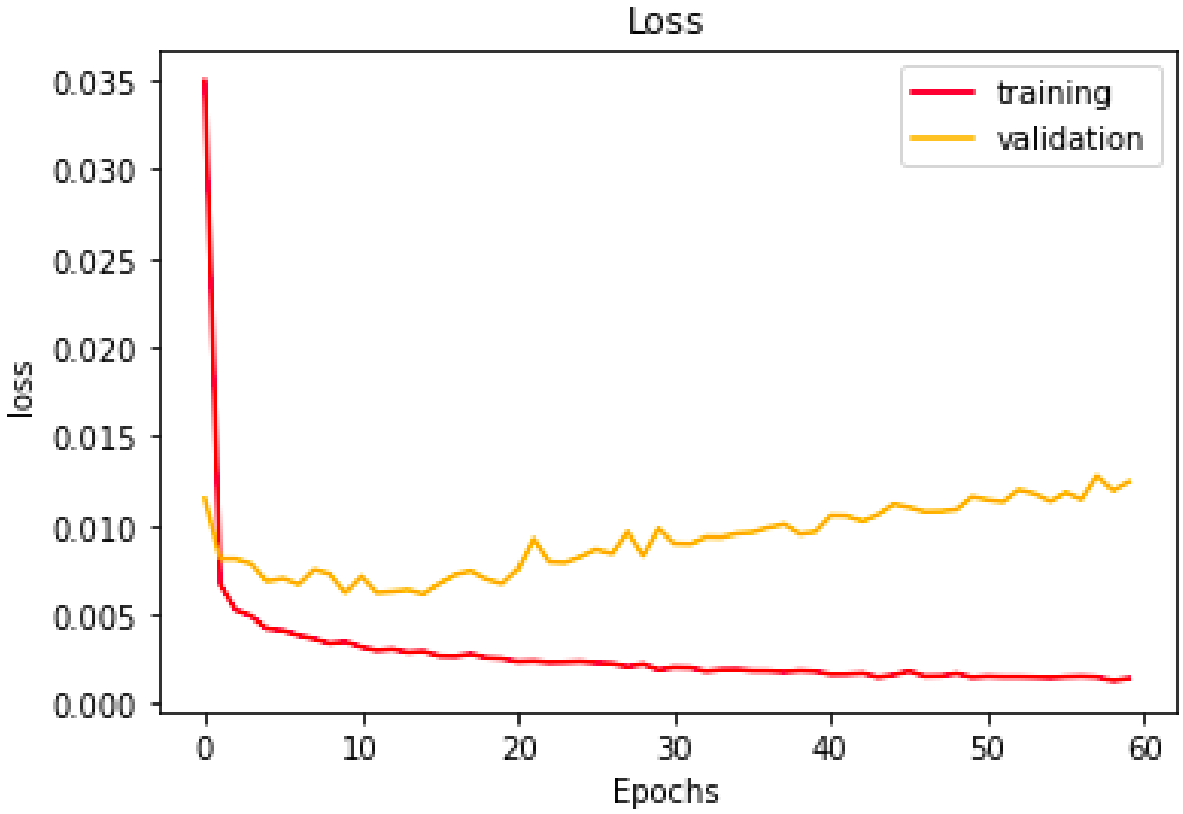


Figure: Funkcija gubitka

Rešenje - Binarna klasifikacija (PCA)
Grafik tačnosti

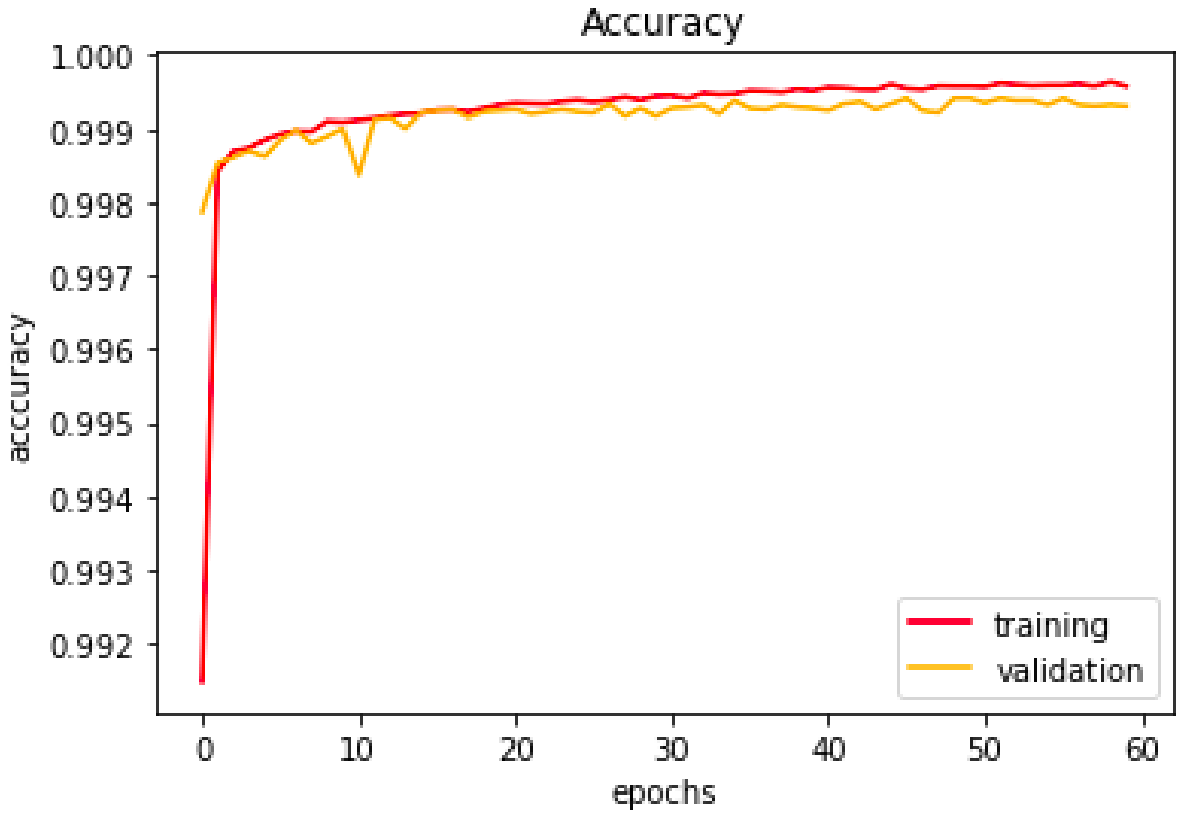


Figure: Tačnost

Rešenje - Binarna klasifikacija (PCA)
Zaključci o modelu

- Na osnovu prethodna dva grafika, uviđa se oscilacija na validacionom skupu što ukazuje na preprilagođavanje. Zato se odlučujemo na treniranje sa manjim brojem epoha.

Rešenje - Binarna klasifikacija (PCA)
Novi model

- Pravi se nova neuronska mreža koja je po svim parametrima ista, izuzev po broju epoha. Isprobano je više različitih vrednosti za broj epoha i donet je zaključak da se mreža najbolje ponaša kad je broj epoha 20.

