Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: https://doi.org/10.24867/29BE25Vajda
DOI: 004.9

PREDIKCIJA HRONIČNE BOLESTI BUBREGA NA OSNOVU SCINTIGRAFSKIH SNIMAKA UPOTREBOM METODA MAŠINSKOG UČENJA

PREDICTION OF CHRONIC KIDNEY DISEASE BASED ON SCINTIGRAPHIC IMAGES USING MACHINE LEARNING METHODS

Kristina Vajda, *Fakultet tehničkih nauka, Novi Sad*

Oblast – Veštačka inteligencija i mašinsko učenje

Kratak sadržaj – *U ovom radu istražena je primena mašinskog učenja za predviđanje stadijuma hronične bolesti bubrega na osnovu podataka ispitnika i slika dobijenih scintigrafijom. Analizirani su klinički i demografski podaci kao i podaci dobijeni iz slika, a primenjeni modeli su: metod k - najbližih suseda, stablo odluke, metod slučajne šume i gradijentnog pojačavanja. Rezultati su pokazali da metod k - najbližih suseda ima najnižu tačnost od 24%, dok su stablo odluke i metod slučajne šume pokazali znatno bolje performanse sa tačnošću od 96%. Model gradijentnog pojačavanja se najbolje pokazao postigavši tačnost od 100%.*

Ključne reči: *Mašinsko učenje, scintigrafija, hronična bolest bubrega*

Abstract – *This paper investigates the application of machine learning data for predicting the stage of chronic kidney disease based on examinee data and images obtained by scintigraphy. Clinical and demographic data, as well as data obtained from images, were analyzed, and the following models were applied: k-nearest neighbors, decision tree, random forest and gradient boosting. The results showed that the k-nearest neighbors method had the lowest accuracy of 24%, while the decision tree and random forest methods showed significantly better performance with an accuracy of 96%. The gradient boosting model performed the best, achieving an accuracy of 100%.*

Keywords: *Machine learning, scintigraphy, chronic kidney disease*

1. UVOD

Veštačka inteligencija i mašinsko učenje i njihova primena predstavljaju jedan od najvećih izazova XXI veka. Na prvi pogled primena deluje poprilično jednostavno i tiče se automatizacije osnovnih čovekovih delatnosti, kao što je rad u fabrici, rad na otvorenom pa čak i prevoz materijala, ali danas mašinsko učenje i veštačka inteligencija sve više i više uspevaju da u pojedinim industrijama zamene čoveka mašinom. Poslednjih godina, veštačka inteligencija je dobila široku primenu u medicini korišćenjem specijalizovanih alata za pronalazak veza unutar velikih skupova podataka. Otkriće teških bolesti u ranim stadijuma, automatizovano prepoznavanje bolesti koje u nekim slučajevima prevazilazi veštine medicinskih stručnjaka govori o važnosti naučnih radova baziranih na veštačkoj inteligenciji i mašinskom učenju u medicini.

Scintigrafija je dijagnostička metoda koja koristi radioaktivne izotope za generisanje slika unutrašnjih organa. Ove slike su važne za identifikaciju različitih zdravstvenih stanja, međutim njihova analiza može biti izazovna. Jedan od najvećih izazova u radu sa medicinskim podacima jeste osiguravanje tačnosti i pouzdanosti rezultata, što zahteva dobre tehnike preprocesiranja i analize podataka. U ovom radu istražuje se kako se mašinsko učenje može koristiti za prevazilaženje ovih izazova i poboljšanje dijagnostičkih procesa.

Predmet istraživanja rada predstavlja metode i tehnike pomoću kojih, na osnovu kliničkih i demografskih podataka kao i podataka dobijenih iz slike, se može predvideti stadijum hronične bolesti bubrega.

Cilj ovog rada je da pruži sveobuhvatan pregled primene mašinskog učenja u analizi scintigrafskih podataka i rešenja u ovom polju.

2. TEORIJSKE OSNOVE I DEFINICIJE

U ovom poglavlju biće predstavljene teorijske osnove algoritama mašinskog učenja koji su korišćeni za predikciju stadijuma hronične bolesti bubrega.

