Predikcija pola korisnika mobilnih telefona
-Projektni zadatak-

Ivana Stojanović, EE59/2014, ivanastojanovic95@gmail.com
Ljiljana Popović, EE72/2014, ljiljana.popovic995@gmail.com

I. UVOD

Mobilni telefoni su danas široko rasprostranjeni i predstavljaju našu svakodnevnicu. Iako im je osnovno svojstvo da obezbede brzu i laku komunikaciju na velikim udaljenostima, danas su razvijeni i za druge primene. Sadrže setove ugrađenih senzora, koji otvaraju vrata ka novim istraživanjima i vode ka razvoju senzorskih aplikacija koje će uneti velike promene u naš svakodnevni život. Na primer, mogu se napraviti aplikacije koje iz velike količine podataka izvlače relevantne informacije koje utiču na dalji tok izvršavanja. Analiza podataka koji su dobijeni na ovaj način doprinosi boljem naučnom razumevanju ljudi i društvenih odnosa, unapređenju mobilnih usluga i tehnološkim inovacijama.

Veoma je važno obezbediti poverljivost podataka, te je neophodno sprečiti svaku zloupotrebu. Iako se korisnik svakodnevno susreće sa ovim problemom često ga nije svestan, jer se zapravo njegovi podaci koriste za različite statističke analize i istraživanja, koja ne utiču na njega. Problem nastaje kada upotreba ovih podataka narušava privatnost i ugoržava identitet korisnika.

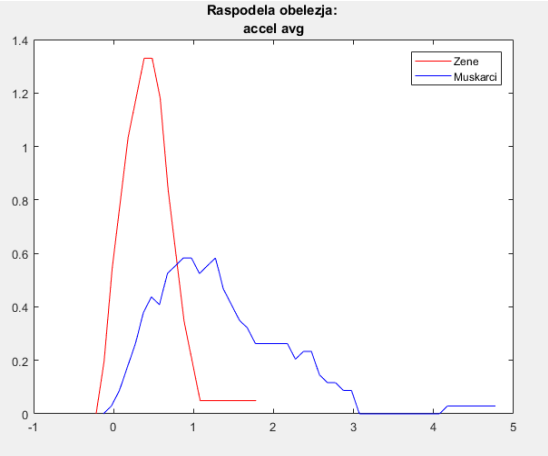
U ovom radu biće predstavljeni različiti modeli za predikciju pola korisnika na osnovu praćenja njegove aktivnosti na mobilnom telefonu.

II. BAZA PODATAKA

Podaci iz date baze su deo Nokia MDC-Mobile Data Challenge-a. Baza se sastoji od 78 korisnika koji su anonimni i to 49 muškaraca i 29 žena. Svaki od njih je opisan sa 40 obeležja koja su numerička. Za potrebe drugog dela zadatka je redukovan skup obeležja na 16. Selekcija je izvršena RelieF algoritmom. Obeležja predstavljaju zabeleženo vreme kada se na telefonu pokreću aplikacije, obavljaju i primaju pozivi ili poruke, menja GSM ćelija, približna rastojanja koja su pređena kao i merenja sa senzora ubrzanja i mnoga druga. Ova baza je potpuna i nema nedostajućih podataka.

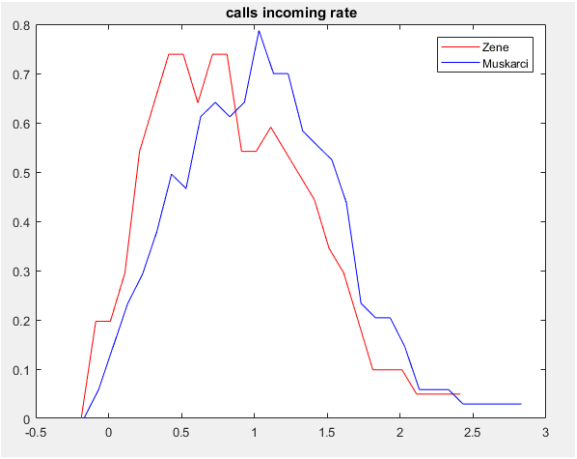
III. ANALIZA PODATAKA

U ovom delu će biti analizirane raspodele pojedinačnih obeležja. Estimacija je vršena Parzenovim prozorom, gde je širina prozora 0.7. Iskorišćena je gotova funkcija *estimacija_1D_parzen.m*. Na slikama ispod prikazane su raspodele najilustrativnijih obeležja.