Rešenje - Binarna klasifikacija (PCA)
Novi model - Grafik funkcije gubitka

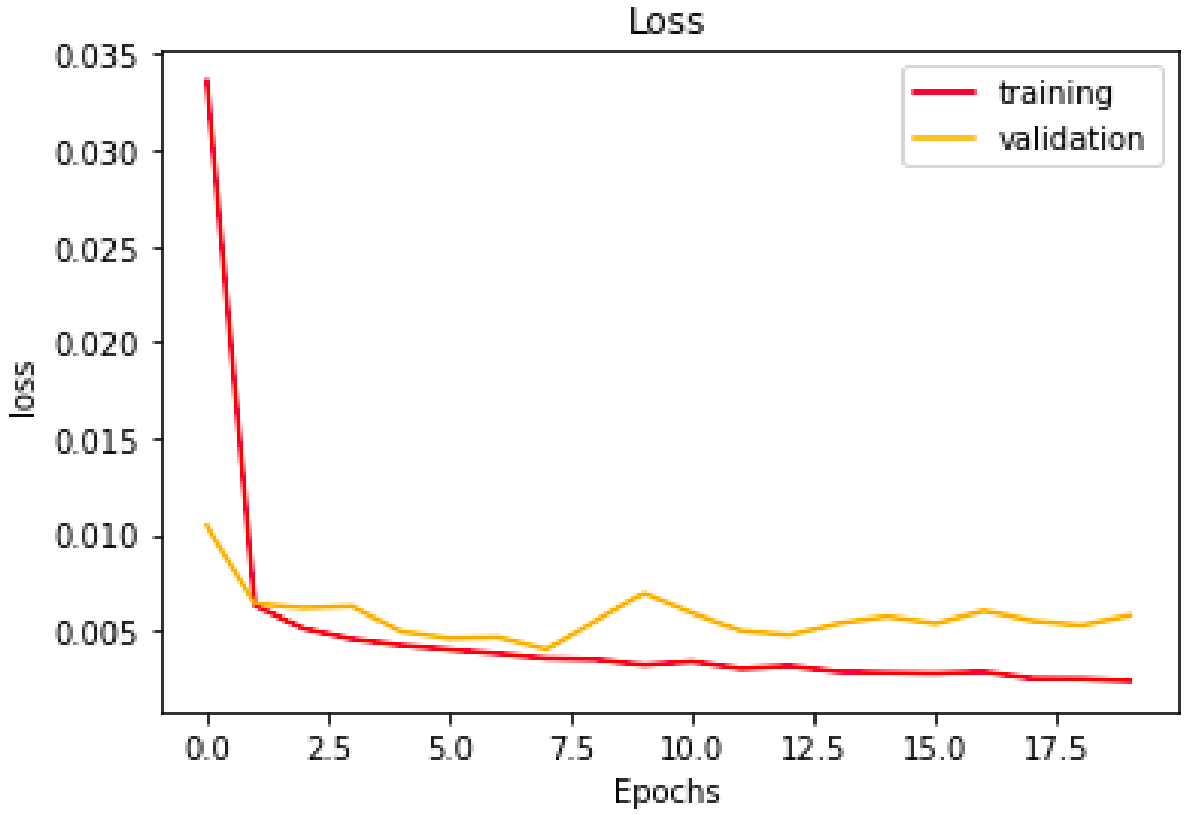


Figure: Funkcija gubitka

Rešenje - Binarna klasifikacija (PCA)
Novi model - Grafik tačnosti

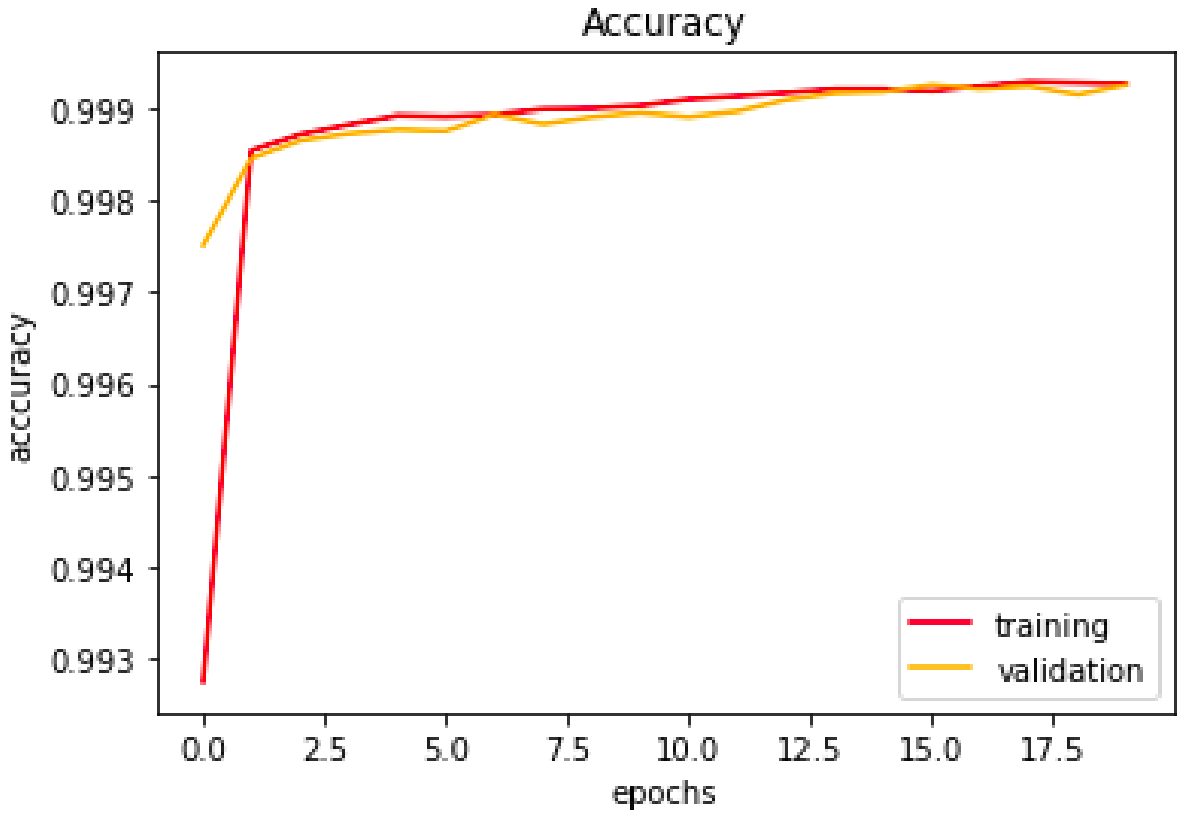


Figure: Tačnost

Rešenje - Binarna klasifikacija (PCA)
Evaluacija

- Na osnovu oba grafika vidimo da su vrednosti na validacionom skupu približne vrednostima na trening skupu. Stoga, smatramo da je ovaj model bolji od prethodnog.
- Vršimo predviđanje na test skupu. S obzirom na to da je aktivaciona funkcija na poslednjem sloju sigmoidna, potrebno je vrednosti veće ili jednake od 0.5 postaviti kao klasa 1, a vrednosti manje od 0.5 kao klasa 0.

Rešenje - Binarna klasifikacija (PCA)
Evaluacija

- Test loss: 0.0026373723189517676
- Test accuracy: 0.9992884397506714
- Train loss: 0.0027256275763326925
- Train accuracy: 0.9994018077850342
- Površina ispod ROC krive : 0.9990514742326431

```
[[ 32059     43]
 [    73 130852]]
```

Rešenje - Binarna klasifikacija (PCA)
Evaluacija

	precision	recall	f1-score	support
False	0.9977281	0.9986605	0.9981941	32102
True	0.9996715	0.9994424	0.9995569	130925
accuracy			0.9992885	163027
macro avg	0.9986998	0.9990515	0.9988755	163027
weighted avg	0.9992888	0.9992885	0.9992886	163027

Rešenje - Binarna klasifikacija (RFE)
Pretprocesiranje

- Pokušavamo sa novim modelom gde umesto PCA koristimo RFE algoritam za selekciju najboljih atributa.
- Za razliku od pripreme za PCA algoritam, u ovom slučaju primenjujemo enkodiranje i standardizaciju nad celim trening skupom i potom te podatke transformišemo pomoću RFE.
- Definišemo model linearne regresije i prosleđujemo RFE algoritmu uz naznaku da želimo da zadržimo 30 atributa.

Rešenje - Binarna klasifikacija (RFE)
Formiranje mreže i treniranje

- Nakon transformacije trening i test skupa formiramo mrežu.
- Mreža ima 3 sloja - prvi ima 150 neurona, drugi 50, a treći jedan. Na prva dva sloja se nalazi ispravljena linearna aktivaciona funkcija, a na poslednjem je sigmoidna.
- Kao optimizator je izabran Adam, sa learning rate 0.0001. Funkcija gubitka je binary crossentropy a kao mera kvaliteta je izabrana tačnost.
- Mreža se trenira u 50 epoha, sa veličinom paketića 32, a među trening podacima je uzeto 20% podataka koji će služiti za validaciju rada mreže.