2.1. Metod k – najbližih suseda

Metod k - najbližih suseda (eng. k – nearest neighbors – kNN) spada u grupu metoda kasnog učenja, koje podrazumevaju odlaganje obrade uzoraka za obuku do trenutka kada treba klasifikovati neobeleženi uzorak. Algoritam k - najbližih suseda je jedan od najjednostavnijih algoritama mašinskog učenja za probleme klasifikacije. Zbog svoje jednostavnosti, on je često i jedan od prvih algoritama mašinskog učenja sa čijom primenom se pokuša rešiti određeni problem. U pitanju je intuitivan algoritam koji klasifikuje nepoznati uzorak na osnovu klase pripadnosti susednih uzoraka iz skupa za obuku. Prilikom klasifikacije neobeleženog uzorka, na osnovu odabrane metrike se meri njegova udaljenost od svakog uzorka iz skupa za obuku. Klasa neobeleženog uzorka predviđa se na osnovu klasne

---

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Dunja Vrbaški, red. prof.

pripadnosti k uzoraka iz skupa za obuku koji su najbliži neobeleženom uzorku, jer se pretpostavlja da je dati uzorak sličniji uzorcima iz skupa za obuku koji su mu bliži u prostoru obeležja [1,2].

2.2. Stablo odluke

Stablo odluke predstavlja jednu od često korišćenih metoda nadgledanog učenja. Može da se koristi za rešavanje i klasifikacionih i regresionih problema i može da radi i sa numeričkim i sa kategoričkim obeležjima. Koren stabla predstavlja čvor koji sadrži skup svih uzoraka, i od njega se stablo grana, odnosno vrši se sukcesivna particija skupa uzoraka na dva disjunktna podskupa. Particija se formira na osnovu vrednosti obeležja koje se promeni kao najznačajnije u datom trenutku. Prema odgovoru na postavljeno pitanje, odnosno prema vrednosti datog obeležja za svaki pojedinačni uzorak, skup pridružen čvoru deli se na dva podskupa, čime se ujedno formiraju i dva nova čvora. Algoritam pronalazi optimalno pitanje na osnovu kojeg vrši particiju čvora prolaskom kroz skup mogućih pitanja i procenom u kojoj meri će čvorovi formirani na osnovu svakog od tih pitanja biti čisti. Čistoća određenog skupa, odnosno čvora, podrazumeva što izraženiju dominaciju jedne od klasa u njemu. Kriterijum za zaustavljanje grananja može biti dobijanje potpuno čistog čvora, a može se zasnivati i na ograničenju u pogledu maksimalnog broja uzoraka u krajnjem čvoru ili maksimalne dubine stabla. Klasa test uzorka kod klasifikacionih problema najčešće se određuje kao klasa koja preovladava među uzorcima u listu [1].

2.3. Metod slučajne šume

Metod slučajne šume (eng. random forest) je metoda ansambalskog učenja koja koristi stabla odluke kao jednostavne klasifikatore. Ideja je da se obuči mnoštvo stabala odluke, a donošenje krajnje odluke o klasi ili vrednosti nepoznatog uzorka vrši se glasanjem u slučaju klasifikacije. Na početku se, primenom metode ponovnog uzorkovanja iz skupa za obuku, formira M novih skupova. Metoda ponovnog uzorkovanja podrazumeva kreiranje novih skupova za obuku nasumičnim izvlačenjem uzoraka sa vraćanjem iz originalnog skupa. Kada je svaki od novodobijenih skupova za obuku iste veličine kao originalni skup, on tipično sadrži oko 2/3 jedinstvenih uzoraka. Svako od stabala odluke u okviru metode slučajne šume zatim se obučava, odnosno formira na osnovu jednog od M novodobijenih skupova za obuku. Kod ovog algoritma greška se procenjuje na osnovu tačnosti predviđanja za uzorke koji u postupku ponovnog uzorkovanja nisu nijednom izvučeni iz polaznog skupa [1].

2.4. Gradijentno pojačavanje

Po pitanju preciznosti, metode gradijentnog pojačavanja (eng. gradient boosting) su među najboljim metodama pojačavanja i mašinskog učenja uopšte. Obučavanje se takođe može smatrati relativno efikasnim. Zbog toga su vrlo popularne u primenama. Osnovna ideja dolazi iz gradijentnih metoda optimizacije, koje počivaju na popravljanju tekućeg rešenja optimizacionog problema dodavanjem vektora proporcionalnog negativnoj vrednosti gradijenta funkcije koja se minimizuje. Ovo ima smisla, pošto negativna vrednost gradijenta pokazuje smer opadanja funkcije. U kontekstu pojačavanja, ova ideja se realizuje tako što se model Fm kojim se dopunjuje ansambl Fm-1 odredi tako da aproksimira gradijent greške po funkciji Fm-1 [3].

