PREDIKCIJA SAOBRAĆAJA U LOKALNOJ RAČUNARSKOJ MREŽI PRIMENOM NADGLEDANOG MAŠINSKOG UČENJA

Slađana Janković, Katarina Kukić, Ana Uzelac, Vladislav Maraš
Univerzitet u Beogradu - Saobraćajni fakultet
s.jankovic@sf.bg.ac.rs; k.mijailovic@sf.bg.ac.rs;
ana.uzelac@sf.bg.ac.rs; v.maras@sf.bg.ac.rs

Rezime: *Mašinsko učenje može se definisati kao generalizacija znanja na osnovu prethodnog iskustva, odnosno podataka o entitetima koji su predmet učenja. Danas, u Big Data eri, mašinsko učenje koristi se kao jedna od vodećih tehnika prediktivne analitike. U ovom radu biće opisana metoda predikcije bazirana na izgradnji i primeni modela nadgledanog mašinskog učenja, kao i implementacija svih faza procesa mašinskog učenja u softverskom alatu Weka (Waikato Environment for Knowledge Analysis). Ovaj data mining softver predstavlja kolekciju algoritama mašinskog učenja koji se koriste u poslovima otkrivanja zakonitosti u podacima. Implementacija procesa mašinskog učenja, korišćenjem softvera Weka, u radu je demonstrirana na studiji slučaja predikcije saobraćaja generisanog od strane različitih korisničkih grupa u okviru posmatrane lokalne računarske mreže.*

Ključne reči: *mašinsko učenje, Big Data, mrežni saobraćaj, Weka*

1. Uvod

Poslovne aktivnosti bazirane na upotrebi informacionih tehnologija karakteriše permanentna dinamičnost, uslovljena stalnim uvođenjem novih poslovnih aktivnosti, kao i unapređenjem postojećih poslovnih procesa. Uticaj poslovnih procesa na iskorišćenje mrežnih kapaciteta lokalne računarske mreže predstavlja važan problem, posmatrano sa aspekta inženjerskih aktivnosti vezanih za alokaciju i proširenje mrežnih kapaciteta u procesima planiranja. Vremenska alokacija poslovnih procesa koji su u direktnoj vezi sa aktivnostima na jednoj lokalnoj mreži predstavlja jedan pristup optimalnom iskorišćenju postojećih mrežnih resursa. Da bi se vremenska alokacija pravilno realizovala, potrebno je izvršiti predikciju iskorišćenja mrežnih resursa na osnovu postojećih podataka dobijenih postupkom monitoringa mrežne infrastrukture. Na taj način mogu se izbeći zagušenja na onim mrežnim linkovima, koji su kritični za efikasnu realizaciju posmatranih poslovnih procesa. Na isti način može se planirati i uvođenje novih poslovnih aktivnosti, koje se mogu sprovoditi u onim vremenskim intervalima za koje je prognoziran niži nivo mrežnih aktivnosti.

Poslovni procesi su, u okviru jedne mreže, po pravilu razdvojeni u skladu sa organizacionom strukturom jedne institucije. Sve službe posmatrane institucije mogu koristiti zajedničke mrežne resurse, poput istog mrežnog linka za vezu ka internetu,

zajedničkih *public* servera za *backup* podataka i sl. Bez obzira što se najveći deo poslovnih aktivnosti jedne službe odvija u okviru pripadajućeg dela mrežne infrastrukture, dok se samo jedan deo aktivnosti reflektuje na zajedničke resurse, u procesu planiranja vremenske alokacije bitno je uočiti kumulativni efekat svih poslovnih procesa na stepen iskorišćenja zajedničkih mrežnih resursa i uključiti ga u proces predikcije njihovog iskorišćenja.

