Analiza i modelovanje klasifikatora za skup podataka o terapijama lečenja bradavica

Ilija Rakočević, IN59/2018, rakocevicilija7@gmail.com

I. Uvod

U medicinskim istraživanjima, jedna od najvažnijih oblasti jeste oblast koja se bavi kožnim oboljenjima. Među kožnim bolestima istraživači uglavnom primenjuju metode mašinskog učenja na lečenje melanoma. Pored ovog tipa, izveden je niz studija i na mnogim drugim tipovima kožnih oboljenja. Međutim, koliko je poznato, istraživanja na polju lečenja bradavica nisu toliko zastupljena. Tema ovog izveštaja jeste upravo analiza i formiranje klasifikatora, čija uloga je da za datog pacijenta predvidi ishod terapije lečenja bradavica. Pacijent ima određene karakteristike, a klasifikator treba da nauči pravilnosti koje postoje među njima. Na osnovu tih pravilnosti, klasifikator treba biti u mogućnosti da donese tačnu odluku o tome da li je terapija bila uspešna ili ne. Takođe, treba da na pravilan način klasifikuje novopristigli uzorak.

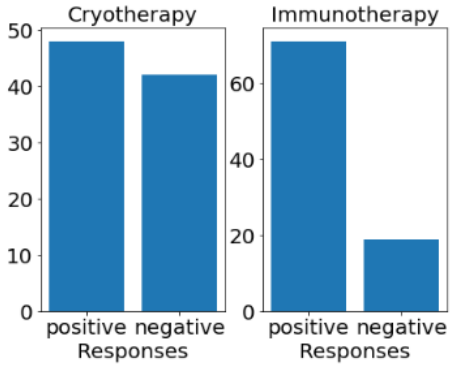
II. Baza podataka

Baza je podeljena na dva skupa i sadrži podatke o 180 pacijenata. Prvi skup sadrži 90 pacijenata na koje je primenjena imunoterapija, a drugi, podatke o preostalih 90 pacijenata sa sprovedenom krioterapijom. Za svakog pacijenta poznata je informacija o polu, starosti, vremenu trajanja bradavice, broju bradavica, vrsti bradavice, površini koju zauzimaju, kao i podatak da li je pacijent odreagovao na terapiju ili nije. Dodatno za imunoterapiju je beležena informacija o prečniku induracije. Iz tog razloga, skup za krioterapiju ima 7 obeležja, a skup za imunoterapiju 8 obeležja.

III. Analiza podataka

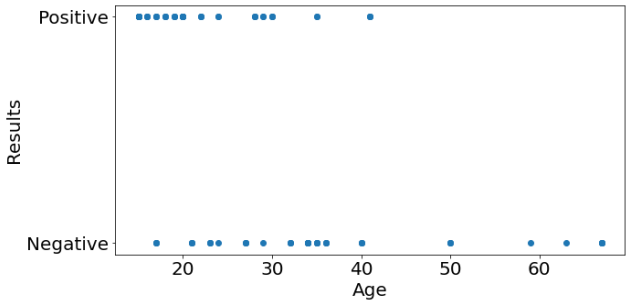
Pošto su podaci bili podeljeni na dva skupa, analiza je vršena na svakom posebno. Kasnije, kada je trebalo primeniti algoritme klasifikacije, skupovi su spojeni u jedan. Prečnik induracije kao obeležje ne postoji u skupu za krioterapiju, tako da su prilikom spajanja te NaN vrednosti popunjene sa vrednošću 0.

Primećuje se da ni jedan skup nema null ili nelogičnih vrednosti.



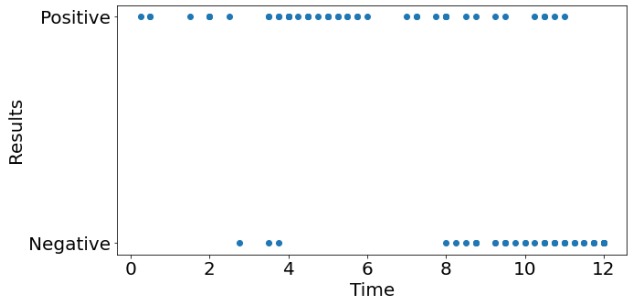
Slika 1. Ishod terapija

Jedan od najvažnijih zapažanja jeste taj da postoji značajna razlika u ishodu terapija (Slika 1.). Analiza pokazuje da krioterapija kod 90 praćenih pacijenata ima pozitivan ishod na svega 48 pacijenata, što je približno 53%. Sa druge strane, imunoterapija ima pozitivan ishod kod 71 pacijenta, što je čini 80% uspešnom.