3. METODOLOGIJA

U ovom poglavlju biće predstavljen proces implemetacije sistema za klasifikaciju stadijuma hronične bolesti bubrega koji se kreće od 1 (početni stadijum) do 5 (napredni stadijum).

Skup podataka je javno dostupan [4] i uključuje demografske i kliničke podatke ispitanika kao što su starost, pol, težina, visina, sistolni i dijastolni krvni pritisak pre i posle dinamičke renalne scintigrafije, BMI, BSA – H (površina tela prema Haycock metodi) i BSA – D (površina tela prema DuBois metodi). Sastoji se od 107 ispitanika od kojih je svaki pojedinačni ispitanik imao minimum 4, a maksimum 6 slika. Slike koje su snimane su posteriorne i anteriorne projekcije dinamičke renalne studije, posteriorne i anteriorne projekcije flood izvora odnosno transmisijski podaci i snimci nakon mokrenja u posteriornim i anteriornim projekcijama.

Stadijum hronične bolesti bubrega je procenjen na osnovu brzine glomuralne filtracije (eng. glomerular filtration rate - GFR). Brzine glomuralne filtracije je procenjena regresijom od klirensa etilendiamintetrasirćetne kiseline (EDTA) dostupnog kod nekih ispitanika i merkaptoacetiltriglicin (MAG3) klirensa dostupnog kod svih ispitanika.

Kako bi se podaci iz ovih skupova podataka pripremili za modele mašinskog učenja, kategoričko obeležje koje se odnosi na pol ispitanika je zamenjeno numeričkim, odnosno kategorička vrednost M menja se numeričkom vrednošću 0, dok se F menja sa 1.

Uzevši u obzir da je srednja vrednost osetljiva na vrednosti koje odstupaju od ostalih, u ovom radu nedostajuće vrednosti su zamenjene medijanom.

Skup podataka proširen je novim atributima koji bi bili zasnovani na podacima dobijenih iz slike. Pre samog izvlačenja obeležja iz slika izvršeno je uklanjanja šuma iz slike i normalizacija intenziteta piksela. Takođe je bilo potrebno primeniti promenu veličine svih slika, da bi slike bile istih dimenzija, što je važno prilikom korišćenja algoritama mašinskog učenja.

Sledeći korak u pripremi podataka je dobijanje segmentiranih slika.

U radu [5], Otsu-ova metoda se analizira kao jedna od najvažnijih globalnih metoda za određivanje praga slike. Osim što je ova metoda istaknuta kao jedna od najpopularnijih i najčešće korišćenih metoda za određivanje praga slike zbog njene jednostavne implementacije i dobrih performansi, pokazala se efikasnom za slike sa dobrim kontrastom između objekta i pozadine. Ova metoda je ocenjena kao robusna i efikasna tehnika za određivanje praga slike u mnogim situacijama, stoga će biti primenjena za dobijanje segmentirane slike u ovom radu. Otsu-ov prag se koristi za automatsko određivanje praga, a prag se izračuvana na osnovu

histograma slike i traži se vrednost koja maksimizira razliku između dve klase piksela [6].

Taj prag se zatim primenjuje na sliku i dobija se binarna slika u kojoj pikseli veći od praga imaju vrednost True, a pikseli manji od praga imaju vrednost False. Na ovaj način je moguće u slici razlikovati objekte od pozadine. Uklonjeni su svi binarni objekti koji se nalaze na ivicama slike da bi se poboljšala tačnost segmentacije. Zatim su obeleženi svi povezani segmenti istom oznakom.

Naredni korak podrazumeva izračunavanje srednje vrednosti i standardne devijacije svih piksela u slici, kao i asimetriju (eng. skewness) i špicastost (eng. kurtosis) rasporede piksela. Ova statistička obeležja pružaju osnovne informacije o distribuciji piksela u slici, koje su korisne za razumevanje globalnih karakteristika slike. Izračunata su svojstva označenih segmenata u segmentiranoj slici i iz svakog segmenta se za obeležja uzimaju površina regiona, obim regiona, solidnost (odnos površine i konveksnog omotača regiona) i eksentričnost regiona (mera odstupanja regiona od idealne elipse). Kombinovanjem svih ovih izračunatih vrednosti sa kliničkim podacima generiše se konačan skup obeležja.

Podaci su podeljeni na trening i test skup u odnosu 80:20.