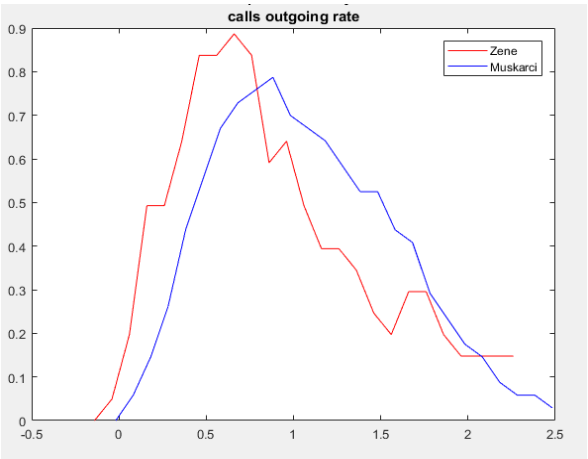
Sl. 1. Raspodele obeležja *accel avg*.

Obeležje *accel avg* predstavlja prosečnu vrednost merenja sa senzora ubrzanja. Na slici 1 se može primetiti da raspodele ovog obeležja za muškarce i žene imaju različite srednje vrednosti. Opsezi vrednosti ovog obeležja za žene i muškarce se ne poklapaju u velikoj meri, te se stoga može zaključiti da je ovo obeležje pogodno za klasifikaciju polova.



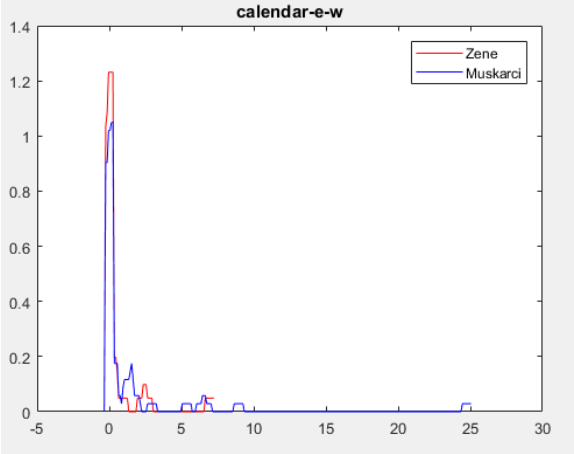
Sl. 2. Raspodele obeležja *calls incoming rate.*

Obeležje *calls incoming rate* predstavlja odnos dolazećih poziva prema ukupnom broju svih konekcija u calllog-u. Na osnovu slike 2 može se primetiti da žene i muškarci imaju sličan broj dolazećih poziva pa se pomoću ovog obeležja može izvršiti lošija klasifikacija nego na primer, kod *accel avg* obeležja.



Sl. 3. Raspodele obeležja *calls outgoing rate*.

Obeležje *calls outgoing rate* predstavlja odnos odlazećih poziva prema ukupnom broju svih konekcija u calllog-u. Sa slike 3 se zaključuje da su muškarci više skloni ka govornim pozivima nego žene, ali i dalje nedovoljno za potrebnu klasifikaciju.



Sl. 4. Raspodele obeležja *calendar-e-w*.

Obeležje *calendar-e-w* predstavlja prosečan broj upisanih događaja u kalendar vikendom. Sa slike 4. se vidi da muškarci i žene podjednako slabo koriste kalendar vikendom i ovo obeležje ne može doprineti klasifikaciji.

IV. KLASIFIKATORI

Za potrebe obuke klasifikatora, iz celog skupa podataka je izdvojen trening skup na kome se vrši obučavanje klasifikatora i test skup koji služi za procenu kvaliteta dobijenog modela klasifikatora. Podela je izvršena ugrađenom Matlab-ovom funkcijom *cvpartition* koja na pametan način uzima podjednak procenat uzoraka obe klase za oba skupa. U ovom radu se za test skup iz svake klase izdvaja po 15% uzoraka.

U nastavku će biti predstavljeni različiti algoritmi korišćeni za predikciju pola, pri čemu su za evaluaciju i odabir parametara modela korišćene *5-fold* i *leave-one-out* krosvalidacione metode.

