Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.7
DOI: https://doi.org/10.24867/17BE03Avramovic

PRIMENA MAŠINSKOG UČENJA ZA IDENTIFIKOVANJE LIMFNIH ČVOROVA KOD TIROIDNE ŽLEZDE

MACHINE LEARNING APPLICATION FOR IDENTIFYING LYMPH NODES IN THE THYROID GLAND

Nataša Avramović, Dejan Nemec, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Ovaj rad opisuje kreiranje i analizu modela za identifikovanje papilarnog karcinoma tiroidne žlezde. Predmet istraživanja su pacijenti kod kojih klinički i ultrazvučno nisu detekovane patološke pojave, odnosno metastaze limfnih čvorova. Operativni nalaz tih pacijenata pokazao je da se preoperativni rezultati, odnosno ultrazvuk, ne može smatrati pouzdanim, i da dovodi do greške kod oko 48% pacijenata. Cilj rada je dostići senzitivnost i specifičnost modela koji će pružiti dovoljno dobar rezultat.*

Ključne reči: *mašinsko učenje, klasifikacija*

Abstract – *The paper describes the creation and analysis of a model for the identification of papillary thyroid cancer. The subject of the research are patients in whom no pathological phenomena or lymph node metastases, have been detected clinically and ultrasound. The surgical outcome findings of these patients point out that preop ultrasound diagonosis could not be considered reliable and that they were wrong in about 40% of patients. The aim of this paper is to achieve the sensitivity and specificity of the model that will provide a good enough result.*

Keywords: *machine learning, classification*

1. UVOD

Kako je era velikih podataka uveliko počela, postoji i izražena potreba za automatizovanim metodama njihove analize i obrade. Dostupnost računara danas i jeftine memorije omogućavaju lak način skladištenja podataka, pa čak i onih nepotrebnih.

Primena obrade velikih podataka u medicini može doprineti veliki značaj. Pružanje najbolje usluge i nega pacijenata dobijaju se kreiranjem modela za personalizovanu, prediktivnu, preventivnu svrhu u medicini, i zasnovani su na korišćenju elektronskih zdravstvenih kartona i generisanju velikih količina podataka.

Analitika velikih podataka obuhvata integraciju heterogenih podataka, analizu, modelovanje, tumačenje i validaciju. Rezultati dobijeni analitikom podataka treba da pruže korist kako pacijentima, tako i lekarima.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Živko Bojović, vanr. prof.

Predmet ovog istraživanja jeste grupa pacijenata sa papilarnim karcinomom štitne žlezde, kod kojih limfni čvorovi u bilo kom delu vrata (centralnom ili lateralnom) klinički (pregled lekara, nalaz krvi) i ultrazvučno nisu opisani kao sumnjive ili patološke pojave. Pacijent je preoperativno išao na ultrazvuk te regije, gde se gledao tumor i procenjivalo se da li su limfni čvorovi maligni ili benigni. Zlatni standard u ovom slučaju je rezultat operacije, odnosno operativni nalaz i on je pokazao stvarno stanje. Kako se ultrazvuk pokazao loše u ovom istraživanju, i pogrešio kod oko 40% pacijenata, potrebno je ispitati koji parametri, dobijeni preoperativnom analizom i operacijom, dovode do metastaza u limfnim čvorovima.

U drugom poglavlju iznete su osnove i aspekti koje predstavljaju potporu za ovaj rad. U trećem poglavlju iznet je koncept celokupnog sistema, kao i neki od ključnih aspekata zašto je određena tehnologija izabrana. Detaljan prikaz algoritma predstavljen je u četvrtom poglavlju, kao i specifičnosti koje su se pojavile pri izradi zadatka a takođe i kako su prevaziđene. Postupak testiranja i ostvareni rezultati prikazani su u petom poglavlju, a zaključak u šestom.