Slika 2. Odnos ishoda i god. pacijenta kod krioterapije

Što se tiče pozitivnog ishoda terapije, uglavnom ga imaju mlađi pacijenti, do nekih 40 godina starosti (Slika 2.). Pacijentima koji imaju preko 40 godina, u velikoj verovatnoći terapija nije pomagala.



Slika 3. Odnos ishoda i proteklog vremena kod krioterapije

Primetno je da skoro svi pacijenti, koji su čekali duže od 8 meseci pre nego što su krenuli na krioterapiju, nisu izlečeni (Slika 3.). Dolazi se do zaključka da pacijenti koji imaju manje od 40 godina i koji nisu mnogo čekali kada su videli oboljenje imaju veće šanse da se izleče krioterapijom.

IV. PRIPREMA

Pre početka obuke modela neophodno je podeliti podatke na trening i test podatke. Trening skup podataka će biti korišćen za dalju obuku modela dok će test podaci biti skriveni od modela i samog stvaraoca modela. Oni će biti primenjeni na samom kraju, kada budu dobijeni optimalni parametri za klasifikator. Na taj način sprečiće se naknadno podešavanje parametara spram uzoraka u test skupu, odnosno preobučavanja modela. Vrednosti u skupovima podataka se, najčešće, dobijaju indirektno na različitim kvantitativnim skalama i to u različitim mernim jedinicama: frekvencijama, bodovima itd. Tako dobijeni podaci nisu kompatibilni, pa ih je neophodno transformisati u neki pogodan oblik, kako bi bilo moguće da se tačnije analiziraju relacije između podataka. Iz tog razloga ćemo standardizovati podatke na nultu srednju vrednost i jediničnu varijasu. Za obuku biće korišćena unakrsna validacija. Metod koji omogućava iskorišćenost svih uzoraka u testiranju. Takođe ovaj metod podrazumeva kreiranje više modela. Broj modela koji se kreira odgovara broju napravljanenih podskupova. U svakoji iteraciji koristi se N-1 podskup za obuku i jedan podskup za testiranje. Kako bi se napravio što bolji model, neophodno je obratiti pažnju da pri podeli na podskupove budu održani klasni odnosi.

Pored ove pripreme, moramo se osvrnuti na tip problema koji razmatramo, kako bismo znali spram koje mere uspešnosti da upravljamo model. Generalno u oblasti medicine, ne tretira se isto greška ako za bolesnog pacijenta kažemo da je zdrav i za zdravog kažemo da je bolestan. Prema tom pravilu moramo i obučiti naš model. Kako je ovo problem binarne klasifikacije, gde prediktujemo ishod terapije, razmatramo binarnu matricu konfuzije, u kojoj je pozitivna klasa „pacijent je izlečen“, a negativna klasa „terapija nije uspela“. Posmatraćemo meru preciznost i nju ćemo maksimizovati, jer nam ona govori da one koje smo proglasili zdravima da su zapravo zdravi.

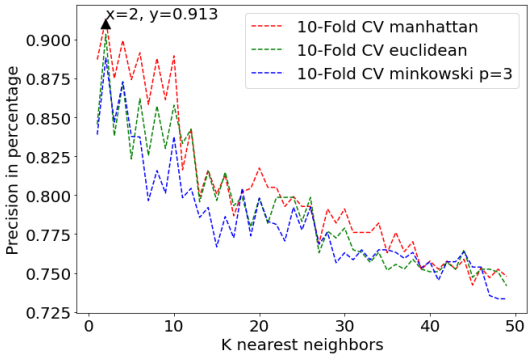
V. KNN

A. *O algoritmu*

Za prvi model upotrebljen je klasifikator k najbližih suseda. Pripada algoritmima nadgledanog učenja. Može se koristiti za rešavanje problema klasifikacije i regresije, s tim što se u kontekstu regresije ređe koristi. U pitanju je intuitivan algoritam koji klasifikuje nepoznati uzorak x na osnovu klasne pripadnosti susednih uzoraka x1, x2, ..., xK iz skupa za obuku. Parametri koji treba da se odrede kako bi klasifikator davao optimalna rešenja jesu: broj suseda koji se uzima u obzir i metrika koja se koristi za određivanje rastojanja.