A. *kNN klasifikator*

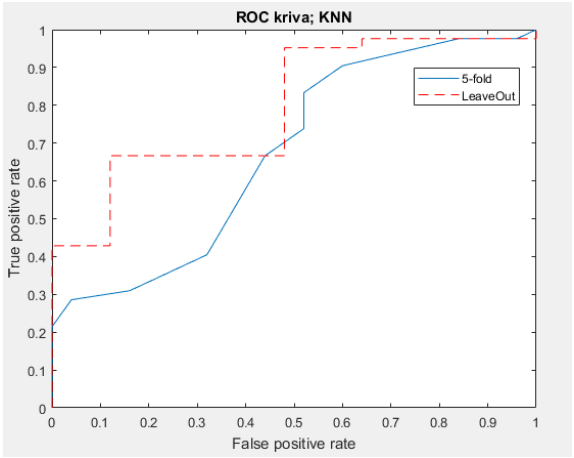
Metod *k* najbližih suseda (kNN) kao pravilo klasifikacije predstavlja veoma intuitivan metod koji klasifikuje neobeležene primerke na osnovu njihove sličnosti sa primercima iz trening skupa, odnosno, odluka o pripadnosti klasi donosi se na osnovu preglasavanja među *k* uzoraka koji su najbliži test uzorku.

Variranjem različitih parametara u funkciji *fitcknn* isprobani su različiti modeli klasifikacije. Kao rastojanje je korišćeno euklidsko. Takođe su uključeni parametri *BreakTies* i *Standardize*. Na osnovu *5-fold* krosvalidacije i mere tačnosti najbolje se pokazao model za k=22, dok je za *leave-one-out* krosvalidaciju to k=11.

Rezultati koji su dobijeni ovim modelima na neredukovanom skupu obeležja su prikazani ispod.

	k	Tačnost	Preciznost	Odziv	F-mera
5-fold	22	0.727	1	0.250	0.400
LeaveOut	11	0.727	1	0.250	0.400

TABELA 1: MERE KNN MODELA ZA CEO SKUP OBELEŽJA.

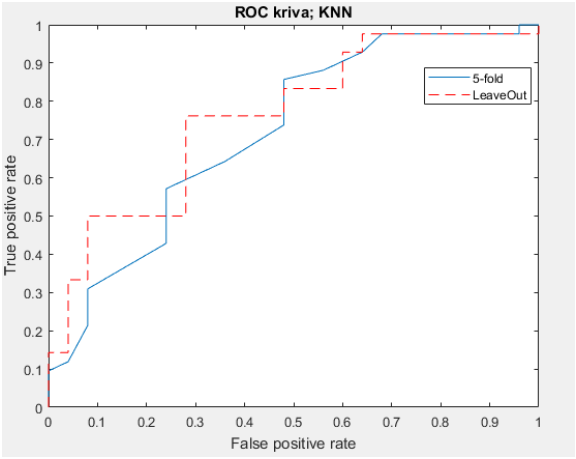


Sl. 5. ROC krive kNN za ceo skup obeležja.

Za iste parametre, na redukovonom skupu, dobijeni su sledeći rezultati:

	k	Tačnost	Preciznost	Odziv	F-mera
5-fold	20	0.727	1	0.250	0.400
LeaveOut	19	0.727	1	0.250	0.400

TABELA 2: MERE KNN MODELA ZA REDUKOVAN SKUP OBELEŽJA.



Sl. 6. ROC krive kNN za redukovan skup obeležja.

Na osnovu prikazanih rezultata, može se primetiti da su sva četiri modela jednakih performansi gde se samo broj suseda menja. Iako su mere tačnosti i preciznosti zadovoljavajuće, odziv ovih modela je neočekivano male vrednosti, pa se može reći da je ovaj model loš klasifikator.