S obzirom na to da monitoring mrežne infrastrukture generiše velike količine podataka koje brzo rastu, za predikciju nivoa mrežnog saobraćaja nametnule su se tehnike *Big Data* analitike. Važnu klasu tehnika *Big Data* analitike čini prediktivna analitika bazirana na nadgledanom mašinskom učenju. Nadgledano učenje karakteriše se time da su uz vrednosti ulaza date i vrednosti izlaza koje im odgovaraju, a ovakvi skupovi podataka nazivaju se označenim skupovima podataka (eng. *labeled dataset*). Potrebno je ustanoviti odnos koji važi između ulaza i izlaza. Na osnovu ovog odnosa najčešće se za neke buduće ulaze vrši predviđanje izlaza [1]. Ulaz i izlaz najčešće se predstavljaju u vektorskom obliku i označavaju sa x i y, pri čemu je x tipično vektor vrednosti nekih nezavisnih promenljivih koje se nazivaju atributima (eng. *attributes*, *features*), dok je y tipično jedna zavisna promenljiva, koja se još naziva i ciljnom promenljivom (eng. *target variable*). Modeli mašinskog učenja obučavani na *Big Data* skupovima podataka uspešno se mogu primeniti za predviđanje ponašanja korisnika i procesa u raznim oblastima. Naravno, postoje i zamke pri korišćenju ovih tehnika. Ova vrsta projektovanja budućeg ponašanja zasnovana je na istorijskim obrascima, zbog čega postoji rizik da se zanemare sadašnji i budući trendovi.

Cilj ovog istraživanja je da ispita mogućnosti predikcije nivoa saobraćaja u lokalnoj računarskoj mreži korišćenjem tehnike nadgledanog mašinskog učenja. Modeli mašinskog učenja razvijani su i primenjivani za predikciju u *data mining* softverskom alatu *Weka* [2].

Druga sekcija rada sadrži kratak uvod u primenu *Big Data* analitike, a naročito mašinskog učenja, u oblasti telekomunikacija. U trećoj sekciji prikazana je primenjena metodologija predikcije, a u četvrtoj rezultati njene primene na *dataset*-u koji opisuje saobraćaj u posmatranoj lokalnoj računarskoj mreži. Poslednja sekcija rada sadrži zaključke o mogućnostima i ograničenjima primene izabrane tehnike za predikciju nivoa saobraćaja u lokalnoj računarskoj mreži.

2. Big Data analitika i mašinsko učenje u telekomunikacijama

U ovom radu biće primenjena jedna od tehnika *Big Data* analitike u telekomunikacijama, stoga se izdvajaju neki od značajnijih radova koji se bave tom tematikom. U studiji [3] analizirani su uticaji *Big Data* tehnologija na dinamiku tržišta, sa detaljnim osvrtom na telekomunikacione kompanije u Švedskoj. U radu [4] prikazana je primena *Big Data* tehnologija kod telekomunikacionih operatera, sa aspekta karakteristika *Big Data* skupova podataka u kominikacijama, arhitekture *Big Data* platformi kao i razvoja aplikacija. Više o opravdanosti i rizicima investiranja telekomunikacionih kompanija u *Big Data* tehnologije, može sa naći u radu [5]. O izazovima sa kojima se suočavamo pri radu sa *Big Data* skupovima podataka može se pročitati u radu [6].

Mašinsko učenje u *Big Data* eri predstavlja jedan od nosećih stubova veštačke inteligencije. Kao početni trenutak u razvoju veštačke inteligencije obično se uzima 1956. godina i čuvena konferencija u *Darthmouth College*-u. Za nešto više od šezdeset godina razvoja, od čega posebno dominira razvoj u poslednje 2 decenije, veštačka inteligencija

prožela je sve aspekte savremenog života. Kratak pregled razvoja veštačke inteligencije prikazan je u radu [7].

Sa ciljem povezivanja univerziteta i istraživačkih ustanova sa problemima iz telekomunikacione prakse pokrenut je i časopis *International Telecomunication Union Jornal, ICT discoveries*. Prvo specijalno izdanje ovog časopisa posvećeno je temi koju i mi iz jednog ugla obrađujemo u ovom radu. Zato bismo zainteresovanog čitaoca uputili na ovo izdanje [8] gde se u nekoliko preglednih radova može upoznati sa raznim aspektima primene veštačke inteligencije u telekomunikacijama.

