UNIVERZITET U BEOGRADU
ELEKTROTEHNIČKI FAKULTET

Primena više modela za mašinsko predviđanje za detekciju eritemo-skvamoznih oboljenja

Seminarski rad iz predmeta Računarstvo u biomedicini na master studijama

Nada Janković 3081/16
Ivan Dimitrov 3232/16

Sadržaj

1. Uvod

Koncept mašinskog učenja ponudio je veliku primenu u oblastima analize i proučavanja kolekcija podataka u biomedicini. On pruža način za prikupljanje i ekstrakciju tih podataka radi formiranja novootkivenih znanja. Najveću prednost koju donose formirana znanja je klasifikacija oboljenja, čiji je cilj određivanje klase oboljenja za bilo kog pacijenta. S obzirom na problem koji postoji prilikom diferencijalnog dijagnostikovanja oboljenja iz grupe eritemo-skvamoznih oboljenja u dermatologiji, kreiranje sistema za davanje dijagnoza vezanih za pomenute bolesti bio bi od velikog značaja.

Prilikom implementacije algoritama koji bi vršili proučavanja podataka vezanih za eritemo-skvamozna oboljenja, nameće se kao glavni zahtev preciznost predviđanja. Pored preciznosti, neophodno je da imaju što kraće vreme predviđanja i treninga zbog efikasnosti. Jedine prepreke ovim zahtevima mogu biti nedostatak informacija vezanih za određene instance i različiti nivoi važnosti pojedinačnih atributa za određene klase.

Korišćena kolekcija podataka sadrži simptome pacijenata sa poznatim dijagnozama. Za formiranje znanja su korišćena sledeća tri algoritma za mašinsko učenje: klasterizacija metodom k-srednjih vrednosti, naivni Bejzov klasifikator i algoritam k-najbližih suseda. Motivacija iza upotrebe različitih algoritama jeste proučavanje efikasnosti algoritma za određivanje dijagnoza vezanih za eritemo-skvamozna oboljenja. Pored toga, poređenje je vršeno i u odnosu na rezultate iz objavljenog rada profesora Elifa Derija Ubejla i Erdogana Dogdua [1]. U ovom radu je isključivo implementirana klasterizacija metodom k-srednjih vrednosti.

Nakon uvoda, u poglavlju dva ovog dokumenta, obrazložen je opis problema koji postoji vezan za dijagnostikovanje eritemo-skvamoznih oboljenja. Opis je izveden iz analize čestih simptoma svakog oboljenja iz navedene grupe. Na osnovu ove analize bi se mogli uočiti najvažniji simptomi za određene bolesti, ali s obzirom na sličnosti oboljenja koja se dijagnostikuju, teško je navesti težine važnosti simptoma.

Poglavlje tri služi za prikaz procesa implementacije algoritama za mašinsko učenje, kao i prethodno sređivanje korišćene kolekcije podataka. Na početku poglavlja su detaljnije obrazloženi atributi iz kolekcije podataka i broj instanci vezani za svaku klasu oboljenja. Svaki korak implementacije je naveden, gde su opisane i ključne metode i paketi za implementirane algoritme u jeziku Python. Na kraju poglavlja, izvršena je validacija dobijenih rezultata kroz analizu preciznosti predviđanja korišćenih algoritama.

U poglavlju četiri su izloženi analiza i poređenje rezultata korišćenih algoritama, kao i rezultata iz pomenutog objavljenog rada. U njemu su precizirani razlozi usled kojih je određeni algoritam imao veću preciznost i zašto je bolje njega upotrebiti za korišćenu kolekciju podataka.

U poslednjem poglavlju prikazana je rekapitulacija rada i implementiranih algoritama, kao i moguća unapređenja prilikom faze implementacije algoritama za mašinsko učenje. Navedena unapređenja se odnose isključivo na kolekciju podataka i algoritme za mašinsko učenje.