B. *SVM klasifikator*

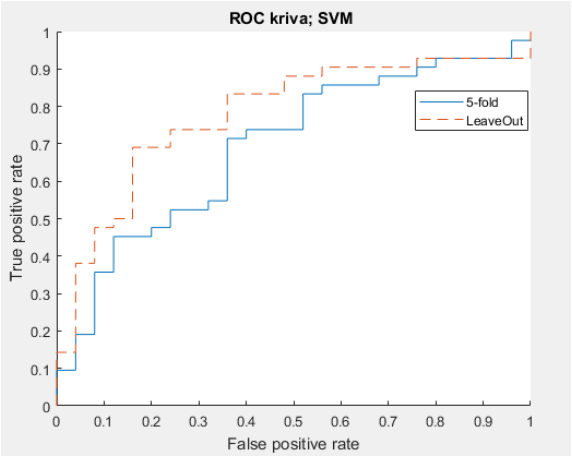
Model vektora nosača (SVM) deli prostor obeležja pomoću jedne hiper-ravni, stoga se koristi za razdvajanje dve klase, kao u razmatranom slučaju. Vektori nosači su uzorci koji se nalaze na ivicama margine.

Obuka se vrši pozivom ugrađene Matlab-ove funkcije *fitcsvm* sa parametrima *KernelFunction* za koji je izabran *polynomial* reda 1 i *Standardize*.

U nastavku su dati rezultati dobijeni za dva skupa obeležja.

	Tačnost	Preciznost	Odziv	F-mera
5-fold	0.727	0.670	0.500	0.570
LeaveOut	0.727	0.670	0.500	0.570

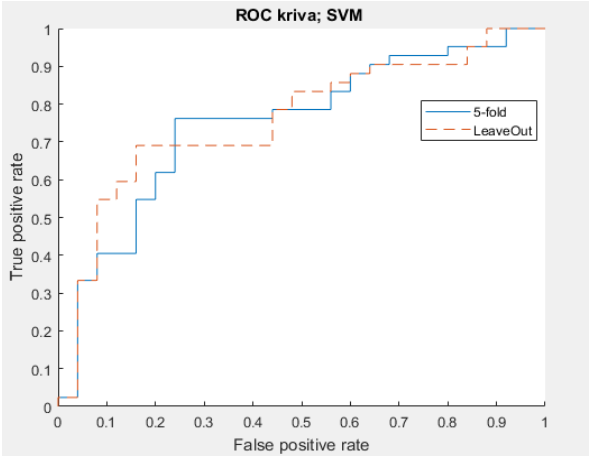
TABELA 3: MERE SVM MODELA ZA CEO SKUP OBELEŽJA.



Sl. 7. ROC krive SVM za ceo skup obeležja.

	Tačnost	Preciznost	Odziv	F-mera
5-fold	0.818	0.750	0.750	0.750
LeaveOut	0.818	0.750	0.750	0.750

TABELA 4: MERE SVM MODELA ZA REDUKOVAN SKUPOBELEŽJA.



Sl. 8. ROC krive SVM za redukovan skup obeležja.

Sa ROC kriva i iz priloženih tabela možemo izvesti zaključak da je bolji klasifikator dobijen nad podacima koji su opisani sa redukovanim skupom obeležja.

C. *Random forest klasifikator*

Random forest algoritam predstavlja jednu od metoda ansambalskog učenja. Ideja je da se obuči mnoštvo stabala odluke i izvrši klasifikacija donošenjem krajnje odluke glasanjem. Iako dimenzionalnost, nedostajuće vrednosti i outlier-i ne predstavljaju problem ovog algoritma javljaju se poteškoće pri interpretaciji i mogućnostima kontrole algoritma.

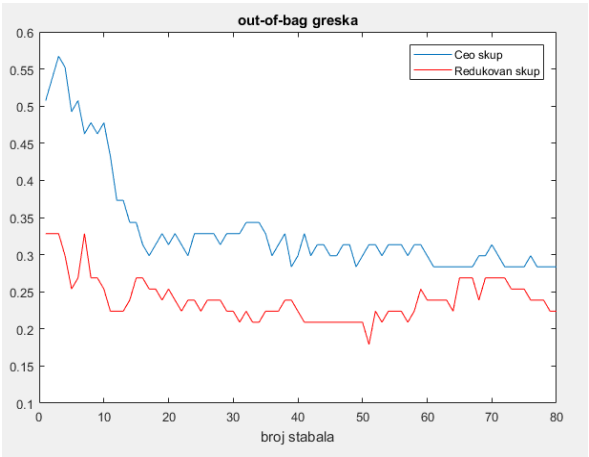
Za potrebe realizacije modela ovog algoritma korišćena je *TreeBagger* funkcija kojoj se kao parametar prosleđuje broj stabala odluke.