2. OSNOVE RADA

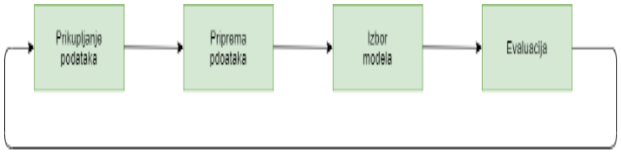
Mašinsko učenje predstavlja jedan od najčešćih oblika veštačke inteligencije, statistike i računarstva, i fokusira se na korišćenje podataka i algoritama tako da imitiraju način učenja kod ljudi. Svakodnevnim prikupljanjem dostupnih podataka omogućava se razvijanje sve većeg broja različitih modela i pametnije donošenje odluka.

Korišćenjem statističkih metoda, ovaj tip učenja obrađuje velike skupove podataka, pronalazi neke obrasce i zakonitosti u njima, pravi predviđanja na osnovu kojih se na kraju donosi odluka. Mašinsko učenje bavi se projektovanjem algoritama koji automatski izvlače bitne informacije iz podataka [1]-[3].

Razvojni ciklus modela mašinskog učenja (Slika 1) je proces izgradnje efikasnog algoritma, tako da se ciklus ponavlja i u svakoj iteraciji dolazi do poboljšanja podataka, modela i bolje procene. Brzina kojom se ciklus ponavlja, predstavlja ono što određuje troškove, a osnovna intervencija je da se izvrši što bolja optimizacija troškova.

U tu svrhu, danas postoje razni dostupni alati. Svrha razvojnog ciklusa je dolazak do rešenja zadatog problema,

tako da ga je najpre potrebno razumeti, jer od toga zavisi krajnji rezultat [4].



Slika 1 *Razvojni ciklus mašinskog učenja*

Prvi korak pri stvaranju i razvoju modela mašinskog učenja je prikupljanje podataka. On ima za cilj identifikaciju i dobijanje svih podataka koji se odnose na problem koji se rešava. Nakon prikupljanja podataka, iste je potrebno pripremiti za dalju obradu. Priprema podataka za dalju obradu odnosi se na statističku analizu podataka, vizuelnu analizu i transformaciju podataka, koje se sprovode sve dok se ne postigne forma potrebna računaru za razumevanje.

Pri razvoju modela mašinskog učenja mora se voditi računa o tome da je cilj da obučeni model dobro radi na novim, neviđenim podacima. Kako bi se došlo do što objektivnije procene performansi modela, podaci se dele na dva dela:

- skup podataka za evaluaciju i
- skup podataka za validaciju.

U ovoj fazi se pripremljeni i obrađeni podaci koriste u izabranom algoritmu za obuku modela, odnosno, model uči iz pripremljenih i obrađenih podataka. Obuka modela se vrši nad trening podacima, a performanse modela se procenjuju na osnovu skupa za validaciju. Na osnovu informacija dobijenih u ovom koraku, može se završiti sa kreiranjem modela i to predstavljanjem zadovoljavajućih rezultata ili vratiti ponovnom obučavanju modela i podešavanju sve dok se ne dobiju zadovoljavajući rezultati [5]-[6].

3. KONCEPT REŠENJA

Grupa pacijenata koja nas interesuje su pacijenti sa papilarnim karcinomom štitne žlezde kod kojih klinički i ultrazvučno nisu opisani sumnjivi ili patološki limfni čvorovi u bilo kom delu vrata, odnosno identifikovan im je benigni tumor, i nisu otkvivene metastaze.

U ovu studiju uključeni su pacijenti sa patohistološki potvđenom dijagnozom papilarnog karcinoma štitaste žlezde koji su operisani na na Institutu za onkologiju i radiologiju Srbije u periodu od januara 2015. godine do septembra 2021. godine.

Podaci su podeljeni tako da se u skupu za obuku nalazi 70% pacijenata, a u skupu za testiranje 30% pacijenata. Obuka je vršena metodom unakrsne validacije, kako bi se izbeglo natprilagođenje modela. U ovoj metodi, skup za obuku deli se na manje podskupove, i u ovom slučaju je odlučeno da u svakom podskupu bude približno 10 pacijenata. Postoji onoliko rundi koliko postoji podskupova, i validacija se vrši tako što je u svakoj rundi jedan podskup korišćen za testiranje, a ostali za obuku.