2. Opis problema

U ovom radu se proučavaju sledeća oboljenja iz grupe eritemo-skvamoznih (lat. *erythemato-squamous*) oboljenja [2]: psorijaza, seboroični dermatitis, lihen planus, *pityriasis rosea,* hronični dermatitis i *pityriasis rubra pilaris*. Ova dermatološka oboljenja imaju istu kliničku prezentaciju po pitanju eritema i deskvamacije sa manjim razlikama. Usled sličnosti simptoma, dijagnoza se teško postavlja. Nekim pacijentima se može postaviti dijagnoza isključivo pomoću kliničkih simptoma, ali je u većini slučajeva neophodna biopsija radi verifikacije. Međutim, histopatološka slika može da prikaže u početnoj fazi karakteristike jednog oboljenja, dok se zapravo radi o drugom.

Psorijazu tipično karakterišu oivičene, ružičaste ili sive naslage, koje mogu izazvati svrab i bol, promene na skalpu i poremećen proces deskvamacije. U zavisnosti od vrste psorijaze, javljaju se dodatni simptomi poput pustula tj. belih gnojnih plikova kod pustularne psorijaze [3].

Seboroični dermatitis se manifestuje u vidu jasno ograničenih crvenih pečata koji mogu biti prekriveni beličasto-žućkastim ljuspama slične kao perut i najčešće je praćen svrabom i crvenilom kože. Kao drugi tip dermatitisa, hronični dermatitis je povezan sa tipičnim simptomima svraba i suve kože [4].

Za kliničku sliku lihen planusa karakteristična je pojava malih lihenskih papula (čvrstih promena iznad ravni kože), koje su prvo okruglog, a potom poligonalnog oblika [5]. Pored papula, dodatno se pojavljuju intenzivan svrab i pozitivan Kobnerov fenomen tj. linearni raspored lihenskih papula usled češanja.

U početnoj fazi *pityriasis rosea* oboljenja pojavljuju se promene na koži u vidu jasno ograničenog eritematoznog ili ovalnog plaka lokalizovanog najčešće na bočnoj strani trupa ili na vratu [6]. Na površini plaka javlja se beličasta skvama naročito na periferiji u vidu ogrlice.

Pityriasis rubra pilaris je grupa retkih poremećaja na koži koji su predstavljeni ljuskavim papulama ili plakovima sa jasno ograničenim ivicama. Mogu da prekriju celo telo ili delove tela poput laktova i kolena [7].

Prethodno navedene karakteristike predstavljaju one koje su primarne za oboljenja pomenuta u ovom radu. Međutim, Kobnerov fenomen naveden je kod lihen planusa, ali je on takođe prisutan kod psorijaze i *pityriasis rubra pilaris*. Na ovaj način možemo primetiti da su ova oboljenja veoma povezana.

3. Implementacija algoritama za mašinsko učenje

Prilikom implementacije algoritama izvršeni su sledeći koraci:

1. Pretprocesiranje;
2. Primena algoritama mašinskog učenja i
3. Validacija dobijenih rezultata.

Procesu implementacije prethodi izbor i formiranje kolekcije podataka (engl. data set) za proučavanje. Kolekcija podataka koja je upotrebljena [2] je specifična za eritemo-skvamozna oboljenja. Ovi podaci su podeljeni za korišćenje prilikom faze treninga (75% kolekcije) i testiranja (25% kolekcije). Sadrži 34 atributa koji se odnose na atribute kliničke i histopatološke reprezentacije za 366 instanci. Kliničke atribute čine:

- eritem;
- deskvamacija;
- ograničene ivice;
- svrab;
- Kobnerov fenomen;
- poligonalne papule;
- folikularne papule;
- promene na oralnoj sluzokoži;
- promene na laktovima i kolenima;
- promene na skalpu;
- porodična istorija i
- godine pacijenta.