Na ovaj način skup podataka od 578 uzoraka podeljen je na trening i test skup, tako da trening skup sadrži 462 uzorka, dok test skup sadrži 116 uzoraka.

Nakon podele na trening i test skup, vrši se dodavanje nula gde je potrebno kako bi se osiguralo da svi skupovi obeležja odnosno svi uzorci imaju istu dužinu.

S obzirom da skup podataka relativno mali, izvršena je augmentacija podataka da bi se proširio trening skup podataka koji će se koristiti za obučavanje modela.

Augmentacija podataka izvršena je rotacijom svake slike za 90, 180 i 270 stepeni. Za svaku rotiranu sliku ponovljen je postupak ukanjanja šuma, normalizacije intenziteta piksela, promena veličine slike, segmentacije slike i izvlačenje obeležja iz slike. Rotacija je oblik augmentacije podataka koja povećava raznolikost skupa podataka, što može poboljšati generalizaciju modela pri upotrebi za neviđene podatke.

Od ukupnih 462 uzorka iz trening skupa nakon izvršene augmentacije podataka dobijeno je 1848 uzoraka.

U ovom radu za optimizaciju hiperparametara korišćena je pretraga po mreži (eng. grid search). Pretraga po mreži je najjednostavnija tehnika koja se koristi kada broj hiperparametara i njihov opseg nisu preveliki [7]. GridSearchCV() pretražuje zadati skup hiperparametara i pronalazi najbolju kombinaciju. Za svaku kombinaciju hiperparametara, GridSearchCV() vrši unakrsnu validaciju i izračunava performanse modela [8].

4. REZULTATI I DISKUSIJA

Prvi korišćeni model je kNN odnosno model k - najbližih suseda. Prema tabeli 1 mogu se viditi mere kvaliteta modela i zaključiti da se ovaj model nije baš najbolje pokazao na ovom skupu podataka. Dobijena je tačnost od 24% što je prilično nisko. Preciznost i osetljivost su takođe niske za većinu klasa. Najviša F1-mera je za klasu 3, ali i to je samo 34% što takođe nije zadovoljavajuće.

Nakon ovog modela primenjen je model stabla odluke koji daje znatno bolje mere kvaliteta, što se može videti u tabeli 2. Tačnost od 96% je vrlo visoka. Preciznost, osetljivost i F1-mera su visoki za sve klase, osim za klasu 5 gde je osetljivost 69%, što je još uvek prihvatljivo. Generalno, stablo odluke daje veoma pouzdane i dobre rezultate.

Tabela 1. *Rezultati modela k – najbližih suseda na test skupu*

	PRECIZNOST	OSETLJIVOST	F1 - MERA
1	0.83	0.16	0.27
2	0.25	0.15	0.19
3	0.29	0.42	0.34
4	0.15	0.26	0.19
5	0.07	0.08	0.07
TAČNOST			0.24
MAKRO PROSECI	0.32	0.22	0.21
TEŽINSKI PROSECI	0.38	0.24	0.24

Tabela 2. *Rezultati modela stabla odluke na test skupu*

	PRECIZNOST	OSETLJIVOST	F1 - MERA
1	0.97	1.00	0.98
2	1.00	1.00	1.00
3	1.00	1.00	1.00
4	0.82	0.95	0.88
5	1.00	0.69	0.82
TAČNOST			0.96
MAKRO PROSECI	0.96	0.93	0.94
TEŽINSKI PROSECI	0.96	0.96	0.96

Tabela 3. *Rezultati metode slučajne šume na test skupu*

	PRECIZNOST	OSETLJIVOST	F1 - MERA
1	0.97	1.00	0.98
2	1.00	1.00	1.00
3	1.00	1.00	1.00
4	0.82	0.95	0.88
5	1.00	0.69	0.82
TAČNOST			0.96
MAKRO PROSECI	0.96	0.96	0.94

TEŽINSKI PROSECI	0.96	0.96	0.96

Tabela 4. *Rezultati modela gradijentnog pojačavanja na test skupu*

	PRECIZNOST	*OSETLJIVOST*	*F1 - MERA*
1	1.00	1.00	1.00
2	1.00	1.00	1.00
3	1.00	1.00	1.00
4	1.00	1.00	1.00
5	1.00	1.00	1.00
TAČNOST			1.00
MAKRO PROSECI	1.00	1.00	1.00
TEŽINSKI PROSECI	1.00	1.00	1.00

Sledeći model je metod slučajne šume koji je postigao identične rezultate kao i stablo odluke, sa tačnošću od 96%, što je pokazano u tabeli 3.