Ukoliko postoji veliki broj obeležja vršena je selekcija obeležja unapred, time su iskorišćena obeležja koja najviše doprinose rezultatu, i uklonjena obeležja koja ništa ne doprinose, a povećavaju računarsku složenost i vreme izvršavanja. Pored selekcije obeležja, broj obeležja je smanjivan i pomoću rezultata dobijenih iz korelacione matrice, tako da ukoliko su dva obeležja visoko korelisana (pozitivno ili negativno) jedno od njih je uklanjano, jer nose istu informaciju.

Pošto su podaci označeni, odnosno za svakog pacijenta se zna ishod, ova vrsta pravljenja modela spada u nadgledano učenje, u klasifikaciju. Model je testiran na različitim klasifikatorima: *kNN, Decision tree, Logistic regression, Support vector machine, Random forest i XGBoost.*

Performanse klasifikatora merene su pomoću matrice konfuzije i AUC ROC krive. Iz matrice konfuzije računati su: senzitivnost, specifičnost i tačnost.

4. ANALIZA PODATAKA

4.1. Baza 1

Baza podataka sadrži 278 pacijenta, kod kojih je mereno 29 obeležja, i jedno dodatno obeležje u bazi podataka predstavlja klasu pacijenta. Na osnovu informacija dobijenih o bazi, uočeno je da postoje nedostajući podaci, i da je broj takvih podataka 613. Takođe, kod obeležja *Bilateralno_st* i *Lokacija_tumora_rezanj_recode* nedostaje oko 50% podataka, tako da se ta obeležja uklanjaju. Obeležja *ID* i *RB* se takođe uklanjaju, ona ne donose praktičan značaj rezultatu.

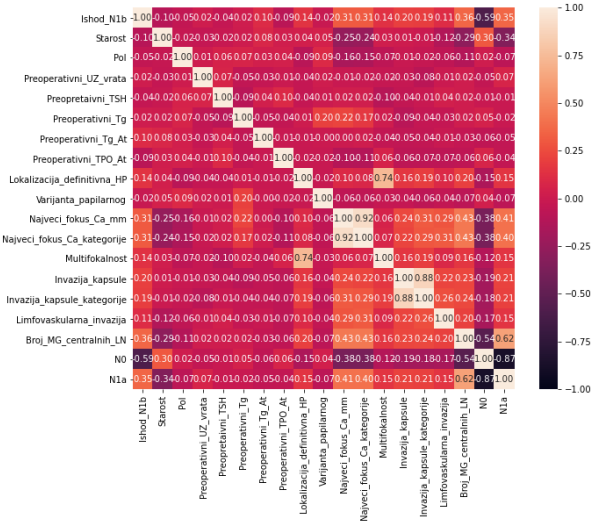
Analizirajući kategorička obeležja iz baze, doneti su dodatni zaključci. Obeležje *Starost* opisuje starost pacijenta, i posmatra se kao numerička varijabla, dok obeležja *Starost_kategorije_55* i *Starost_kategorije_45* opisuju takođe starost, samo se posmatraju kao kategoričke varijable i odnose se na pripadnost određenoj starosnoj grupi. Odavde, radi analize statističkih vrednosti zdravih i bolesnih pacijenata odlučeno je da se samo obeležje *Starost* zadrži. Dalje, obeležja *Preoperativni_UZ_vrata_nominal* i *Preoperativni_UZ_vrata*, takođe se odnose na isto, tako da će u ovom slučaju samo obeležje *Preoperativni_UZ_vrata* biti zadržano. Obeležja *Multifokalnost*, *Broj_fokusa* i *Broj_fokusa_kategorije* takođe se odnose na isto, odnosno na to da li postoji jedan ili više tumora na žlezdi, tako da se samo obeležje *Multifokalnost* zadržava. Obeležja *Najveci_fokus_Ca_i_micro_Ca* i *Najveci_fokus_Ca_kategorije* se odnose na veličinu tumora, tako da se samo jedno zadržava, u ovom slučaju *Najveci_fokus_Ca_kategorije.*

Međuzavisnost obeležja prikazana je pomoću hitmape na slici 2.