B. *Odabir optimalnih parametara*

Kao što je već pomenuto, evaluacija parametara je rađena uz pomoć metode unakrsne validacije i to deljenjem na 10 *foldova*. Metrike koje su isprobavane su: *manhattan*, *euclidean* i *minkowski* sa parametrom p=3. Takođe testirane su različite vrednosti za broj suseda. Najbolja vrednost za preciznost je postignuta u iteraciji sa 2 suseda i *manhattan* metrikom (Slika 4.). Takođe isti postupak unakrsne validacije sproveden je nad metrikama *„euclidean“* i *„minkowski“*. U narednim korišćenjima algoritma su korišćeni ovi parametri.



Slika 4. Testiranje parametara

VI. RANDOMFOREST

A. *O algoritmu*

Ideja je da se obuči mnoštvo stabala odluke. Pritom, donošenje krajnje odluke o klasi ili vrednosti nepoznatog uzorka vrši se glasanjem u slučaju klasifikacije, odnosno usrednjavanjem dobijenih rezultata u slučaju regresije. Na početku se, primenom *bootstrap* metode formira *M* novih skupova. *Bootstrap* metoda 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.

B. *Odabir optimalnih parametara*

Za pronalazak optimalnih parametara korišćena je

ugrađena funkcija iz paketa *sklearn* pod nazivom *GridSearchCV*. Ovoj funkciji smo kao parametre prosledili model koji obučavamo, broj estimatora, kriterijum grananja, maksimalnu dubinu stabala, minimalni broj uzoraka da bi nastavio sa podelom unutrašnjih čvorova, da li da koristi *bootstrap* metod ili ne i klasne težine. Takođe, postavili smo broj *foldova* da bude 10 za unakrsnu validaciju. Optimalni parametri koji su dobijeni: broj estimatora : 20, kriterijum: *gini*, maksimalna dubina: 5, minimalni broj uzoraka u čvoru: 0.05, klasne težine: *balanced*, da li da koristi bootstrap metod: *False*. Ovako dobijeni parametri su korišćeni u sledećim korišćenjima algoritma.

VII. Logistic regression

A. O algoritmu

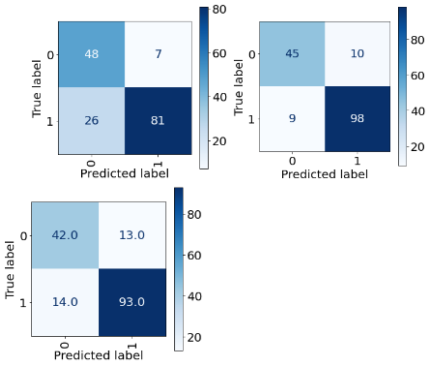
U slučaju binarne klasifikacije (0/1), model linearne regresije mogao bi biti primenjen na klasifikaciju postavljanjem odgovarajućeg praga odlučivanja, npr. T = 0.5 za izlaznu promenljivu. Međutim, opseg izlazne promenljive nije ograničen na [0, 1] i sam model je zavisan od uzoraka u skupu za obuku, jer dodavanjem novih uzoraka granica odlučivanja može značajno da se promeni. Zaključuje se da linearna regresija nije pogodan model za klasifikaciju. Kako bi se uočeni problem prevazišao, umesto osnovnog modela linearne regresije potrebno je naći funkciju čiji kodomen pokriva opseg [0, 1]. Logistička regresija definiše se kao generalizovan linearni regresioni model koji za dati uzorak na ulazu predviđa verovatnoću da uzorak pripada klasi y = 1 koristeći pogodnu nelinearnu funkciju (sigmoidalna ili logistička funkcija).