Sledi prikaz dobijenih rezultata za oba skupa obeležja.

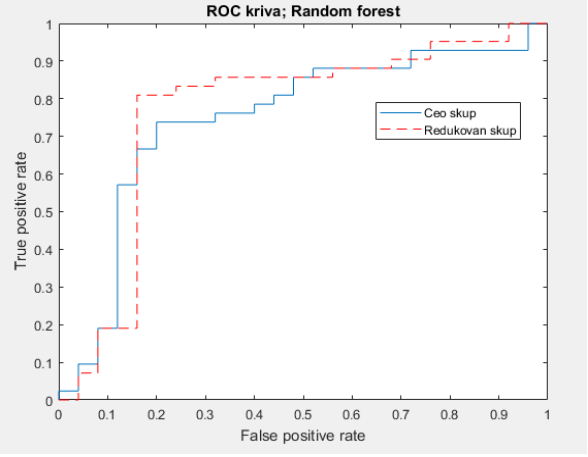
Br. obeležja	Tačnost	Preciznost	Odziv	F-mera
40	0.818	1	0.500	0.670
16	0.909	1	0.750	0.857

TABELA 5: MERE RANDOM FOREST ALGORITMA ZA OBA SKUPA OBELEŽJA.

S obzirom da je dati skup podataka veoma mali, tačnost postignuta ovim algoritmom je zadivljujuća.



Sl. 9. Out-of-bag greška *Random forest* za oba skupa obeležja.



Sl. 10. ROC krive *Random forest* algoritma za oba skupa obeležja.

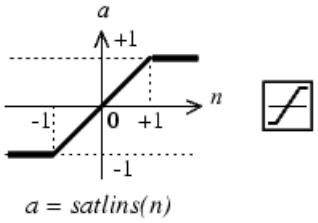
Random forest algoritam koristi *bootstraping* i na osnovu toga dolazi se do takozvane *out-of-bag* greške koje su prethodno prikazane. Sa slike 9. se vidi, da posle svega 10 stabala dolazi do naglog pada greške na vrednost oko 0.3, dok je u samom startu ona bila oko 0.5, i to sve kada je u pitanju ceo skup obeležja. Time se zaključuje da veći broj stabala uzrokuje manju grešku, odnosno bolju klasifikaciju. Nakon 80 stabala, greška postaje konstantna, te je u redu stati sa povećavanjem broja stabala. Za razliku od pomenute, *out-of-bag* greška na redukovanom skupu ima znatno manje vrednosti na celom opsegu i može se stati sa povećavanjem broja stabala nakon 40.

Na osnovu svega navedenog, zaključuje se da je bolje koristiti redukovan skup podataka za klasifikaciju pomoću *Random forest* algoritma, kao što se može videti iz prikazanih rezultata.

D. *Neuralne mreže*

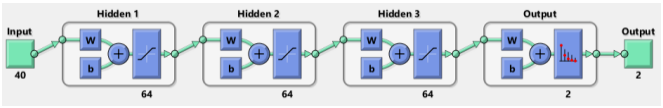
Neuralne mreže predstavljaju novu generaciju sistema za informaciono procesiranje koje pokazuju osobinu učenja, memorisanja i generalizacije na osnovu podataka kojima se obučavaju. Arhitektura neuralne mreže se sastoji od ulaznog, skrivenog i izlaznog sloja. Svaki sloj se sastoji od određenog broja neurona, a svaki neuron okarakterisan je težinama i pristrasnošću. Izlazi prethodnog sloja se sumiraju i rezultat prolazi kroz aktivacionu funkciju neurona.

Za generisanje neuralne mreže korišćen je Matlab-ov alat *Neural Network Toolbox*, čiji je kod prilagođen zahtevima ovog zadatka. Algoritam zasnovan na propagaciji unazad koji je korišćen za obuku mreže, je skaliran konjugovani gradijent propagacije unazad. Kao aktivaciona funkcija korišćena je *satlins* (Simetrična saturaciona linearna funkcija) koja je prikazana na slici 11.



Sl. 11. *Satlins* funkcija.