Rešenje - Binarna klasifikacija (RFE)
Grafik funkcije gubitka

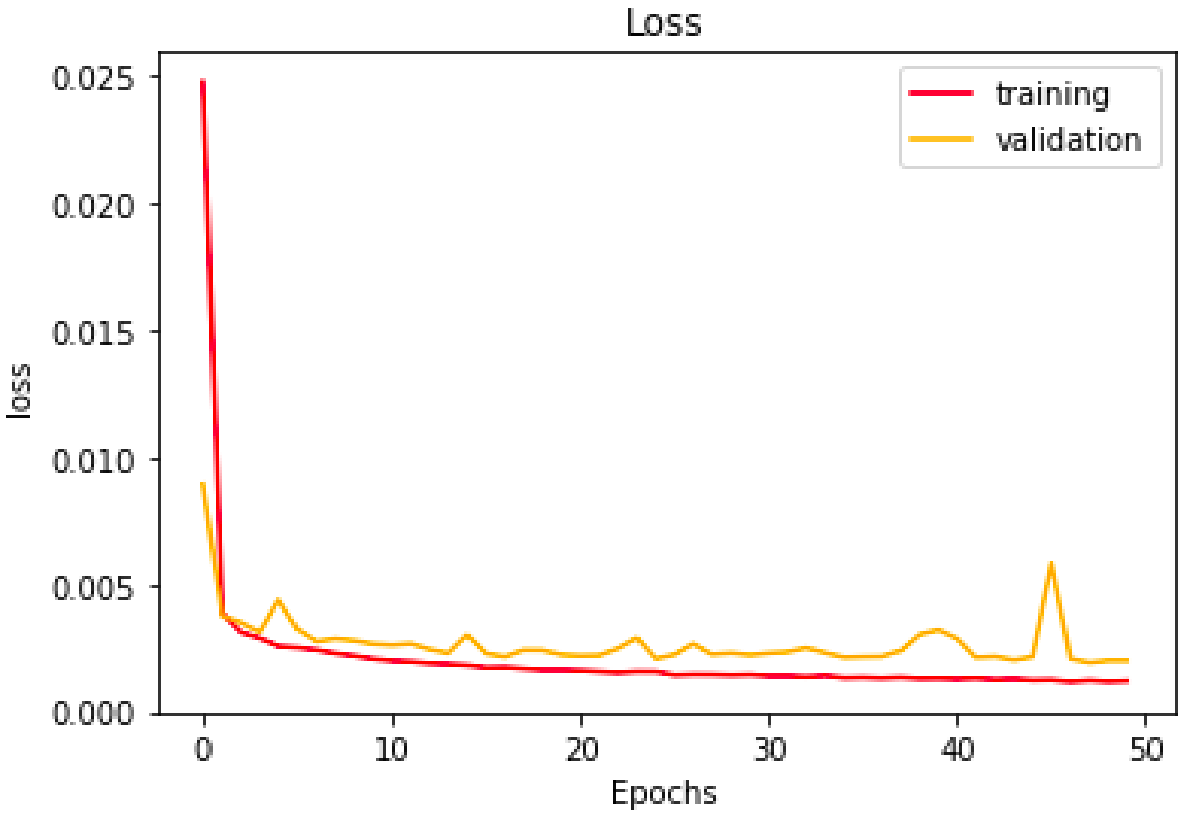


Figure: Funkcija gubitka

Rešenje - Binarna klasifikacija (RFE)
Grafik tačnosti

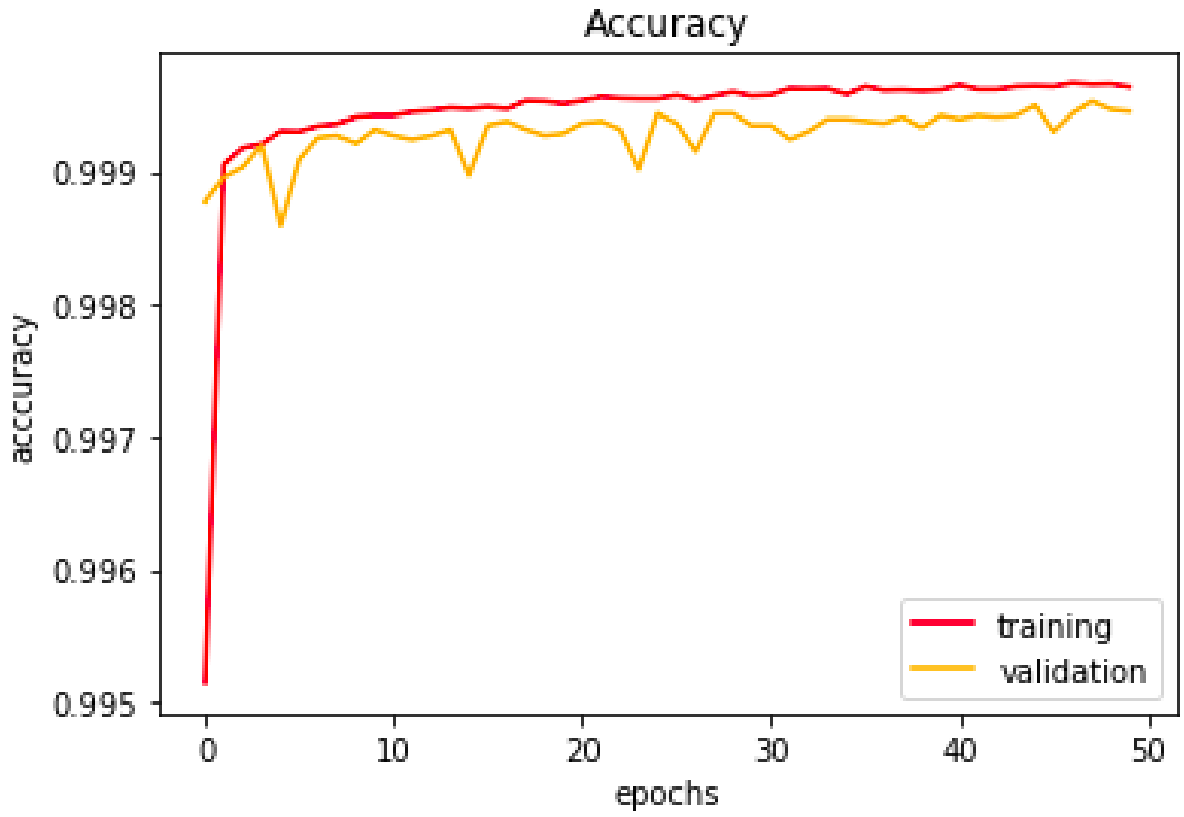


Figure: Funkcija gubitka

Rešenje - Binarna klasifikacija (RFE)
Zaključci o modelu

- Na osnovu prethodna dva grafika, uviđa se oscilacija na validacionom skupu što ukazuje na preprilagođavanje. Zato se odlučujemo na treniranje sa manjim brojem epoha.

Rešenje - Binarna klasifikacija (RFE)
Novi model

- Pravi se nova neuronska mreža koja je po svim parametrima ista, izuzev po broju epoha. Isprobano je više različitih vrednosti za broj epoha i donet je zaključak da se mreža najbolje ponaša kad je broj epoha 20.