Više o različitim mogućnostima statističkog učenja i na njemu zasnovane predikcije, može se videti u [9]. U *Cisco*-voj studiji [10] navodi se da će globalni IP saobraćaj (protok podataka putem interneta) do 2021. godine utrostručiti vrednost u odnosu na 2016. godinu. Pri tom se predviđa da će se značajno povećati broj uređaja povezanih na internet. Pre svega se misli na porast broja bežičnih uređaja i rastući trend *Internet of Things* koncepta. Prognoze su da će broj uređaja povezanih na internet do 2021. trostruko nadmašiti brojnost ljudske populacije u tom trenutku [10], što nam ukazuje da će mašinsko učenje tek dobiti svoje mesto u sferi telekomunikacija.

3. Metodologija

Proces mašinskog učenja sastoji se od faza: priprema podataka, izgradnja modela, validacija modela, testiranje modela i primena modela. Mašinsko učenje je iterativan proces u kojem se sve gore nabrojane faze ponavljaju onoliko puta koliko je potrebno. Ponavljanje ovih faza završava se kada se iscrpu sve kombinacije atributa, svi raspoloživi algoritmi i vrednosti parametara algoritama, ili kada se dođe do modela zadovoljavajućih performansi. Jednom, kada testiranje modela pokaže da model ima zadovoljavajuće performanse, može se započeti korišćenje modela u predikciji izabrane varijable.

Priprema podataka sastoji se od: čišćenja sirovih podataka od nepotpunih zapisa ili zapisa sa neispravnim vrednostima, konvertovanja podataka u odgovarajući format i sl.

Izgradnja svakog od modela mašinskog učenja sastojala se od sledećih faza:

1. Definisanje cilja modela, u skladu sa ciljevima prediktivne analitike;
2. Izbor ciljne promenljive, tj. atributa iz skupa podataka čiju vrednost želimo da predvidimo primenom modela mašinskog učenja;
3. Izbor algoritma nadgledanog mašinskog učenja, u skladu sa prirodom ciljne promenljive i atributa;
4. Izbor relevantnih atributa skupa podataka;
5. Priprema skupova podataka za učenje i za testiranje modela, prema zahtevima izabranog algoritma;
6. Podešavanje modela, tj. vrednosti hiperparametara specifičnih za svaku vrstu algoritma mašinskog učenja;
7. Učenje (treniranje) modela – primena izabranog algoritma mašinskog učenja na skup podataka za učenje, u cilju dobijanja hiperparametara modela.

Za potrebe formiranja skupova podataka za treniranje i testiranje modela iskorišćene su log datoteke programskog paketa za monitoring mreže *The Multi Router Traffic Grapher* (MRTG) [11], koje su zabeležile aktivnosti 14 korisničkih grupa, kao i aktivnost na zajedničkom linku za pristup internetu. Snimanje podataka za svaku pojedinačnu grupu realizovano je primenom monitoringa upravljivog *Layer 3* (L3) sviča, koji je zadužen za segmentaciju lokalne računarske mreže (LAN) na 15 virtuelnih

podmreža (*Virtual Local Area Network* – VLAN). VLAN segmentacija podrazumeva postojanje jednog (*default*) VLAN-a, koji nije uključen u podelu LAN mreže po organizacionoj strukturi, kao i 14 ostalih VLAN-ova koji su formirani u skladu sa organizacionom strukturom posmatrane institucije. Za snimanje nivoa saobraćaja predefinisani interval je 300 sekundi, odnosno 5 minuta. U svakom intervalu se beleže podaci o prosečnim i maksimalnim vrednostima realizovanog saobraćaja u oba smera (odlazni i dolazni). Ovi podaci se beleže u jednoj liniji MRTG log datoteke, koja se sastoji od ukupno 5 podataka: vremenski žig, prosečna vrednost odlaznog saobraćaja, prosečna vrednost dolaznog saobraćaja, maksimalna vrednost odlaznog saobraćaja i maksimalna vrednost dolaznog saobraćaja. Ove vrednosti mogu biti zabeležene u jedinicama bit/s ili byte/s, zavisno od MRTG konfiguracije.