Na osnovu slike 2 uočava se visoka korelacija između obeležja *N1a* i *N0*, tako da se jedno od tih obeležja uklanja, ovde je odabrano da to bude obeležje *N1a*. Takođe, visoka korelacija se primećuje kod obeležja *N0* sa ishodom, tako da se ipak i *N0* uklanja. Korelacija između obeležja *Najveci_fokus_Ca_kategorije* i *Najveci_fokus_Ca_mm* je jasna, jer prva predstavlja kategoričko obeležje, a druga isto to, samo numeričko, tako da se zadržava numeričko obeležje. Obeležje

Lokalizacija_definitivna_HP ima visoku korelaciju sa obeležjem *Multifokalnost*, tako da će i ono biti uklonjeno. Visoka međusobna korelacija postoji i kod obeležja *Invazija_kapsule* i *Invazija_kapsule_kategorije*, tako da se prvo obeležje uklanja.

Nakon uklanjanja visoko korelisanih obeležja, rađene su statističke vrednosti za numerička obeležja.



Slika 2 *Korelaciona matrica za bazu 1*

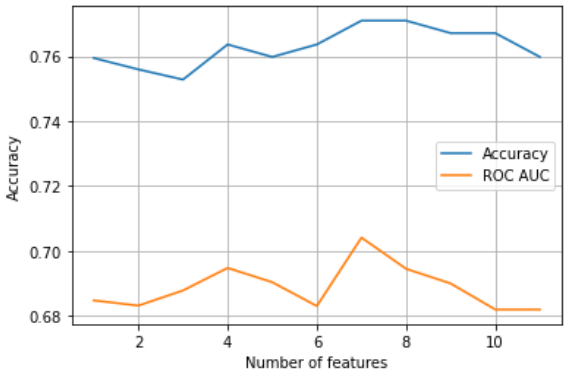
Na osnovu tabele 1, može se zaključiti da su statističke vrednosti zdravih i bolesnih osoba u pogledu ispitivane starosti približne, tako da ovo obeležje ne nosi značajnu informaciju. Obeležje *Preoperativni_TSH* ima duplo veću maksimalnu vrednost kod zdravih osoba, tako da može biti od značaja pri dobijanju rezultata. Obeležje *Preoperativni_Tg* pokazuje približne statističke vrednosti za zdrave i bolesne pacijente, tako da se i ono može isključiti iz daljeg razmatranja.

Obeležje *Preoperativni_Tg_At* pokazuje značajnu razliku u statističkim vrednostima kod zdravih i bolesnih pacijenata, gde pokazuje više od 10 puta veću vrednost maksimuma, i oko 5 puta veću srednju vrednost kod bolesnih pacijenata. Obeležje *Preoperativni_TPO_At* pokazuje 5 puta veću srednju vrednost, i približno duplo veću vrednost kod zdravih pacijenata. Obeležje *Najveci_fokus_Ca_mm* pokazuje razliku u statističkim parametrima, dok obeležje *Broj_MG_centralnih_LN* ne pokazuje značajnu razliku između zdravih i bolesnih pacijenata, tako da se ni ono ne može smatrati bitnim za donošenje rezultata.

Tabela 1 *Statističke vrednosti za bazu 1*

Obeležje	Zdrave osobe			Bolesne osobe		
	Min	Max	Sr. vr.	Min	Max	Sr. vr.
Starost	19	75	47,8	19	75	44,6
Preoperativni_TSH	0.01	11.93	2.15	0.01	5.56	2.0
Preoperativni_Tg	0.04	780	84.2	0.2	732	88.6
Preoperativni_Tg_At	0.2	1128	159.9	0.1	44061	799.89
Preoperativni_TPO_At	0.1	8663	360.32	0.5	1685.5	172.4
Najveci_fokus_Ca_mm	1	45	11.7	1	65	19.2
Broj_MG_centralnih_LN	0	17	1	0	15	3

Kada se nad preostalim obeležjima izvrši selekcija obeležja dobija se vrednost AUC ROC krive oko 71%, i postiže tačnost od oko 77% pri radu sa samim obeležjem (Slika 3).