Rešenje - Binarna klasifikacija (RFE)
Novi model - Graf funkcije gubitka

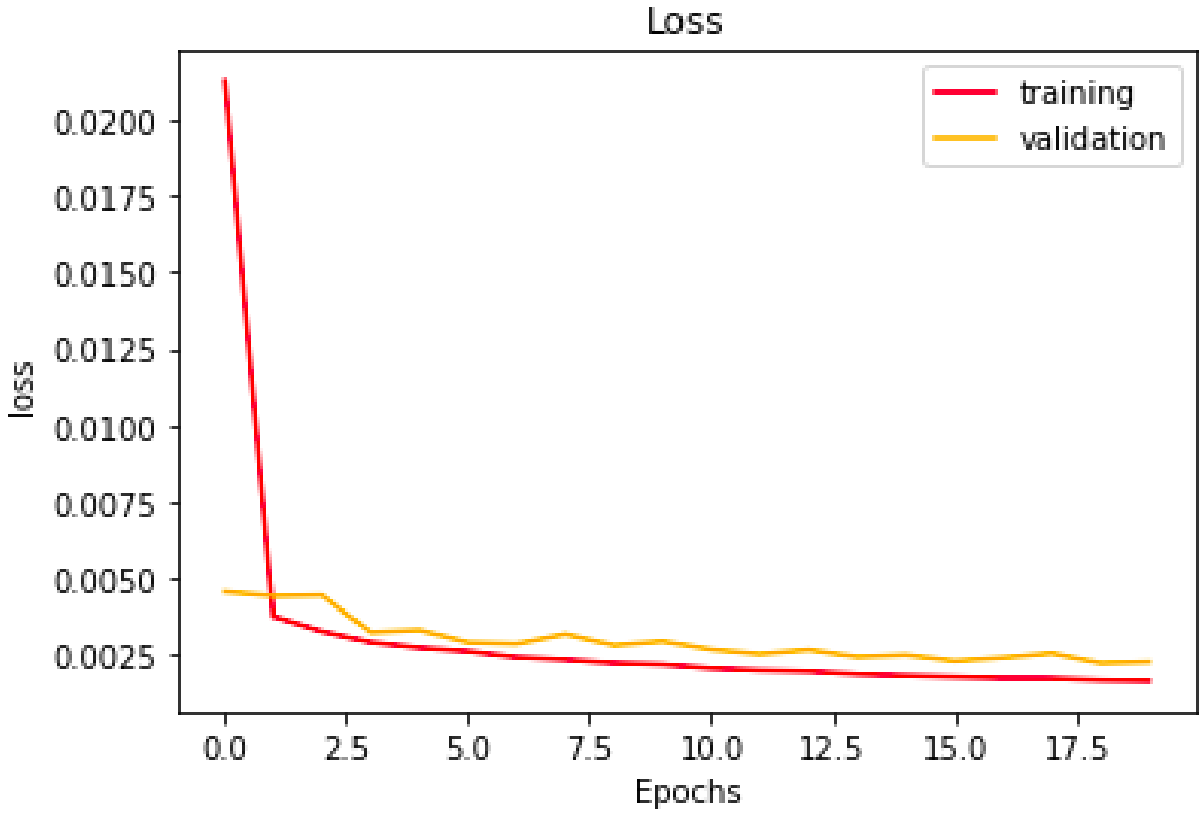


Figure: Funkcija gubitka

Rešenje - Binarna klasifikacija (RFE)
Novi model - Graf tačnosti

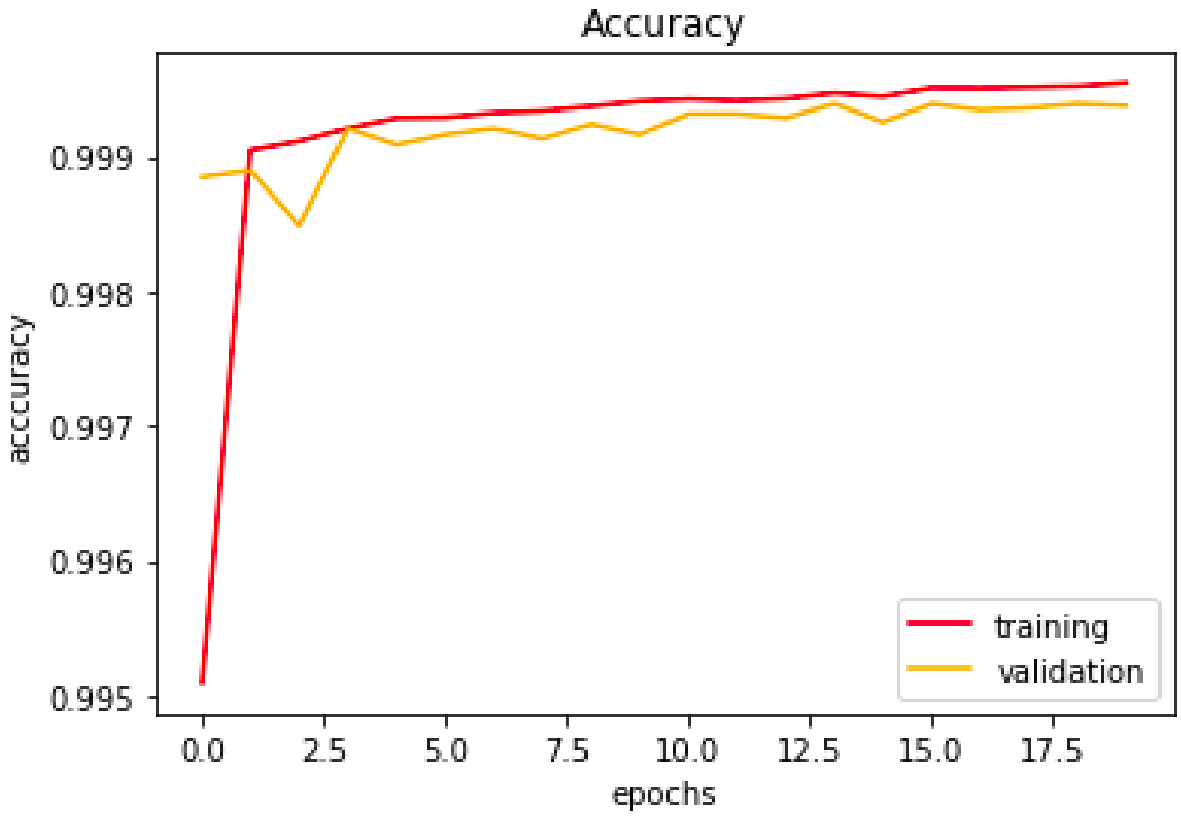


Figure: Tačnost

Rešenje - Binarna klasifikacija (RFE)
Evaluacija

- Test loss: 0.0023511093301084884
- Test accuracy: 0.999417245388031
- Train loss: 0.0016062885237643056
- Train accuracy: 0.999589256286621
- Površina ispod ROC krive : 0.999343287537729

```
[[ 32077     25]
 [    70 130855]]
```

МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Binarna klasifikacija (RFE)
Evaluacija

	precision	recall	f1-score	support
False	0.9978225	0.9992212	0.9985214	32102
True	0.9998090	0.9994653	0.9996371	130925
accuracy			0.9994173	163027
macro avg	0.9988157	0.9993433	0.9990793	163027
weighted avg	0.9994178	0.9994173	0.9994174	163027

МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Višeklasna klasifikacija
Pretprocesiranje

- Na sledećoj slici je dat raspored klasa po broju instanci.

```
smurf.              280790
neptune.            107201
normal.              97278
back.                 2203
satan.                1589
ipsweep.              1247
portsweep.            1040
warezclient.          1020
teardrop.              979
pod.                   264
nmap.                  231
guess_passwd.           53
buffer_overflow.        30
land.                   21
warezmaster.            20
imap.                   12
rootkit.                10
loadmodule.              9
ftp_write.               8
multihop.                7
phf.                     4
perl.                    3
spy.                     2
Name: class, dtype: int64
```

Rešenje - Višeklasna klasifikacija
Pretprocesiranje

- Nakon primene neuronske mreže na ovakav skup podataka utvrđeno je da mreža veoma loše uči klase sa malim brojem instanci, što je i očekivano. Stoga su klase sa jednocifrenim i dvocifrenim brojem instanci objedinjene u novu klasu pod imenom "other_attacks." Raspored klasa po broju instanci sada izgleda ovako:

```
smurf.           280790
neptune.         107201
normal.           97278
back.              2203
satan.             1589
ipsweep.           1247
portsweep.         1040
warezclient.       1020
teardrop.           979
pod.                264
nmap.               231
other_attacks.      179
Name: class, dtype: int64
```

Rešenje - Višeklasna klasifikacija
Pretprocesiranje

- Klase kodiramo pomoću cat.codes i kodiranje izgleda ovako:

```
{0: 'back.',
 1: 'ipsweep.',
 2: 'neptune.',
 3: 'nmap.',
 4: 'normal.',
 5: 'other_attacks.',
 6: 'pod.',
 7: 'portsweep.',
 8: 'satan.',
 9: 'smurf.',
 10: 'teardrop.',
 11: 'warezclient.'}
```

Rešenje - Višeklasna klasifikacija
Pretprocesiranje

- Skup podataka delimo na trening i test skup u odnosu 2:1 pri čemu vodimo računa o stratifikaciji ciljne promenljive.
- Kao i u binarnoj klasifikaciji izbacujemo atribute "is_host_login" i "num_outbound_cmds".
- Da bismo mogli primeniti PCA algoritam na podatke, isto kao u binarnoj klasifikaciji, delimo skup na dva skupa: sa kategoričkim i numeričkim atributima.
- Na kategoričke podatke primenjujemo binarno kodiranje, a numeričke standardizujemo.