S obzirom na strukturu raspoloživog skupa podataka o dolaznom i odlaznom saobraćaju u posmatranoj lokalnoj računarskoj mreži, definisani su sledeći ciljevi budućih modela: predikcija prosečnog nivoa dolaznog saobraćaja u petominutnim intervalima, predikcija prosečnog nivoa odlaznog saobraćaja u petominutnim intervalima. Kao ciljne promenljive, izabrani su sledeći atributi: Odlazni saobraćaj – petominutni prosek i Dolazni saobraćaj – petominutni prosek.

S obzirom na to da su ciljne promenljive našeg skupa podataka neprekidne, izgrađeni su modeli mašinskog učenja bazirani na najpopularnijim regresionim algoritmima: linearna regresija (eng. *Linear Regression*), *k* najbližih suseda (eng. *K-Nearest Neighbors*), drvo odlučivanja (eng. *Decision Tree*), metoda potpornih vektora za regresiju (eng. *Support Vector Machines for Regression, SMOreg*), neuronska mreža (eng. *Neural Network*).

Pored treniranja i testiranja modela, urađena je i validacija modela, kako bi se izabrao najbolji model između više kandidata, odredila optimalna konfiguracija parametara modela i izbegli problemi prevelikog podudaranja (eng. *overfitting*) i nedovoljnog podudaranja (eng. *underfitting*). Preveliko podudaranje odnosi se na situaciju u kojoj model savršeno nauči da vrši predikciju za instance iz trening skupa, ali ima veoma slabu sposobnost predikcije za instance koje se i malo razlikuju od naučenih. Nedovoljno podudaranje odnosi se na slučaj kada model ne uspeva da aproksimira podatke za trening, tako da ima slabe performance čak i na trening skupu podataka.

Za validaciju modela korišćen je pristup poznat pod nazivom unakrsno vrednovanje (eng. *cross-validation*). Ovaj pristup za ocenu performansi modela, koristi samo podatke za trening, a sastoji se od sledećih faza:

1. Raspoloživi skup podataka za treniranje modela deli se na K jednakih delova-podskupova (eng. *folds*). Najčešće se deli na 10 podskupova (eng. *10-fold cross validation*).
2. Model se trenira na K-1 podskupova podataka (npr. na prvih K-1 podskupova).
3. Model se ocenjuje na jedinom preostalom (K-tom) podskupu podataka.
4. Koraci 2 i 3 ponavljaju se K puta. U svakoj iteraciji uzima se jedan deo podataka za potrebe validacije modela, dok se ostatak (K-1 delova) koristi za učenje. Bira se uvek različit podskup koji će se koristiti za validaciju modela.
5. Performanse modela izračunavaju se kao aritmetičke sredine performansi dobijenih u K iteracija.

Nekoliko različitih mera može se koristiti za ocenjivanje uspešnosti numeričke predikcije [12]. Projektovane vrednosti ciljne promenljive, dobijene za skup instanci za

validaciju modela su: $p_1, p_2, \ldots, p_n$; dok su stvarne vrednosti ciljne promenljive: $a_1, a_2, \ldots, a_n$.