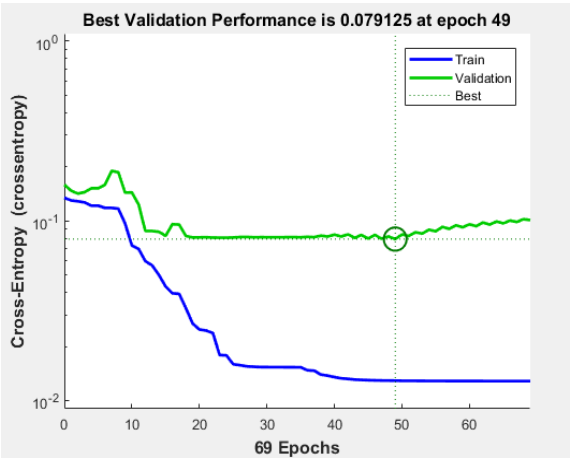
Generisan je model od tri skrivena sloja sa po 64 neurona.



Sl. 12. Model neuralne mreže za 40 obeležja.

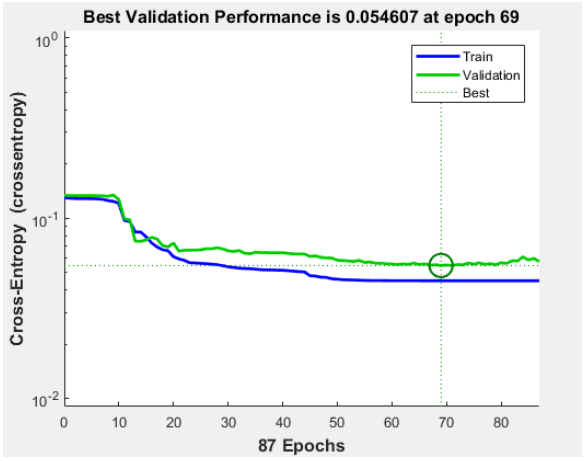
Model je primenjen na oba skupa obeležja i u nastavku su prikazani dobijeni rezultati.

Grafici performanse, prikazani na slikama 13. i 14. pokazuju promenu funkcije cene kroz epohe nad svakim skupom. Epoha predstavlja jedan prolazak celokupnog skupa za obuku kroz mrežu. Mreža treba da se u svakoj novoj epohi bolje obuči. Ukoliko algoritam dostigne najbolji rezultat i u narednih 20 epoha ne premaši postignuti rezultat na validacionom skupu, obuka staje.



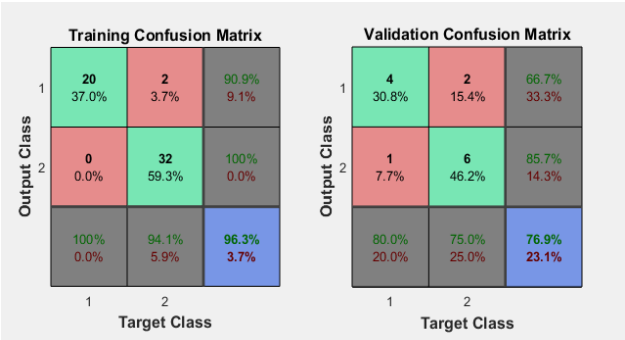
Sl. 13. Grafik performanse za 40 obeležja.

Sa ove slike se još vidi i da se mreža nije mnogo obučila, kao i da je došlo do natprilagođenja algoritma na trening podacima. Najbolji rezultat je postignut u 49. epohi.



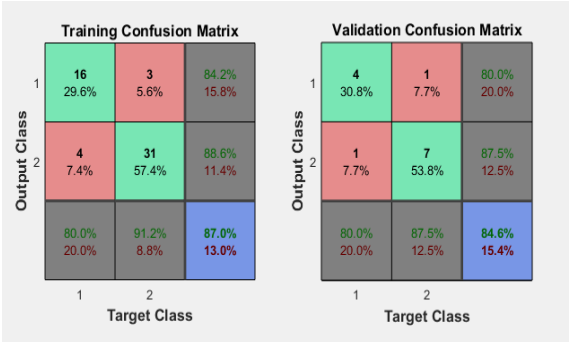
Sl. 14. Grafik performanse za 16 obeležja.