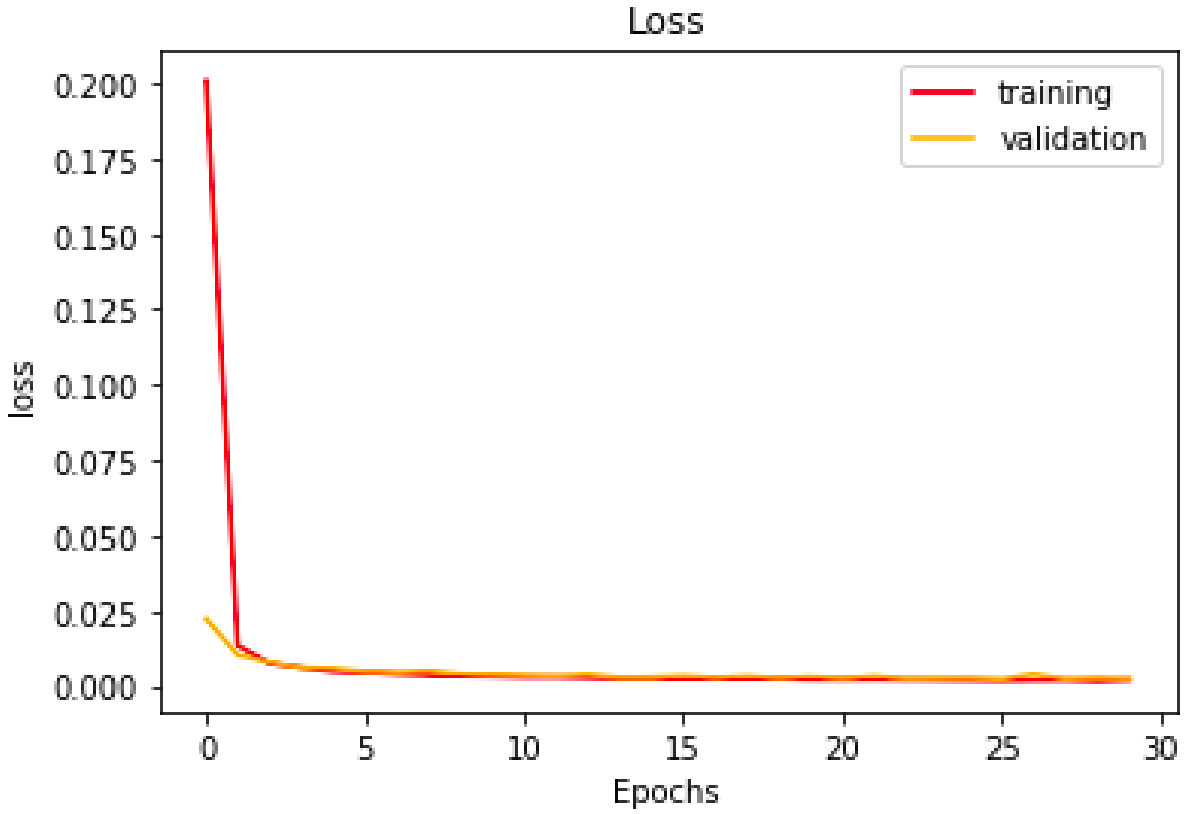
Rešenje - Višeklasna klasifikacija
Pretprocesiranje

- Primenjujemo PCA algoritam na numeričke podatke i biramo da nam broj komponenti bude 20. Ovaj broj je eksperimentalno utvrđen. Primenom PCA algoritma sačuvali smo 96% informacija.
- Na skup podataka dobijen primenom PCA algoritma nadovezujemo enkodirane kategoričke atribute.

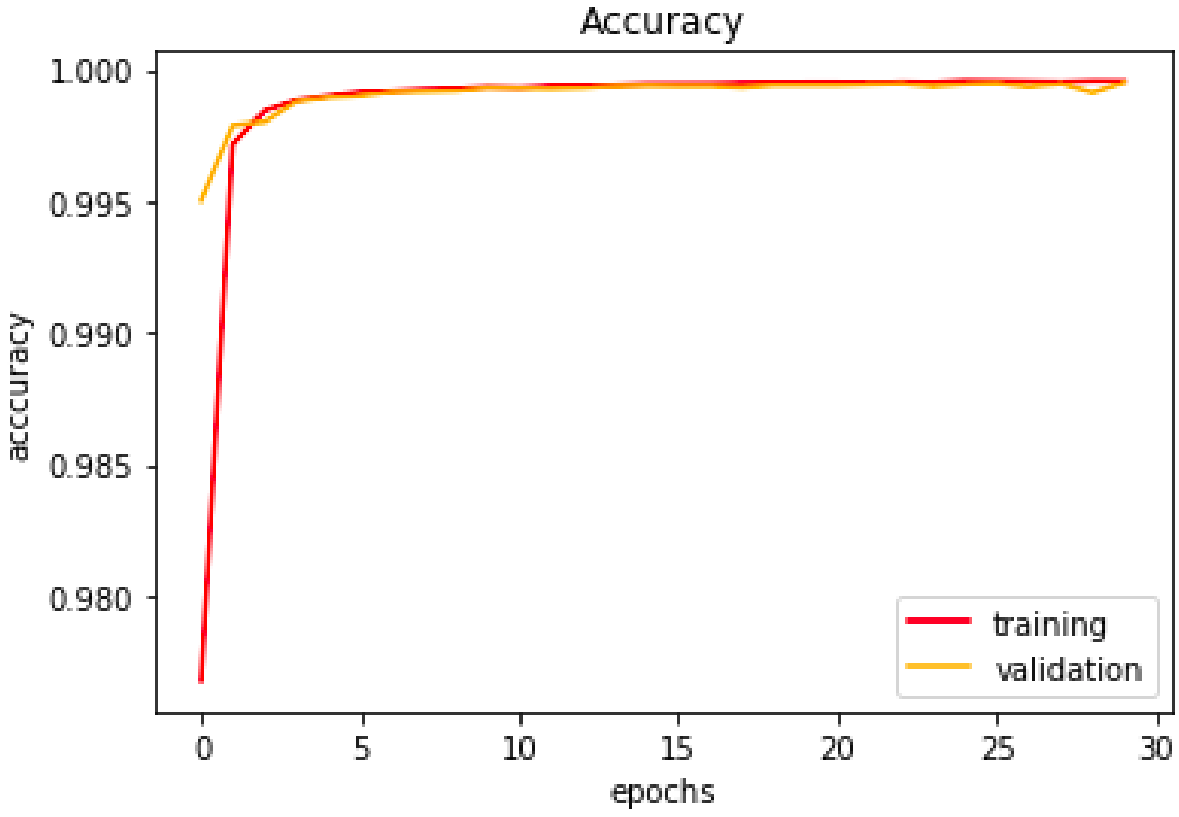
Rešenje - Višeklasna klasifikacija - Prvi model
Formiranje i treniranje mreže

- Neuronska mreža ima 3 sloja - prvi ima 150 neurona, drugi 50, a treći 12. Na prva dva sloja se nalazi ispravljena linearna aktivaciona funkcija, a na poslednjem je softmax.
- Kao optimizator je izabran Adam, sa learning rate 0.0001. Funkcija gubitka je category crossentropy a kao mera kvaliteta je izabrana tačnost.
- Mreža se trenira u 30 epoha, sa veličinom paketića 128, a među trening podacima je uzeto 20% podataka koji će služiti za validaciju rada mreže.