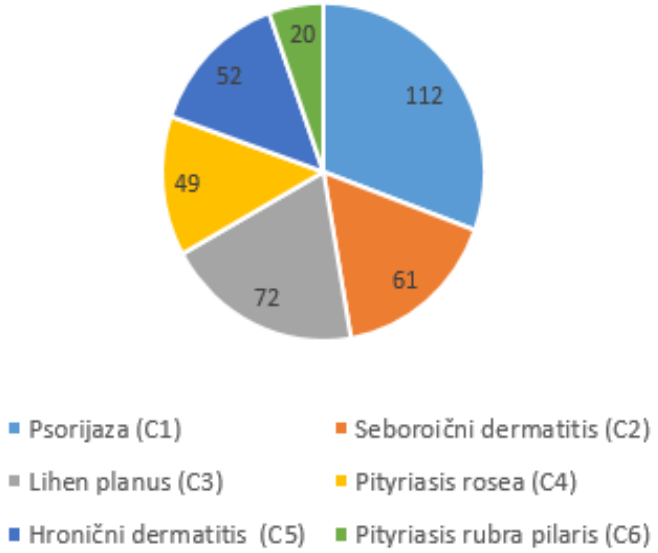
Histopatološke atribute čine:

- nedostatak melanina;
- eozinofilni infiltrat;
- eozinofilni plućni infiltrat;
- fibroza papilarnog dermisa;
- egzocitoza;
- akantoza;
- hiperkeratoza;
- parakeratoza;
- grupisanje interpapilarnih nastavaka;
- izduzeni interpapilarni (epidermalni) nastavci;
- istanjivanje suprapapilarnog epidermisa;
- sunđeraste pustule;
- „Munrov“ mikroabsces;
- fokalna hipergranuloza;
- nestajanje granularnog sloja;
- vakualizacija i oštećenje bazalnog sloja;
- spongioza;
- testerasti interpapilarni nastavci;
- folikularni čepovi;
- perifolikularna parakeratoza;
- inflamatorni mononuklearni infiltrat i
- trakasti infiltrat.

U ovoj kolekciji podataka, atribut porodične istorije ima vrednost 1 ukoliko je neka od ovih bolesti posmatrana u okviru porodice pacijenta, u suprotnom je jednaka 0. Svi

atributi izuzev godina pacijenta predstavljaju se vrednošću u opsegu od 0 do 3. Vrednost 0 predstavlja nedostatak atributa, vrednosti 1 i 2 srednju prisutnost atributa, dok vrednost 3 predstavlja najveću moguću vrednost.

Na slici 1 je predstavljen grafik sa brojem pacijenata iz korišćene kolekcije podataka čije su dijagnoze vezane za određena eritemo-skvamozna oboljenja. Uz svako oboljenje je prikazan na grafiku identifikator klase tog oboljenja (npr. psorijaza je predstavljena klasom sa identifikatorom 1). U kolekciji podataka je poslednjom kolonom naznačena dijagnoza svakog pacijenta, na osnovu čega je izračunat prikazan broj instanci za svako oboljenje. S obzirom da je za oboljenje *pityriasis rubra pilaris* vezano samo 20 instanci, ono je izbačeno iz razmatranja, što je objašnjeno u poglavlju 3.1.2.



Slika 1: Broj pacijenata vezanih za određena oboljenja

3.1. Pretprocesiranje

Pre početka učenja je potrebno analizirati ulazne podatke i odraditi početnu obradu nad njima, kako bi oni bili pogodniji za korišćenje u svrhu mašinskog učenja. Obrade koje su primenjene su:

1. Normalizacija podataka,
2. Rukovanje vrednostima koje nedostaju i
3. Biranje atributa.

3.1.1. Normalizacija podataka

Normalizacija podataka je neophodna kod većine algoritama mašinskog učenja. Ona podrazumeva da se vrednosti svih atributa svedu na isti opseg, koji se tipično nalazi u rangu vrednosti od 0.0 do 1.0. Normalizacijom se postiže ravnopravnost između svih atributa iz kolekcije podataka i na taj način svaki atribut podjednako utiče na krajnji rezultat. Kako sama priroda atributa ne bi uticala na njegovu važnost prilikom upotrebe, neophodno je da se izvrši obrada nad njim. U našem slučaju se može uočiti kao primer atribut „Godine pacijenta“, koji predstavlja numeričku vrednost čiji opseg može preći opseg ostalih atributa koji uobičajeno predstavlja vrednosti [0-3]. Bez upotrebe normalizacije, algoritam za učenje bi visoke vrednosti ovog atributa tumačio

kao dominantne prilikom kalkulacija. Posledica neželjene dominantnosti atributa može dovesti do loših rezultata.