Srednja kvadratna greška (1) je glavna i najčešće korišćena mera.

$$Srednja\ kvadratna\ greška = \frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n} \quad (1)$$

Srednja apsolutna greška (2) je prosečna veličina individualnih grešaka bez uzimanja u obzir njihovog znaka. Srednja kvadratna greška ima tendenciju da preuveličava efekat izuzetaka - slučajeva kod kojih je greška predviđanja veća nego kod drugih, dok apsolutna greška nema ovaj efekat: sve veličine greške se tretiraju ravnomerno prema njihovoj veličini.

$$Srednja\ apsolutna\ greška = \frac{|p_1-a_1|+\cdots+|p_n-a_n|}{n} \quad (2)$$

Kvadratni koren srednje kvadratne greške izračunava se na osnovu jednačine (3):

$$Kvadratni\ koren\ srednje\ kvadratne\ greške = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}} \quad (3)$$

Relativna kvadratna greška (4) računa se u odnosu na ono što bi bilo da se koristi jednostavni prediktor. Jednostavni prediktor je samo prosek stvarnih vrednosti iz podataka za trening, označen sa $\bar{a}$. Dakle, relativna kvadratna greška uzima ukupnu kvadratnu grešku, i normalizuje je tako što je deli ukupnom kvadratnom greškom podrazumevanog prediktora.

$$Relativna\ kvadratna\ greška = \frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2} \quad (4)$$

Kvadratni koren relativne kvadratne greške izračunava se na osnovu jednačine (5):

$$Kvadratni\ koren\ relativne\ kvadratne\ greške = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2}} \quad (5)$$

Relativna apsolutna greška (6) je ukupna apsolutna greška, sa istom vrstom normalizacije.

$$Relativna\ apsolutna\ greška = \frac{|p_1-a_1|+\cdots+|p_n-a_n|}{|a_1-\bar{a}|+\cdots+|a_n-\bar{a}|} \quad (6)$$

Poslednja mera tačnosti predikcije je koeficijent korelacije (7), koji meri statističku korelaciju između vrednosti a i p. Koeficijent korelacije uzima vrednosti od 1 za rezultate koji su u potpunoj korelaciji, preko 0 kada nema korelacije, do -1 kada su rezultati u savršeno negativnoj korelaciji.

$$Koeficijent\ korelacije = \frac{S_{PA}}{\sqrt{S_P S_A}}, \quad (7)$$

gde je:

$$S_{PA} = \frac{\sum_{i=1}^{n}(p_i-\bar{p})(a_i-\bar{a})}{n-1}, \qquad S_P = \frac{\sum_{i=1}^{n}(p_i-\bar{p})^2}{n-1}, \qquad S_A = \frac{\sum_{i=1}^{n}(a_i-\bar{a})^2}{n-1} \quad (8)$$

U većini praktičnih situacija prediktivni model koji je najbolji prema jednoj meri, istovremeno je najbolji i prema svim ostalim merama greške.

Da bi predvideli performanse modela na novim podacima, potrebno je proceniti mere njihovih performansi na skupu podataka koji nije igrao nikakvu ulogu u formiranju modela. Ovaj nezavisni skup podataka naziva se skup podataka za testiranje.

Sledeća faza je poređenje performansi modela dobijenih na skupu podataka za testiranje sa performansama dobijenim na skupu podataka za trening. Ono omogućava da se izbegne problem prevelikog podudaranja. Ako model radi veoma dobro na podacima za trening, ali slabo na podacima za testiranje, onda postoji problem prevelikog podudaranja.

Da bi se predvidele vrednosti izabranih ciljnih promenljivih u budućnosti, potrebno je pripremiti odgovarajući skup podataka i na njega primeniti model mašinskog učenja koji je izabran kao najbolji.

4. Rezultati i analiza rezultata

U Tabelama 1 i 2 prikazane su performanse šest najboljih modela predikcije prosečnog nivoa odlaznog saobraćaja merene na skupu podataka za trening, odnosno testiranje, respektivno.