Rešenje - Višeklasna klasifikacija - Prvi model
Grafik funkcije gubitka



Rešenje - Višeklasna klasifikacija - Prvi model
Grafik tačnosti



Rešenje - Višeklasna klasifikacija - Prvi model
Evaluacija

- Test loss: 0.002280461698022376
- Test accuracy: 0.9994847774505615

```
[[  727     0     0     0     0     0     0     0     0     0     0     0]
 [    0   407     0     2     2     0     0     0     1     0     0     0]
 [    0     0 35375     0     1     0     0     0     0     0     0     0]
 [    0     1     0    72     3     0     0     0     0     0     0     0]
 [    3     6     2     0 32071     2     0     1     4     0     0    13]
 [    0     0     0     0    11    47     0     0     0     0     0     1]
 [    0     0     0     0     0     0    87     0     0     0     0     0]
 [    0     0     1     0     1     0     0   341     0     0     0     0]
 [    0     0     0     2     5     0     0     0   517     0     0     0]
 [    0     0     0     0     1     0     0     0     0 92660     0     0]
 [    0     0     0     0     0     0     0     0     0     0   323     0]
 [    0     0     0     0    21     0     0     0     0     0     0   316]]
```

МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Višeklasna klasifikacija - Prvi model
Evaluacija

	precision	recall	f1-score	support
0	1.00	1.00	1.00	727
1	0.98	0.98	0.98	412
2	1.00	1.00	1.00	35376
3	0.95	0.95	0.95	76
4	1.00	1.00	1.00	32102
5	0.98	0.68	0.80	59
6	1.00	1.00	1.00	87
7	0.99	0.99	0.99	343
8	0.98	0.98	0.98	524
9	1.00	1.00	1.00	92661
10	1.00	1.00	1.00	323
11	0.97	0.93	0.95	337
accuracy			1.00	163027
macro avg	0.99	0.96	0.97	163027
weighted avg	1.00	1.00	1.00	163027

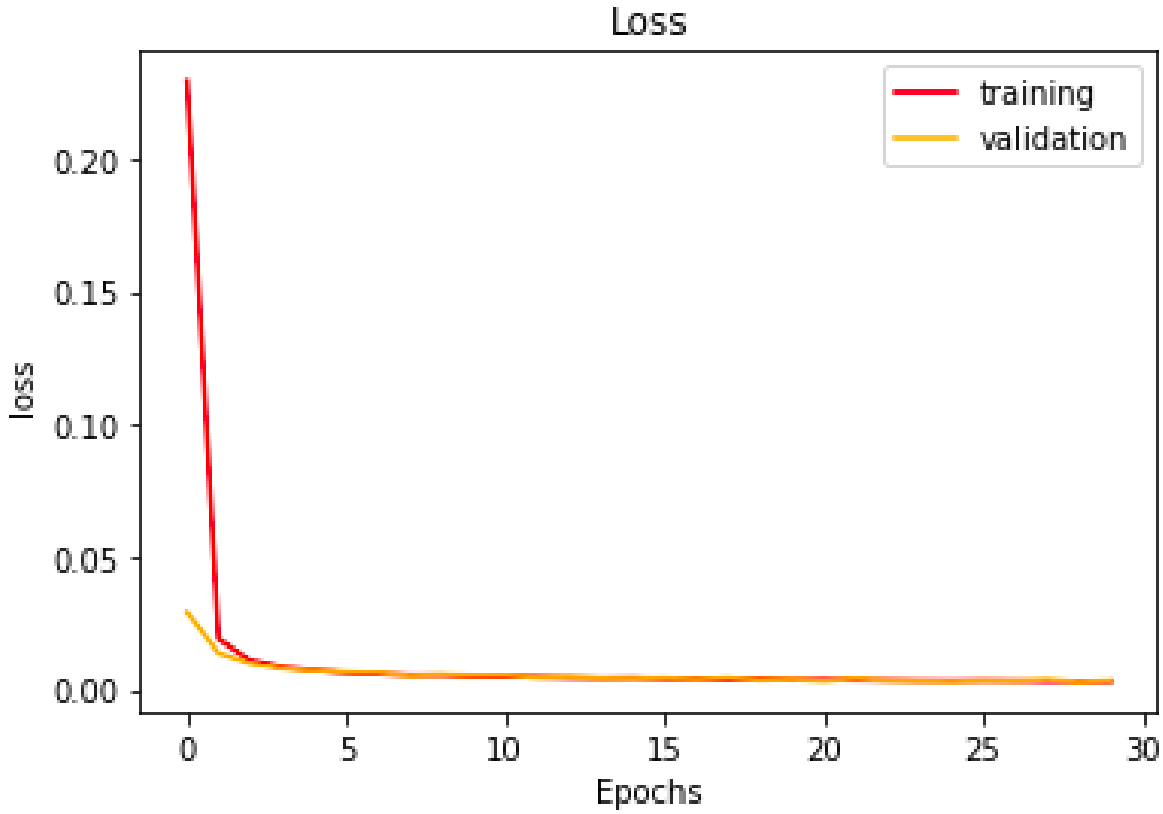
Rešenje - Višeklasna klasifikacija - Prvi model
Gde model greši

- 21 instancu koja je zapravo napad 'warezclient.' model je klasifikovao kao normalnu.
- 13 instanci koje su zapravo normalna konekcija model je klasifikovao kao napad 'warezclient.'
- 11 instanci koje pripadaju klasi 'other_attacks.' model je klasifikovao kao normalnu.
- 6 instanci koje pripadaju normalnoj konekciji model je klasifikovao kao napad 'ipsweep.'
- 5 instanci koje pripadaju napadu 'satan.' model je klasifikovao kao normalnu konekciju.

Rešenje - Višeklasna klasifikacija - PCA
Formiranje i treniranje mreže

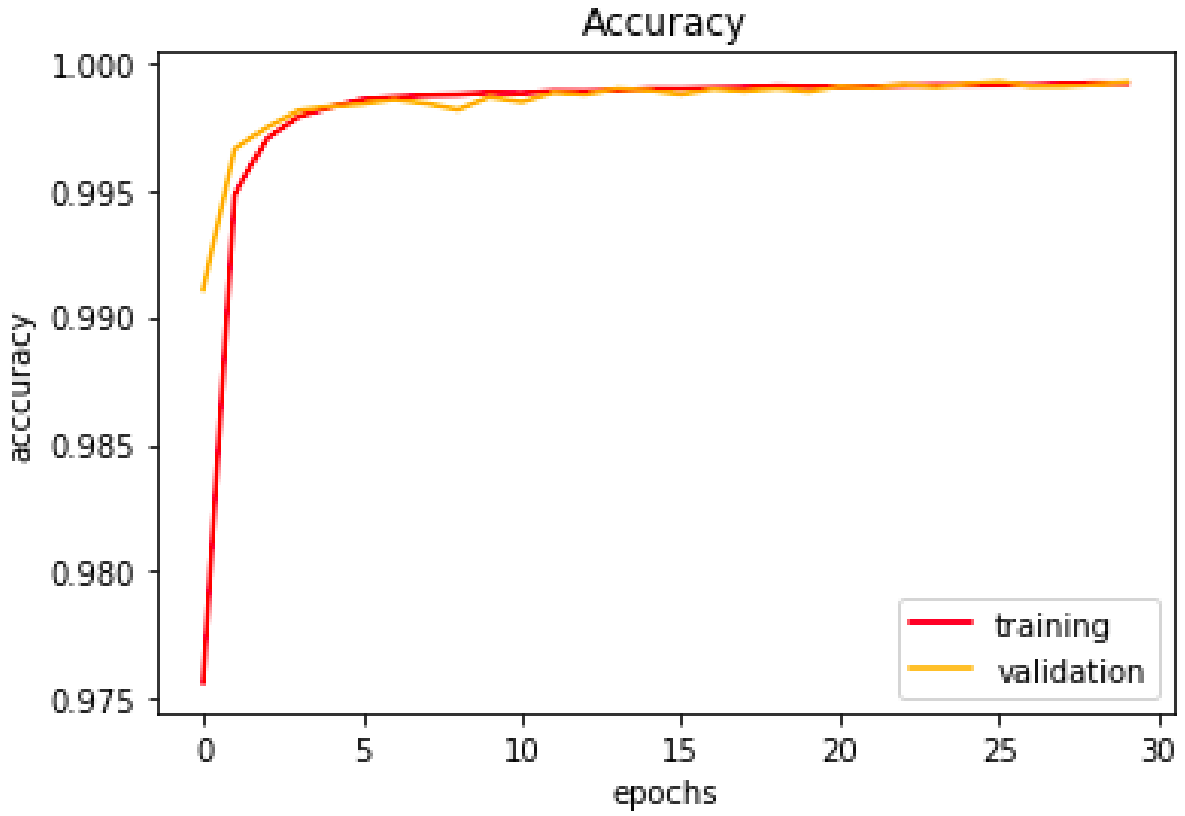
- Neuronska mreža ima 3 sloja - prvi ima 150 neurona, drugi 50, a treći 12. Na prva dva sloja se nalazi ispravljena linearna aktivaciona funkcija, a na poslednjem je softmax.
- Kao optimizator je izabran Adam, sa learning rate 0.0001. Funkcija gubitka je category crossentropy a kao mera kvaliteta je izabrana tačnost.
- Mreža se trenira u 30 epoha, sa veličinom paketića 128, a među trening podacima je uzeto 20% podataka koji će služiti za validaciju rada mreže.
- Ova mreža je ista kao i prethodna, osim što je ulazna dimenzija prilagođena podacima nakon PCA i trenira se na podacima nad kojima je primenjen PCA.