Za normalizaciju je korišćen *MinMaxScaler* iz *python* paketa *sklearn.preprocessing*, koji predstavlja Min-Max normalizaciju [8]. Na ovaj način se vrši transformacija svakog atributa skaliranjem vrednosti u opsegu [0,1].

3.1.2. Rukovanje vrednostima koje nedostaju

Rukovanje vrednostima koje nedostaju je veoma komplikovana tema jer postoji više načina na koji se mogu zameniti vrednosti koje nedostaju u kolekciji podataka. Međutim, u našem slučaju je broj ovakvih vrednosti relativno mali (8) i radi se o vrednostima samo jednog atributa - „Godine pacijenta“. Iz tih razloga su instance koje sadrže atribut nepoznate vrednosti zanemarene i izbačene iz kolekcije podataka.

Pored navedenih pretprocesnih obrada urađena je još jedna obrada, koja je izvršena zbog referenci iz eksternih radova. Naime, iz analiza korišćene kolekcije podataka je utvrđeno je da je broj primeraka poslednje klase oboljenja „*pityriasis rubra pilaris*“ isuviše mali da bi algoritam mogao na osnovu njega da donosi relevantne zaključke. Zbog toga su instance ove klase izbačene iz inicijalnih podataka.

3.1.3. Biranje atributa

U fazi pretprocesiranja treba odrediti koji će atributi biti korišćeni, odnosno da li će radi smanjivanja šuma neki atributi biti isključeni iz faze učenja. Da bi se odredilo koliko je neki atribut značajan, u kolekciji podataka se koristi koeficijent korelacije posmatranog atributa i atributa koji prikazuje kojoj klasi instanci pripada. U tabeli 1 je dat prikaz korelacije svakog od ulaznih atributa.

Tabela 1: Prikaz koeficijenta korelacije svakog od ulaznih atributa

Naziv atributa	**Koeficijent korelacije**
eritem	-0.37
deskvamacija	-0.52
ograničene ivice	-0.36
svrab	0.15
kobnerov fenomen	-0.01
poligonalne papule	0.13
folikularne papule	0.17
promene na oralnoj sluzokoži	0.13
promene na laktovima i kolenima	-0.62
promene na skalpu	-0.62
porodična istorija	-0.35
godine pacijenta	-0.063
nedostatak melanina	0.14
eozinofilni infiltrat	-0.02

Naziv atributa	Koeficijent korelacije
eozinofilni plućni infiltrat	-0.56
fibroza papilarnog dermisa	0.65
egzocitoza	0.31
akantoza	-0.06
hiperkeratoza	-0.14
parakeratoza	-0.50
grupisanje interpapilarnih nastavaka	-0.71
izduzeni interpapilarni nastavci	-0.35
istanjivanje suprapapilarnog epidermisa	-0.70
sunđeraste pustule	-0.47
„Munrov“ mikroabsces	-0.53
fokalna hipergranuloza	0.14
nestajanje granularnog sloja	-0.42
vakualizacija i oštećenje bazalnog sloja	0.14
spongioza	0.20
testerasti interpapilarni nastavci	0.14
folikularni čepovi	0.05
perifolikularna parakeratoza	-0.02
inflamatorni mononuklearni infiltrat	0.01
trakasti infiltrat	0.14

Međutim, izbacivanje atributa iz kolekcije podataka sa najmanjom korelacijom nije donelo poboljšanje na preciznosti, čak je donelo i pogoršanja. Jedno objašnjenje za ovakvo ponašanje je što je i sam raspored korelacija dosta ravnomeran i veoma mali broj atributa ima zanemarljivo male korelacije.