Sa slike 14. se vidi da greška na trening skupu opada skoro isto kao na validacionom skupu, kao i da se mreža malo bolje obučila nego u prethodnom slučaju. Najbolji rezultat je postignut u 69. epohi.



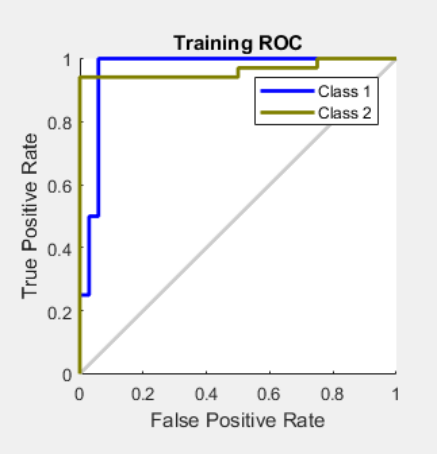
Sl. 15. Matrice konfuzije za trening i validacioni skup (40 obeležja).

Iz prikazanih matrica konfuzija primećujemo da je mreža pogrešila na veoma malo trening uzoraka, kao i da bolje pogađa muškarce.

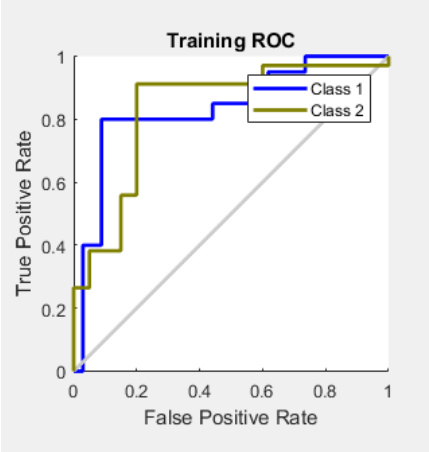


Sl. 16. Matrice konfuzije za trening i validacioni skup (16 obeležja).

Na selektovanom skupu obeležja, mreža je pri treningu više grešila na oba pola u odnosu na mrežu koja je trenirana za ceo skup obeležja.



Sl. 17. ROC krive za trening skup (40 obeležja).



Sl. 18. ROC krive za trening skup (16 obeležja).

Prikazane slike ilustruju performanse modela dobijene na trening i validacionom skupu, dok se krajnje mere dobijaju na test skupu, koji model do tad nije obradio. Rezultati su prikazani u tabeli 6.

Br. obeležja	Tačnost	Preciznost	Odziv	F-mera
40	0.800	0.800	0.600	0.685
16	0.800	0.800	0.600	0.685

TABELA 6: MERE NEURALNE MREŽE ZA OBA SKUPA OBELEŽJA.

Iz priložene tabele se može primetiti da se ovaj model podjednako dobro snašao sa podacima, kako celog, tako i redukovanog skupa obeležja.

V. ZAKLJUČAK

U ovom radu predstavljene su najčešće korišćene metode klasifikacije. Na raspolaganju je mali skup podataka koji sadrži 62,8% uzoraka muškaraca i 37,2% žena, te se iz tog razloga pokazalo da svi klasifikatori bolje odlučuju u korist muškaraca.

Poređenjem performansi dobijenih klasifikatora, kao najbolji model za problem predikcije pola korisnika, pokazao se model algoritma *Random forest*. Tačnost modela za oba skupa obeležja je najveća u poređenju sa ostalim modelima. Ovaj algoritam postiže bolju tačnost na redukovanom skupu, što je i očekivano, s obzirom da se u njemu nalaze diskriminatornija obeležja.

Klasifikator dobijen metodom najbližih suseda na oba skupa obeležja postiže identične rezultate, iako broj suseda nije isti. Kako je broj suseda oko 20, a broj uzoraka za obučavanje modela 54, klasifikacija ovim modelom daje dobre rezultate s obzirom da se oslanja na okolinu u kojoj je znatno veći broj muškaraca.

SVM klasifikator je rezultovao boljim performansama na redukovanom skupu obeležja, kao i u odnosu na kNN.

Neuralnoj mreži dimenzionalnost podataka nije problem pa je postigla jednake performanse na oba skupa obeležja.

S obzirom da je Random forest algoritam težak za interpretaciju, teško je ući u neku dublju analizu zašto je dao najbolji rezultat.