Rešenje - Višeklasna klasifikacija - PCA
Grafik funkcije gubitka



МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Višeklasna klasifikacija - PCA
Grafik tačnosti



Rešenje - Višeklasna klasifikacija - PCA
Evaluacija

- Test loss: 0.0039198086932115675
- Test accuracy: 0.9990983093383795

```
[[  724     0     0     0     3     0     0     0     0     0     0     0]
 [    0   407     0     2     3     0     0     0     0     0     0     0]
 [    0     0 35375     1     0     0     0     0     0     0     0     0]
 [    0     1     0    64    11     0     0     0     0     0     0     0]
 [    9    11     3     0 32041     5     0     2     1     1     0    29]
 [    0     0     1     0    13    42     0     0     0     0     0     3]
 [    0     0     0     0     3     0    84     0     0     0     0     0]
 [    0     0     2     0     1     0     0   340     0     0     0     0]
 [    0     0     0     0    10     0     0     7   507     0     0     0]
 [    0     0     0     0     2     0     0     0     0 92659     0     0]
 [    0     0     0     0     0     0     0     0     0     0   323     0]
 [    0     1     0     0    22     0     0     0     0     0     0   314]]
```

МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Višeklasna klasifikacija - PCA
Evaluacija

	precision	recall	f1-score	support
0	0.99	1.00	0.99	727
1	0.97	0.99	0.98	412
2	1.00	1.00	1.00	35376
3	0.96	0.84	0.90	76
4	1.00	1.00	1.00	32102
5	0.89	0.71	0.79	59
6	1.00	0.97	0.98	87
7	0.97	0.99	0.98	343
8	1.00	0.97	0.98	524
9	1.00	1.00	1.00	92661
10	1.00	1.00	1.00	323
11	0.91	0.93	0.92	337
accuracy			1.00	163027
macro avg	0.97	0.95	0.96	163027
weighted avg	1.00	1.00	1.00	163027

Rešenje - Višeklasna klasifikacija - PCA
Gde model greši

- 29 instanci koje pripadaju normalnoj konekciji model je klasifikovao kao napad 'warezclient.'
- 22 instance koje pripadaju napadu 'warezclient.' model je klasifikovao kao normalnu konekciju.
- 13 instanci koje pripadaju napadu 'other_attacks' model je klasifikovao kao normalnu konekciju.
- 11 instanci koje pripadaju napadu 'nmap' model je klasifikovao kao normalnu konekciju.
- 11 instanci koje pripadaju normalnoj konekciji model je klasifikovao kao napad 'ipsweep.'.
- 10 instanci koje pripadaju napadu 'satan.' model je klasifikovao kao normalnu konekciju.

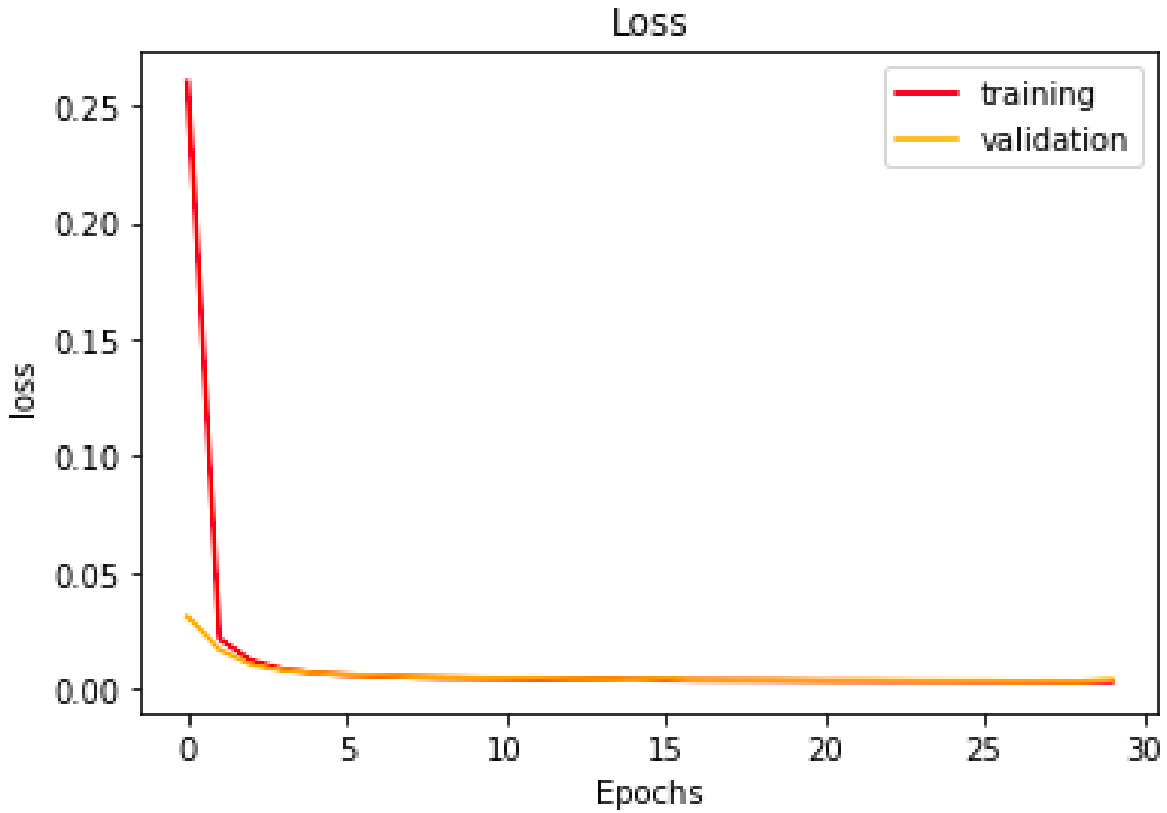
Rešenje - Višeklasna klasifikacija - RFE
Pretprocesiranje

- Podaci su enkodirani i standardizovani na odgovarajući način.
- Formiramo model linearne regresije i prosleđujemo algoritmu RFE kao i broj atributa koje želimo da izabere. Eksperimentalno je utvrđeno da najbolje rezultate daje za 30 atributa. Nakon toga se podaci transformišu u skladu sa rezultatima RFE-a.

Rešenje - Višeklasna klasifikacija - RFE
Formiranje i treniranje mreže

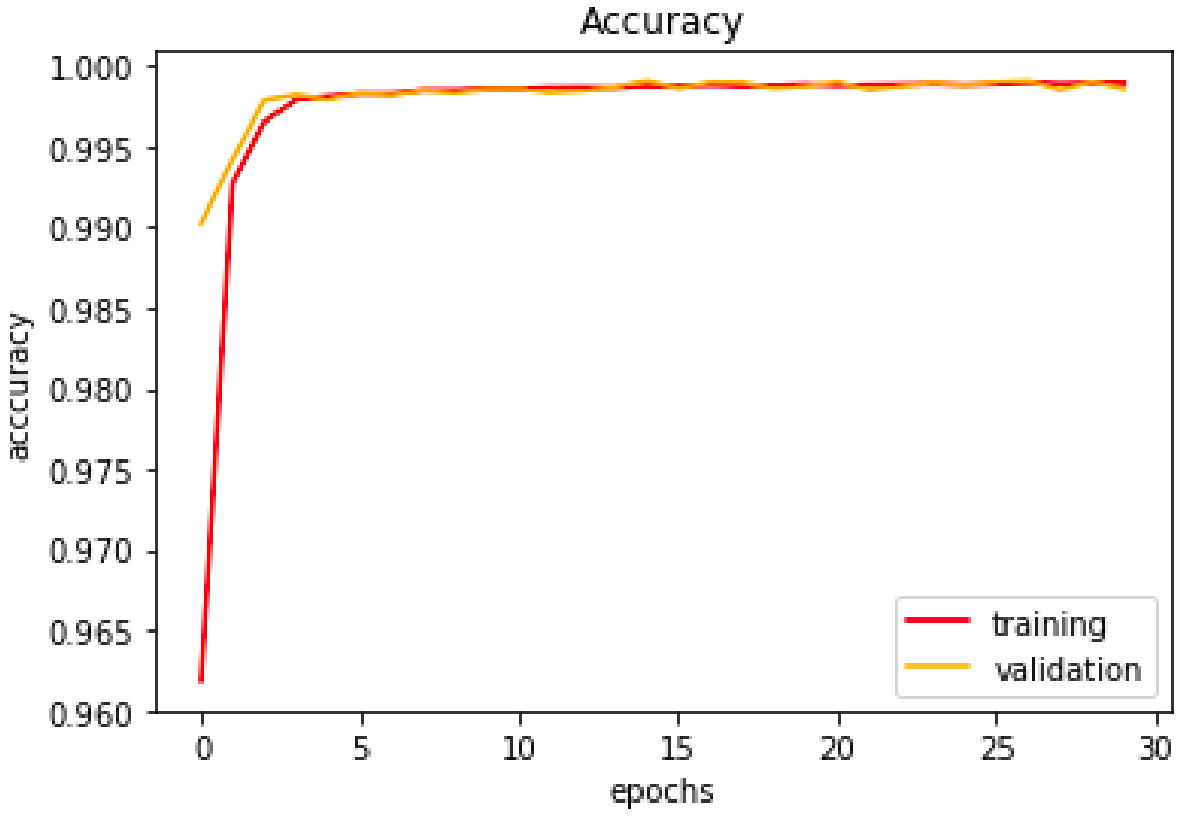
- Neuronska mreža ima 3 sloja - prvi ima 150 neurona, drugi 50, a treći 12. Na prva dva sloja se nalazi ispravljena linearna aktivaciona funkcija, a na poslednjem je softmax.
- Kao optimizator je izabran Adam, sa learning rate 0.0001. Funkcija gubitka je category crossentropy a kao mera kvaliteta je izabrana tačnost.
- Mreža se trenira u 30 epoha, sa veličinom paketića 128, a među trening podacima je uzeto 20% podataka koji će služiti za validaciju rada mreže.
- Ova mreža je ista kao i prethodna, osim što je ulazna dimenzija prilagođena podacima nakon RFE i trenira se na podacima nad kojima je primenjen RFE.