Takođe treba dodati da je korelacija jedan metod koji može pomoći, ali ona ne može biti korišćena za određivanje konačnih uzrok-posledica veza u kolekciji podataka.

3.2. Primena algoritama mašinskog učenja

U okviru faze primene algoritama mašinskog učenja su upotrebljena sledeća četiri algoritma:

1. Klasterizacija metodom k-srednjih vrednosti (engl. k-means clustering),
2. Naivni Bejzov klasifikator (engl. Naive Bayes classifier),
3. Algoritam k-najbližih suseda (engl. k-Nearest Neighbors algorithm) i
4. Hibridni algoritam na bazi glasanja.

U nastavku su detaljnije objašnjeni pomenuti algoritmi i način upotrebe istih.

4.2.1. Klasterizacija metodom k-srednjih vrednosti

Klasterizacija metodom k-srednjih vrednosti je jedan od najpoznatijih i najjednostavnijih algoritama klasterizacije. Njena prednost leži u pogodnosti za primenu nad problemima u kojim je potrebno izvršiti grupisanje instanci po sličnosti. Ovaj metod predstavlja particionisanje datih objekata u k klastera u kojem svaki objekat pripada klasteru sa najbližom srednjom vrednošću. Ulazne podatke algoritma čine:

- k – broj klastera (definisan je vrednošću 5 usled postavljanja dijagnoze za 5 oboljenja) i
- skup za trening sa određenim brojem instanci, gde svaka instanca sadrži odgovarajući skup atributa (definisan je sa 253 instance sa po 34 atributa).

Algoritam počinje nasumičnim izborom k instanci iz skupa za trening koja predstavljaju centroide. Dok algoritam ne konvergira tj. dok se barem jedna instanca pomera iz jednog u novi klaster, vrši se grupisanje instanci po klasterima identifikovanjem najbližeg centroida za određenu instancu i definisanje novih centroida usrednjavanjem svih instanci u okviru klastera.

4.2.2. Naivni Bejzov klasifikator

Naivni Bejzov klasifikator se naziva „naivnim“ jer posmatra sve atribute instance nezavisno tj. zanemaruje korelacije između atributa. Njegova prednost je što brzo predviđa rezultat. Kao i klasterizacija metodom k-srednjih vrednosti opisana u poglavlju 3.2.1, ovaj klasifikator ima za ulazni podatak skup za trening sa instancama.

Ovaj klasifikator je zasnovan na Bejzovoj teoremi (engl. Bayes' theorem), koja pruža način da se izračuna verovatnoća P(c|x) koristeći verovatnoće P(c), P(x) i P(x|c) na sledeći način:

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

gde su P(x) i P(c) verovatnoće slučajeva x i c, a P(x|c) verovatnoća slučaja x pod uslovom da se c dogodi. Ova formula se izračunava za svaku klasu c i klasa sa najvećom verovatnoćom predstavlja ishod predikcije.

Prilikom faze implementacije klasifikatora, korišćen je tip naivnog Bejzovog modela sa Gausovom tj. normalnom distribucijom za klasifikaciju iz *python* paketa *sklearn.naive_bayes* [8].

4.2.3. Algoritam k-najbližih suseda

Algoritam k-najbližih suseda se zasniva na upoređivanju sa najsličnijim instancama u trening skupu. Ulazne podatke algoritma čine:

- k – broj suseda koje razmatra (definisan je vrednošću 10) i
- skup za trening sa određenim brojem instanci, gde svaka instanca sadrži odgovarajući skup atributa (definisan je sa 253 instance sa po 34 atributa).

Nova instanca će pripadati klasi kojoj pripada najbliži sused iz trening skupa. S obzirom da svi atributi imaju linearne vrednosti, distanca između dve instance meri se korišćenjem euklidskog rastojanja. Ovo rastojanje definiše se sledećom formulom:

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

gde vrednosti x i y predstavljaju vrednosti atributa instanci. Kako bi se zaobišla značajna razlika u opsezima za atribut „godine pacijenta“, neophodno je da ovom algoritmu prethodi normalizacija opisana u poglavlju 3.1.