B. Odabir optimalnih parametara

Parametri koji su došli u obzir: *solver* koji predstavlja algoritam kojim će se vršiti optimizacija i *max_iter* sa kojim se iteracije algoritma koji optimizuje ograničavaju baš na tu vrednost. Konkretno za solver su ponuđeni sledeći algoritmi optimizacije: *lbfgs*, *newton-cg*, *liblinear*, *sag*, *saga*, a za broj iteracija vrednosti: 100, 200, 400, 500. Dobijena je najveća preciznost za kombinaciju solver: *liblinear* i broj iteracija 100. Od sada će ove vrednosti biti korišćene za logističku regresiju.

VIII. Poređenje performansi na trening skupu

Za sva tri klasifikatora je urađena priprema, nađeni su optimalni parametri i obučeni su na trening skupu. Matrice konfuzije (Slika 5), koje su dobijene akumulacijom matrica iz svake od iteracija unakrsne validacije, govore koliko su klasifikatori uspešni. Ovde se može uočiti da je posmatrajući svaku vrstu najtamnije polje ono koje se nalazi na glavnoj dijagonali, a elementi na glavnoj dijagonali govore broj uspesnih klasifikacija. Ova činjenica ukazuje da dobijeni klasifikatori u nekoj meri uspešno rade posao. Leva matrica je rezultat KNN klasifikatora, desna predstavlja rezultat logističke regresije, dok srednja predstavlja rezultat RF klasifikatora. U tabeli (Tabela 2.) možemo videti vrednosti mera uspešnosti za svaki klasifikator.



Slika 5. Matrice konfuzije za obučenost na trening skupu

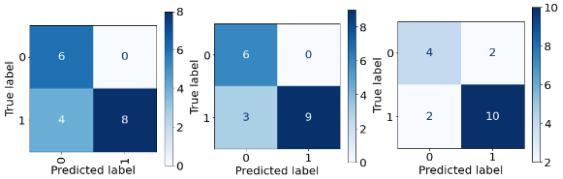
Tabela 2. Vrednosti mera posle obuke na trening skupu

	KNN	RF	LR
Precision	0.9204	0.9074	0.8774
Accuracy	0.7963	0.8827	0.8334
Recall	0.7570	0.9159	0.8692
Specificity	0.8728	0.8182	0.7636
F score	0.8308	0.9116	0.8732

Kao što se može primetiti, najveću tačnost ima KNN klasifikator. Od ukupno 180 pacijenata, pogrešno je predvideo izlečenje kod 7 pacijenata. Izgleda sitno, ali je i dalje veliki broj za tako krupnu grešku. Međutim, KNN rezultat je za skoro 50% bolji od klasifikatora linearne regresije, a za 3 pacijenta bolji od RF klasifikatora.

IX. Poređenje performansi na test skupu

Matrice konfuzije na slici (Slika 6.) predstavljaju uspešnost klasifikatora prilikom testiranja na test skupu. Raspored je isti kao i na prošloj slici, gde leva matrica predstavlja rezultat KNN klasifikatora, desna rezultat logičke regresije, a matrica u sredini rezultat RF klasifikatora.



Slika 6. Matrice konfuzije za testiranje na test skupu

Tabela 3. Vrednosti mera posle testiranja na test skupu

	KNN	RF	LR

Precision	0.9204	1.0	0.8334
Accuracy	0.7963	0.8334	0.7778
Recall	0.7570	0.7500	0.8334
Specificity	0.8728	1.0	0.6667
F score	0.8308	0.8571	0.8334

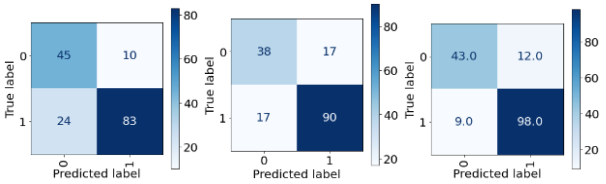
Kod rezultata za test skup je situacija drugačija. KNN klasifikator i RF nemaju ni jednu pogrešnu predikciju o izlečenosti pacijenta, dok LR klasifikator na 18 pacijenata, 2 pacijenta pogrešno klasifikuje.