Tabela 1. Performanse šest najboljih modela predikcije prosečnog nivoa odlaznog saobraćaja merene na skupu podataka za trening

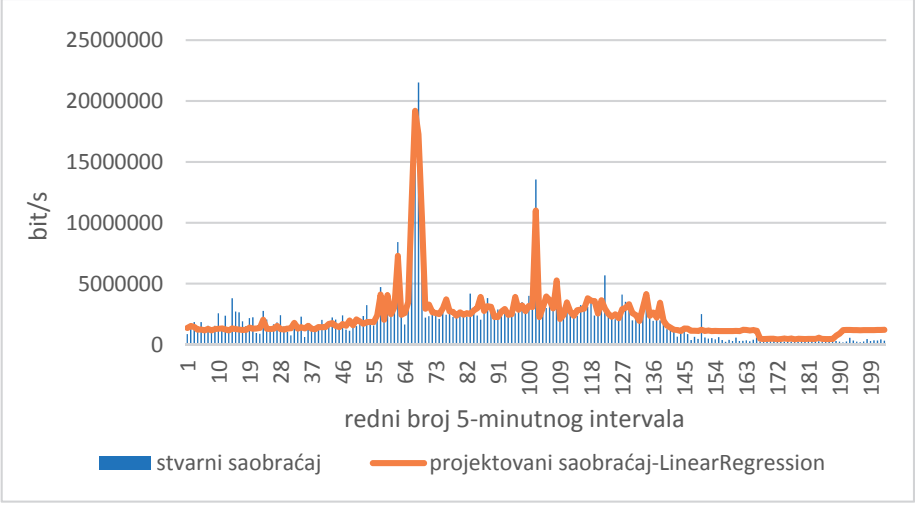
Algoritam	Koeficijent korelacije	Srednja apsolutna greška	Kvadratni koren srednje kvadratne greške	Relativna apsolutna greška (%)	Kvadratni koren relativne kvadratne greške (%)
Linear Regression	0.8746	593043.27	891542.44	45.0225	48.3222
Multilayer Perceptron	0.8242	843481.49	1176387.31	64.0353	63.7610
REPTree	0.8646	546339.21	924474.65	41.4769	50.1071
Random Forest	0.8868	475532.81	855691.52	36.1014	46.3790
SMOreg	0.8650	598262.23	938358.50	45.4188	50.8596
IBk	0.8051	661880.7	1153057.66	50.2485	62.4965

Kao što se može videti u Tabeli 2, kod modela baziranih na algoritmima *REPTree*, *Random Forest* i *IBk* zabeležen je problem prevelikog podudaranja. Najbolji model tražen je među preostalim modelima – modelima baziranim na regresionim algoritmima: *Linear Regression*, *Multilayer Perceptron* i *SMOreg*. Najbolje performanse, prema svim metrikama, pokazao je model baziran na algoritmu *Linear Regression* (Tabela 2), tako da je predikcija zavisne varijable izvršena korišćenjem ovog modela mašinskog učenja.

Tabela 2. Performanse šest najboljih modela predikcije prosečnog nivoa odlaznog saobraćaja merene na skupu podataka za testiranje

Algoritam	Koeficijent korelacije	Srednja apsolutna greška	Kvadratni koren srednje kvadratne greške	Relativna apsolutna greška (%)	Kvadratni koren relativne kvadratne greške (%)
Linear Regression	0.9528	544977.34	773545.82	41.3489	31.1008
Multilayer Perceptron	0.9027	754544.87	1349199.31	57.2494	54.2452
REPTree	0.5494	858547.17	2075471.79	65.1404	83.4454
Random Forest	0.5913	916947.39	2017344.07	69.5713	81.1083
SMOreg	0.9468	549976.18	820946.21	41.7282	33.0066
IBk	0.5363	901703.37	2103566.89	68.4147	84.5750

Odnos stvarnog i projektovanog prosečnog nivoa odlaznog saobraćaja na petominutnim intervalima prikazan je na Slici 1.



Slika 1. Stvarni i projektovani prosečni nivo odlaznog saobraćaja na petominutnim intervalima

U Tabelama 3 i 4 prikazane su performanse šest najboljih modela predikcije prosečnog nivoa dolaznog saobraćaja merene na skupu podataka za trening, odnosno testiranje, respektivno. Ni kod jednog modela nije registrovan problem prevelikog podudaranja.