Rešenje - Višeklasna klasifikacija - RFE
Grafik funkcije gubitka



МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Višeklasna klasifikacija - RFE
Grafik tačnosti



Rešenje - Višeklasna klasifikacija - RFE
Evaluacija

- Test loss: 0.004489319530296752
- Test accuracy: 0.9984971880912781

```
[[  725     0     0     0     0     2     0     0     0     0     0     0]
 [    0   382     0    26     3     0     0     0     1     0     0     0]
 [    0     0 35372     0     1     2     0     1     0     0     0     0]
 [    0     1     0    72     3     0     0     0     0     0     0     0]
 [    2     3     2     3 32076     4     0     1     7     0     0     4]
 [    0     0     1     0    19    37     0     0     0     0     0     2]
 [    0     0     0     0     0     0    87     0     0     0     0     0]
 [    0     0     0     0     3     0     0   340     0     0     0     0]
 [    1     0     0     2     6     0     0     0   515     0     0     0]
 [    0     0     0     0    17     0     0     0     0 92644     0     0]
 [    0     0     0     0     0     0     0     0     0     0   323     0]
 [    0     0     0     0   128     0     0     0     0     0     0   209]]
```

МАТЕМАТИЧКИ ФАКУЛТЕТ

Rešenje - Višeklasna klasifikacija - RFE
Evaluacija

	precision	recall	f1-score	support
0	1.00	1.00	1.00	727
1	0.99	0.93	0.96	412
2	1.00	1.00	1.00	35376
3	0.70	0.95	0.80	76
4	0.99	1.00	1.00	32102
5	0.82	0.63	0.71	59
6	1.00	1.00	1.00	87
7	0.99	0.99	0.99	343
8	0.98	0.98	0.98	524
9	1.00	1.00	1.00	92661
10	1.00	1.00	1.00	323
11	0.97	0.62	0.76	337
accuracy			1.00	163027
macro avg	0.95	0.92	0.93	163027
weighted avg	1.00	1.00	1.00	163027

Rešenje - Višeklasna klasifikacija - RFE
Gde model greši

- 128 instanci koje pripadaju napadu 'warezclient.' model je klasifikovao kao normalnu konekciju.
- 26 instanci koje pripadaju napadu 'ipsweep.' model je klasifikovao kao napad 'nmap.'.
- 19 instanci koje pripadaju napadu 'other_attacks.' model je klasifovao kao normalnu konekciju.
- 17 instanci koje pripadaju napadu 'smurf.' model je klasifikovao kao normalnu konekciju.

Karakteristike sistema

- Processor: Intel Core i3-6100U CPU @ 2.30GHz x 4
- Graphics: Intel HD Graphics 520 (Skylake GT2)
- Memory: 7,2 GiB
- GNOME: 3.28.2
- OS: Ubuntu 18.04 LTS
- OS type: 64-bit
- Disk: 130,1 GB

МАТЕМАТИЧКИ ФАКУЛТЕТ

Eksperimentalni rezultati i izazovi

- Šta se dešava kada promenimo i kombinujemo nekeparametre? Menjali smo broj slojeva, čvorova u sloju, iteracija,kao i funkciju aktivacije.
- Broj čvorova u sloju kao i broj iteracija ne utiče znatno napreciznost i udeo stvarno pozitivnih vrednosti u mreži
- Međutim, broj iteracija utiče na preprilagođavanje mreže, stoga je bilo potrebno isprobati različite vrednosti. U svim modelima je započeto treniranje sa većim brojem epoha i ta vrednost se u svim modelima pokazala sklona preprilagođavanju. Kod binarne klasifikacije PCA, kao i kod binarne klasifikacije RFE, eksperimentalno je utvrđeno da najbolji rezultat daje za 20 epoha. Kod višeklasne klasifikcije se 30 epoha pokazalo kao dobar izbor u svim modelima.

Eksperimentalni rezultati i izazovi

- Isprobana je binarna klasifikacija bez smanjivanja dimenzionalnosti podataka, međutim, rezultati su mnogo bolji kada se prethodno primeni PCA ili RFE algoritam.
- Kod višeklasne isprobana je višeklasna nad neizmenjenim skupom podataka i očekivano je veoma loše klasifikovala klase sa malim brojem instanci. Isprobana je i višeklasna sa PCA, ali ni to nije dalo mnogo bolje rezultate.
- Potom je pokušano da se samo objedine klase sa jednocifrenim brojem instanci i da se primeni obična višeklasna, to je bolje radilo od prethodnih modela, međutim, i dalje su neke manje klase bile mnogo lošije klasifikovane. Nakon toga su objedinjene klase sa jednocifrenim i dvocifrenim brojem instanci i to je pokazalo daleko bolje rezulate.

Eksperimentalni rezultati i izazovi

- Na osnovu toga je donet zaključak da je objedinjavanje klasa neophodno pa su sva tri modela koja su izabrana u projektu rađena nad tim skupom podataka.

Zaključak

- Veštačke neuronske mreže su veoma moćan alat, ali treba paziti na preprilagođavanje i potprilagođavanje.
- Network Intrusion Data je i dalje jedan on najpopularnijih i najboljih skupova podataka za detekciju upada na mreže.
- Veći izazov je predstavljala binarna klasifikacija, verovatno zbog razlika u instancama koje pripadaju različitim napadima.
- Višeklasna klasifikacija je nakon objedinjavanja klasa sa jednocifrenim i dvocifrenim brojem instanci bila bolja od binarne.

Zaključak

- Kod višeklasne sva tri modela su najviše mešala normalnu konekciju i napad "warezclient.". Warezclient je napad koji se sa stoji u dovlačenju (skidanju) datoteka sa FTP servera, a takva dinamika napada izgleda kao legalan proces. S obzirom da to nije ilegalan proces, nije začuđujuće da je model propustio da ga prepozna kao napad.
- PCA i RFE algoritmi su veoma moćni alati i doprinose stabilnosti rešenja.

Literatura

[1] Atributi. on-line at: https://kdd.ics.uci.edu/databases/kddcup99/task.html.
[2] Napadi. on-line at: https://web.cs.dal.ca/zincir/bildiri/pst05-gnm.pdf.
[3] Muniyandi R. C. Azzawi M. A. Abdulmajed, E. S. A Novel Method of Preprocessing and Evaluating the KDD CUP 99 Data Set. Adv Research in Dynamical Control Systems, 9(10), 2017.
[4] A. Engelbrecht. Computational Intelligence - An Introduction. John Willey Sons, 2007.

МАТЕМАТИЧКИ ФАКУЛТЕТ