Prilikom faze implementacije algoritama, korišćena je varijanta algoritma gde težina suseda ne zavisi od distance od posmatrane instance. Na ovaj način, veća težina neće biti data bližim susedima.

4.2.4. Hibridni algoritam na bazi glasanja

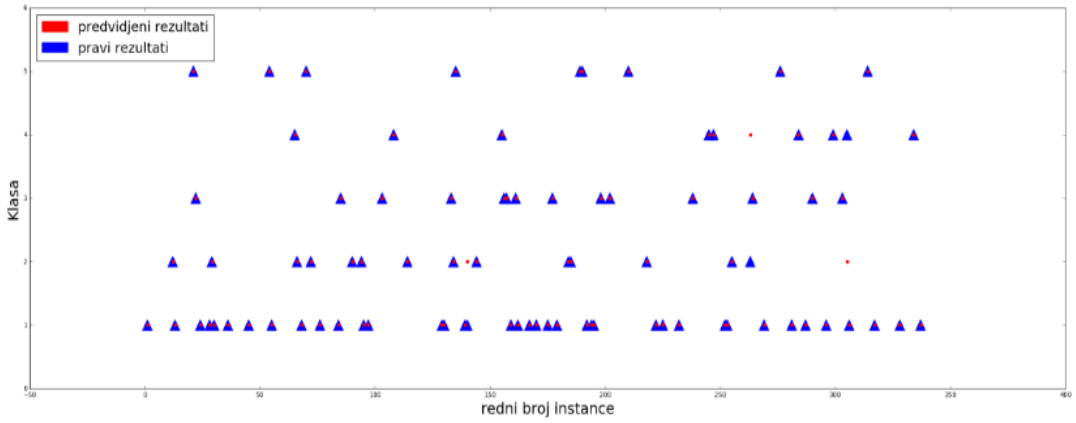
Ovaj klasifikator predstavlja kombinaciju prethodna tri algoritma, gde svaki algoritam trenira sa istim „trening setom“ podataka. Kada se radi predikcija, svi algoritmi rade u paraleli i na kraju svaki da svoj rezultat, a konačan rezultat se bira po težini glasa svakog algoritma.

Osnovni parametar prilikom rada ovog algoritma su težine koje se dodeljuju svakom pojedinačnom algoritmu. U našem primeru smo uradili merenja sa dve varijante težina: 1-1-1 i 1-2-1, gde smo prednost dali algoritmu k-najbližih suseda. Rezultati koje smo dobili kod obe postavke se nisu značajno razlikovali, a bili su u rangu rezultata algoritma k-najbližih suseda.

3.3. Validacija dobijenih rezultata

Prilikom faze validacije dobijenih rezultata, vrši se komparacija dobijenih rezultata sa pravim rezultatima koji se nalaze u kolekciji podataka. Za ulazne podatke modela koji je upotrebljen za predikciju, definisano je 25 posto instanci iz kolekcije podataka i vrši se navedeno poređenje rezultata.

Jedan primer rezultata koji je predvideo model za našu kolekciju podataka vezan za algoritam k-najbližih suseda se nalazi na slici 2. Na mestima gde se poklapaju plavi trougao i crveni krug je tačno predviđanje, dok je prikazan promašaj na mestima gde se nalazi isključivo crveni krug. Za ovaj primer je preciznost predviđanja jednaka 96.5%.



Slika 2: Primer raspodele rezultata predikcije nad instancama za testiranje (klasifikator k-najbližih suseda)

4. Analiza i poređenje rezultata

Rezultati preciznosti koji su dobijeni kada je primenjen model klasterizacije metodom k-srednjih vrednosti su u približni vrednostima koje su dobijene u radu *"Automatic Detection of Erythemato-Squamous Diseases Using k-Means Clustering"* [1]. U navedenom radu, kreiran je sistem koji isključivo sadrži primenu algoritma klasterizacije metodom k-srednjih vrednosti i njihova preciznost predviđanja iznosi 94.22%, gde je najveća stopa greške vezana za seboroični dermatitis. Međutim, smatra se da je ta stopa greške viša u odnosu na druga oboljenja usled nedostatka podataka tj. instanci u kolekciji podataka.