Slika 3 *Selekcija obeležja*

Pravljenjem modela sa obeležjima koja doprinose rezultatu, dobijaju se rezultati prikazani u Tabeli 2.

Tabela 2 *Evaluaciona metrika za bazu 1*

Klasifikator	Tačnost	Senzitivnost	Specifičnost	Srednja apsolutna greška
kNN	70,24	21,74	88,52	11,31
DT	67,86	34,78	80,33	15,82
LR	72,62	21,74	91,8	4,46
RF	70,24	21,74	88,52	0,01
SVM	73,81	21,74	93,44	16,64
XGBoost	71,43	13,04	93,44	0,01

Na osnovu tabele 2, zaključeno je da nije postignuta odgovarajuća senzitivnost što je bio jedan od uslova.

Analizom dobijenih rezultata sa doktorima koji su radili na prikupljanju podataka, zaključeno je da je došlo do značajnih propusta prilikom pravljenja baze, tako da je ponovna analiza rađena na bazi 2, u kojoj su uočeni problemi rešeni.

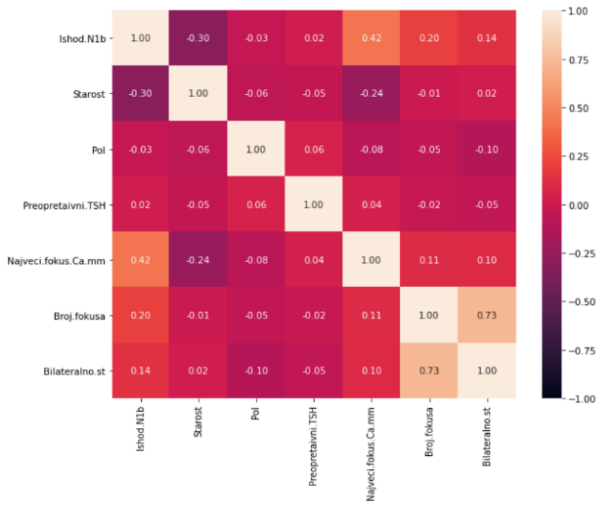
4.2. Baza 2

Baza podataka sadrži 273 pacijenta, kod kojih je mereno 7 obeležja, i jedno dodatno obeležje u bazi podataka predstavlja klasu pacijenta. Osnovnom analizom baze podataka utvrđeno je da svi pacijenti uglavnom imaju sve ispitivane parametre, i da samo u nekim slučajevima parametar *Preoperativni.TSH* nedostaje. Odnosno, nedostajuće vrednosti postoje samo za taj parametar. U zavisnosti od tipa varijable nedostajuće vrednosti se popunjavaju. Kako je parametar *Preoprativni.TSH* numerička varijabla, i kako nedostaje samo 6 vrednosti, što je oko dva posto, odlučeno je da se nedostajuće vrednosti popune metodom srednje vrednošću kolone, jer tako mala količina ubačenih podataka neće mnogo uticati na rezultat.

Međuzavisnost obeležja prikazana je pomoću matrice korelacije na slici 4.

Kao najviše korelisana obeležja mogu se izdvojiti *Broj.fokusa* i *Bilateralno.st*, dok ostala obeležja uglavnom ne pokazuju visoku korelisanost. Između kategorija *Broj.fokusa* i *Bilateralno.st* postoji veza kada se utvrdi da

je tumor multifokalan, odnosno kada postoji više od jednog tumorskog fokusa, a bilateralan znači da je prisutan u oba režnja.



Slika 4 *Matrica korelacije za bazu 2*

Statističke vrednosti za numerička obeležja baze 2 date su u tabeli 3.