Tabela 3. Performanse šest najboljih modela predikcije prosečnog nivoa dolaznog saobraćaja merene na skupu podataka za trening

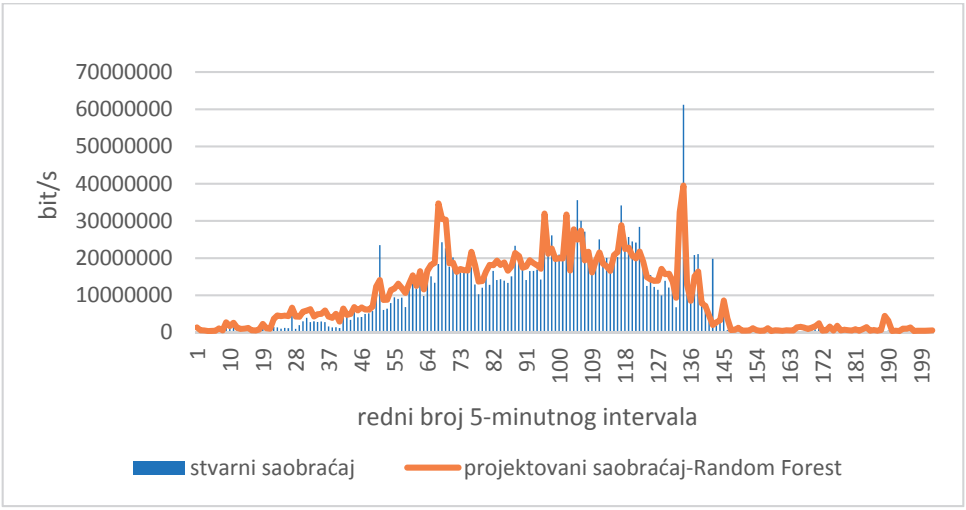
Algoritam	**Koeficijent korelacije**	**Srednja apsolutna greška**	**Kvadratni koren srednje kvadratne greške**	**Relativna apsolutna greška (%)**	**Kvadratni koren relativne kvadratne greške (%)**
Linear Regression	0.9509	1507324.82	4214549.70	15.3450	32.6182
M5P	0.9739	1242286.00	3038584.03	12.6469	23.5169
REPTree	0.9719	1612842.24	3046403.61	16.4192	23.5774
Random Forest	0.9621	1565535.13	3723759.20	15.9376	28.8197
SMOreg	0.9740	1340474.14	3007146.39	13.6465	23.2736
IBk	0.9269	2112275.80	4867982.69	21.5036	37.6754

Tabela 4. Performanse šest najboljih modela predikcije prosečnog nivoa dolaznog saobraćaja merene na skupu podataka za testiranje

Algoritam	**Koeficijent korelacije**	**Srednja apsolutna greška**	**Kvadratni koren srednje kvadratne greške**	**Relativna apsolutna greška (%)**	**Kvadratni koren relativne kvadratne greške (%)**
Linear Regression	0.9214	2161150.38	3993296.35	26.0767	39.9954
M5P	0.9179	2249927.57	4086328.19	27.1479	40.9272
REPTree	0.8923	2502801.66	4631026.55	30.1991	46.3827
Random Forest	0.9388	2087918.36	3482984.23	25.1931	34.8843
SMOreg	0.9112	2170523.99	4180936.22	26.1898	41.8748
IBk	0.8915	2546940.08	4544325.21	30.7317	45.5143

Kao najbolji, izabran je model baziran na algoritmu *Random Forest*, jer je na skupu podataka za trening pokazao dobre performanse (Tabela 3), a na skupu podataka za testiranje najbolje performanse (Tabela 4). Stoga je predikcija zavisne varijable izvršena

korišćenjem ovog modela mašinskog učenja. Odnos stvarnog i projektovanog prosečnog nivoa dolaznog saobraćaja na petominutnim intervalima prikazan je na Slici 2.