X. Redukcija dimenzionalnosti

Nakon odrađene obuke i testiranja iste klasifikatore ćemo ponoviti, ali uz redukciju dimenzionalnosti. Redukcija podataka ukratko predstavlja formiranje manjeg skupa po obimu, ali sa istim rezultatima analize. Dakle, proverićemo da li otklanjanjem neki obeležja remetimo predikcije koje donose ovi klasifikatori i ako remetimo, u kolikoj meri se to dešava. Algoritam koji će se koristiti je LDA, skraćeno od *Linear Discriminant Analysis*. Koristi se kao tehnika smanjenja dimenzionalnosti obezbeđujući projekciju koja najbolje odvaja primere prema njihovoj dodeljenoj klasi.

XI. Poređenje performansi na trening skupu posle redukcije

Po prikazanim rezultatima (Slika 7.), redukcija je vidno narušila predikcije za KNN i RF klasifikatore. Međutim, kod klasifikatora logističke regresije, predviđanje se poboljšalo. Sa pogrešnih 13 predikcija se spustilo na 12.



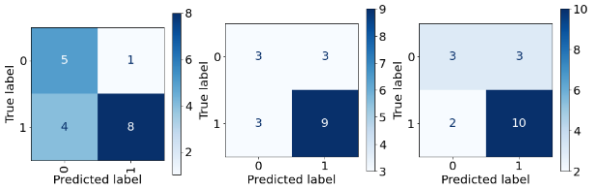
Slika 7. Matrice konfuzije za obučenost na trening skupu nakon redukcije

Tabela 4. Vrednosti mera posle obuke na trening skupu

	KNN	RF	LR
Precision	0.8925	0.8411	0.8909
Accuracy	0.7901	0.7901	0.8704
Recall	0.7757	0.8411	0.9159
Specificity	0.8182	0.6910	0.7818
F score	0.8300	0.8411	0.9032

XII. Poređenje performansi na test skupu posle redukcije

Na test skupu, posle redukcije, vidne su lošije performance kod svih klasifikatora (Slika 8.). Najveći pad je kod RF, koji se sa 0 loših predikcija izlečenja podigao na 3 loše predikcije. Pored ovih predikcija, spustio je pogodak terapija koje nisu imale efekta. KNN je doneo samo jednu lošu odluku o izlečenju pacijenta.



Slika 8. Matrice konfuzije za testiranje na test skupu nakon redukcije

Tabela 5. Vrednosti mera posle testiranja na test skupu

	KNN	RF	LR
Precision	0.8889	0.7500	0.7692
Accuracy	0.7223	0.6667	0.7223
Recall	0.6667	0.7500	0.8334
Specificity	0.8334	0.5000	0.5000
F score	0.7619	0.7500	0.8000

XIII. Zaključak

Na osnovu svega što je ispitano, uključujući analizu podataka i kasnije treniranje modela, može se reći da se ishodi terapija mogu uspešno predvideti korišćenjem metoda mašinskog učenja. Šta više, mogu biti jako korisni sa svojim predikcijama i metodama koje koriste. Što se tiče terapija, immunoterapija je pokazala bolje rezultate u lečenju pacijenata.

XIV. Reference

[An expert system for selecting wart treatment method - ScienceDirect]
[Skin Lesion Analyzer + Tensorflow.js Web App | Kaggle]
[Melanoma:DetailAnalysis,EDA,IP,Augmentation,Model | Kaggle]
[How to find the optimal value of K in KNN? | by Amey Band | Towards Data Science]
[STANDARDIZACIJA I NORMALIZACIJA PODATAKA (dokumen.tips)]
[k-nearest neighbors algorithm - Wikipedia]
[PowerPoint Presentation (ucg.ac.me)]
[LinearDiscriminantAnalysisforDimensionalityeductionin Python(machinelearningmastery.co)]
[sklearn.model_selection.cross_val_scorescikit-learn 1.0.2 documentation]
[3.3. Metrics and scoring: quantifying the quality ofpredictions scikit-learn 1.0.2 documentation]
[sklearn.linear_model.LogisticRegressionCV — scikit-learn 1.0.2 documentation]
[Literatura koja je dobijena uz skup podataka]
[MU_praktikum]