Tabela 3 *Statističke vrednosti za bazu 2*

Obeležje	Zdrave osobe			Bolesne osobe		
	Min	Max	Sr. vr.	Min	Max	Sr. vr.
Starost	20	78	51,13	19	77	42,8
Preoperativni .TSH	0.01	11.93	2.07	0.01	10,26	2,15
Najveci.fokus .Ca.mm	1	40	8,85	1	40	15,84
Broj.fokusa	1	7	1	1	6	2

Iz dobijenih rezultata ne uočavaju se značajne razlike između zdravih i bolesnih pacijenata. Selekcija obeležja za bazu 2 nije rađena, pošto je broj obeležja već smanjen na minimum. Rezultati dobijeni testiranjem modela nad različitim klasifikatorima dati su u tabeli 4.

Tabela 4 *Evaluaciona metrika za bazu 2*

Klasifikator	Tačnost	Senzitivnost	Specifičnost	Srednja apsolutna greška
kNN	67,07	75,0	59,52	24,41
DT	69,51	58,41	82,05	23,24
LR	76,83	76,7	77,08	10,91
RF	79,27	80,95	77,5	0,01
SVM	70,73	72,09	69,23	23,20
XGBoost	75,61	69,44	80,43	0,01

Na osnovu tabele 4 primećuje se da je klasifikator sa najboljim performansama RF, odnosno daje najveću senzitivnost što je najznačajniji ispitivani parametar metrike za ovaj problem.

5. ZAKLJUČAK

Preoperativna ultrazvučna dijagnostika nije dovoljno senzitivna u detekciji metastaza u limfnim čvorovima vrata kod pacijenata sa karcinomom štitaste žlezde, čime je onemogućeno adekvatno stažiranje bolesti. U našem istraživanju je na osnovu postoperativno dobijenog patohistološkog nalaza pokazano da su limfonodalne metastaze bile prisutne kod oko 48% pacijenata kod kojih je preoperativni ultrazvučni nalaz bio uredan.

Rezultat dobijen ovom studijom se pokazuje kao veoma dobar jer pruža senzitivnost oko 80%, sa približnom specifičnošću i tačnošću. Takođe, dobijeni su i demografski i patohistološki parametri koji su se pokazali kao najzanačajniji za postizanje rezultata. Veća baza podataka, i univerzalna identifikacija ispitivanih parametara će dovesti do boljih rezultata u budućnosti.

6. LITERATURA

[1] Andread C. Muller, Sarah Guido, *Introduction to Machine Learning with Python*, O'Reilly Media, Inc. 2016.

[2] Kevin P. Murphy, *Machine Learning: a Probabilistic Perspective*, MIT Press, 2012.

[3] Trevor Hastie, Robert Tibishirani, Jeronime Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[4] Martin Krzwinski, Naomi Altman, *Classification and regression trees*, Nature America, Nature Methods, Vol. 14 No. 8, August 2017.

[5] Jake Lever, Matrin Krzywinski, Naomi Altman, *Model selection and overfitting*. Nature America, Nature Methods, Vol. 13 No. 9, September 2016.

[6] Candice Bentejac, Anna Csorgo, Gonzalo Matrinez-Munoz, *A Comparative Analysis of XGBoost*., ResearchGate, November 2019.

Kratka biografija:

Nataša Avramović rođena je u Loznici 1997. godine. Osnovne akademske studije, na smeru Komunikacione tehnologije i obrada signala, završila je 2020. godine, sa odbranom bečelor rada pod naslovom „Praćenje kretanja i zadržavanja ljudi u zatvorenom prostoru primenom kompjuterske vizije". Do sada je objavila jedan naučni rad na međunarodnom skupu TELFOR.

kontakt: avramovicnatasa97@gmail.com

Dejan Nemec rođen je 1972. god. Diplomirao, specijalizirao i magistrirao je na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva. Oblast interesovanja su telekomunikacije i obrada signala.

Zahvalnica:

Izradu ovog rada pomogao je Fakultet tehničkih nauka u Novom Sadu, Departman za energetiku elektroniku i telekomunikacije, u okviru projekta pod nazivom „Razvoj i primena savremenih metoda u nastavi i istraživačkim aktivnostima na Departmanu za energetiku, elektroniku i telekomunikacije".