Slika 2. Stvarni i projektovani prosečni nivo dolaznog saobraćaja na petominutnim intervalima

5. Zaključak

U okviru ovog istraživanja, za predikciju nivoa dolaznog i odlaznog saobraćaja u posmatranoj lokalnoj računarskoj mreži, predložena je metoda prediktivne analize bazirana na nadgledanom mašinskom učenju. Implementacija opisane metode u studiji slučaja pokazala je da za predikciju odlaznog saobraćaja regresioni algoritmi daju najbolje rezultate, dok je kod ostalih algoritama uočen problem prevelikog podudaranja. Za predikciju dolaznog saobraćaja najbolje performanse pokazao je model baziran na algoritmu iz klase "drvo odlučivanja". Kod predikcije dolaznog saobraćaja ni kod jednog algoritma nije zabeležen problem prevelikog podudaranja. Dakle, predložena metoda predikcije verifikovana je u realizovanoj studiji slučaja.

Zahvalnica

Ovaj rad delimično je podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, u okviru projekta pod brojem 036012.

Literatura

[1] A. Burkov, *The Hundred-page Machine Learning Book*. Andriy Burkov, 2019.

[2] M. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update", *ACM SIGKDD Explorations Newsletter*, Vol. 11(1), pp. 10-18, 2009.

[3] Claes Dalén, Fredric Dahlblom “Big data in the Telekom Industry”, Stockholm School of Economics Department of Marketing and Strategy, May 2014
[4] Z. Wang, G. F. Wei, Y. L. Zhan, et al. “Big data in telecommunication operators: data, platform and practices”, *Journal of communications and information networks*, vol. 2(3), pp.78-91, 2017.
[5] J. Bughin, “Reaping the benefts of big data in telecom”, *Journal of Big Data*, vol. 3(14), https://doi.org/10.1186/s4053 7-016-0048-1, 2016.
[6] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods”, *Journal of Business Research*, vol. 70, pp. 263-286, 2017.
[7] B. G. Buchanan, “A very brief history of artificial intelligence”, *AI Magazine*, vol. 26(4), pp. 53-60, 2005.
[8] X. Guibao, M. Yubo, and L. Jialiang, “Inclusion of Artificial Intelligence in Communication Networks and Services”, *ITU Journal, ICT discoveries*, vol 1(1), pp. 33-38, 2018.
[9] Z. Q. J. Lu, “The elements of statistical learning: data mining, inference, and prediction”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 173(3), pp. 693–694, 2010.
[10] Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, Cisco White Paper, June 2017. [Online]. Available at: https://www.reinvention.be/webhdfs/v1/docs/complete-white-paper-c11-481360.pdf
[11] T. Oetiker, “MRTG-logfile - description of the mrtg-2 logfile format”, May 2017. [Online]. Available at: https://oss.oetiker.ch/mrtg/index.en.html.
[12] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, *Fourth Edition*. Burlington, USA: Morgan Kaufmann, 2017.

Abstract: *Machine learning can be defined as generation of knowledge based on the previous experience, such as data related entities that are subjects of learning. Today, in the Big Data era, machine learning is used as the leading technique in predictive analytics. In this paper a prediction method based on the building and application of supervised machine learning models is described, altogether with the implementation of all stages of the machine learning process in the software tool called Weka (Waikato Environment for Knowledge Analysis). This data mining software is a collection of machine learning algorithms that are used in data mining tasks. The implementation of machine learning process using Weka toolkit is demonstrated on a case study of the traffic prediction. The traffic is generated by different user groups within the observed local area network.*

Keywords: *machine learning, Big Data, network traffic, Weka*

LOCAL AREA NETWORK TRAFFIC PREDICTION USING SUPERVISED MACHINE LEARNING

Slađana Janković, Katarina Kukić, Ana Uzelac, Vladislav Maraš