Nakon modela slučajnih šuma, testiran je i model gradijentnog pojačavanja koji je postigao savršene rezultate sa tačnošću od 100%, prikazan u tabeli 4. Sve klase su tačno klasifikovane sa preciznošću, osetljivošću i F1 - merom od 100%. Ovo može ukazivati da je gradijentno pojačavanje previše prilagođen na trening skupu, što može biti znak natprilagođenja, ali ovde je očigledno model tačno klasifikovao sve primere iz test skupa.

5. ZAKLJUČAK

U ovom radu istražena je primena mašinskog učenja u analizi podataka dobijenih scintigrafijom sa ciljem predviđanja stadijuma hronične bolesti bubrega.

Nakon preprocesiranja podataka i odabira atributa, kreirani su modeli. Trenirani su metod k – najbližih suseda, stablo odluke, metod slučajne šume i gradijentno pojačavanje.

Zatim su evaluirane su njihove performanse.

Rezultati su pokazali da metod k - najbližih suseda nije dao zadovoljavajuće performanse, sa tačnošću od samo 24% i niskim vrednostima preciznosti, osetljivosti i F1-mere za većinu klasa. S druge strane, stablo odluke i metod slučajne šume pokazali su značajno bolje rezultate sa tačnošću od 96%, visokim preciznostima i osetljivostima za sve klase osim klase 5, gde je osetljivost bila nešto niža, ali i dalje prihvatljiva.

Model gradijentnog pojačavanja dao je savršene rezultate sa tačnošću od 100% za sve klase.

Analizom ovih rezultata zaključuje se da su stablo odluke, metod slučajnih šuma i gradijentno pojačavanje veoma efikasni modeli za klasifikaciju scintigrafskih podataka i predviđanje stadijuma hronične bolesti bubrega na osnovu podataka dobijenih scintigrafijom. Ovi modeli mogu značajno doprineti poboljšanju dijagnostičkih procesa i pružiti podršku medicinskim stručnjacima u donošenju preciznijih odluka.

Primena mašinskog učenja u analizi scintigrafskih podataka pokazuje veliki potencijal za unapređenje medicinske dijagnostike. Dalja istraživanja bi trebalo da se fokusiraju na testiranje ovih modela na većem i raznovrsnijem skupu podataka. Takođe, bilo bi korisno istražiti kombinaciju ovih modela sa drugim tehnikama mašinskog učenja i integrisati ih u šire sisteme za podršku dijagnostici.

6 LITERATURA

[1] T. Nosek, B. Brkljač, D. Despotović, M. Sečujski, T. Lončar-Turukalo, Praktikum iz mašinskog učenja, 1st ed., Novi Sad: Univerzitet u Novom Sadu, 2020, pp. 122.

[2] P. Cunningham, S. J. Delany., ”k-Nearest neighbour classifiers.”, ACM Computing Surveys, vol. 54(6), pp. 1-25, April, 2007.

[3] M. Nikolić, A. Zečević, Mašinsko učenje, Beograd: Matematički fakultet Univerziteta u Beogradu, 2019.

[4] https://dynamicrenalstudy.org/ (pristupljeno u maju 2024.)

[5] M. Sezgin, B. Sankur., "Survey over image thresholding techniques and quantitative performance evaluation." Journal of Electronic Imaging, vol. 13(1), pp. 146-165, 2004.

[6] N. Otsu., (1979). "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, vol. 9(1), pp. 62-66,1979.

[7] D. Vidaković, Tehnologije i alati u mašinskom učenju, Novi Sad: Fakultet tehničkih nauka Univerziteta u Novom Sadu, 2023.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (pristupljeno u junu 2024.)

Kratka biografija:

Kristina Vajda je rođena 1999. godine u Vrbasu. Završila je gimnaziju u Bečeju, 2018. godine. Nakon završene srednje škole, upisuje Fakultet tehničkih nauka u Novom Sadu, smer Biomedicinsko inženjerstvo. Nakon završenih osnovnih akademskih studija, upisuje 2022. godine master akademske studije na Fakultetu tehničkih nauka, smer Veštačka inteligencija i mašinsko učenje, modul Veštačka inteligencija u medicini. Ispunila je sve obaveze i položila je sve ispite predviđene studijskim programom sa prosečnom ocenom od 9,76.
kontakt: vajda.kristina2@gmail.com