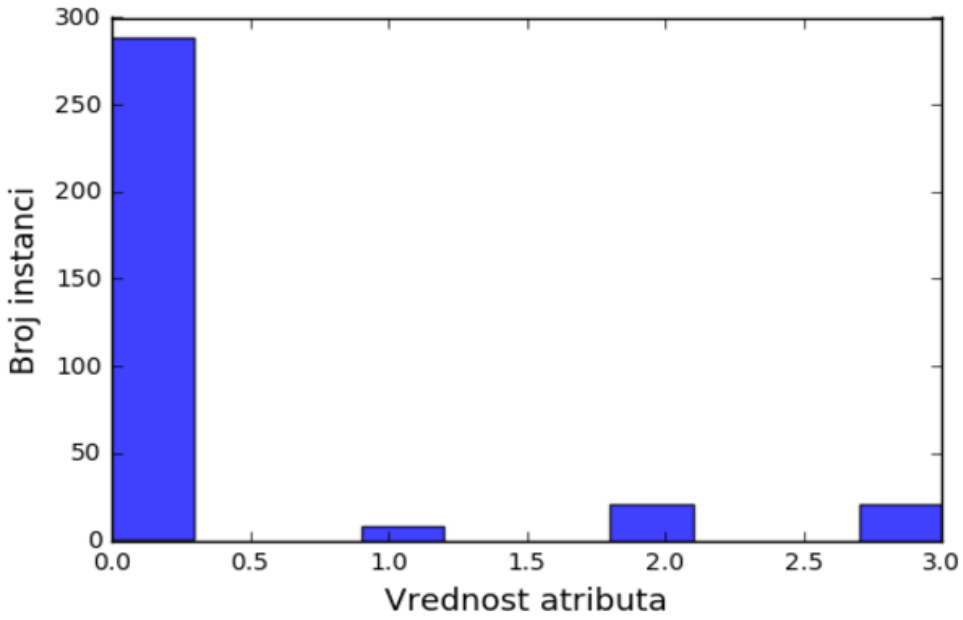
Pored pomenute preciznosti predviđanja, značajni su pojmovi senzitivnosti i specifičnosti. Senzitivnost se odnosi na sposobnost algoritma da identifikuje one kod kojih se dijagnoza sastoji od određenog oboljenja. Izračunava se kao odnos tačno pozitivnih od ukupnog broja obolelih pacijenata. Specifičnost se odnosi na sposobnost algoritma da identifikuje one kod kojih dijagnoza ne sadrži određeno oboljenje. Izračunava se kao odnos tačno negativnih od ukupnog broja pacijenata koji nisu oboleli. Tačno pozitivni rezultati se odnose na slučaj kada algoritam ima pozitivno predviđanje, tj. kada potvrdi dijagnozu iz kolekcije podataka, dok se tačno negativni rezultati odnose na slučaj kada algoritam potvrdi nepostojanje oboljenja koje je navedeno u kolekciji podataka. Vrednosti za preciznosti predviđanja, senzitivnosti i specifičnosti koje smo dobili nakon implementacije su prikazane u tabeli 2.

Tabela 2: Vrednosti parametara rezultata korišćenih algoritama

Algoritam za mašinsko učenje	**Klase**	**Specifičnost**	**Senzitivnost**	**Preciznost predviđanja**
Klasterizacija metodom k-srednjih vrednosti	C1	100 %	100 %	85.75 %
	C2	100 %	50 %	
	C3	100 %	100 %	
	C4	87.65 %	100 %	
	C5	100 %	100 %	
Algoritam k-najbližih suseda	C1	100 %	100 %	98.8 %
	C2	98.65 %	81.82 %	
	C3	100 %	100 %	

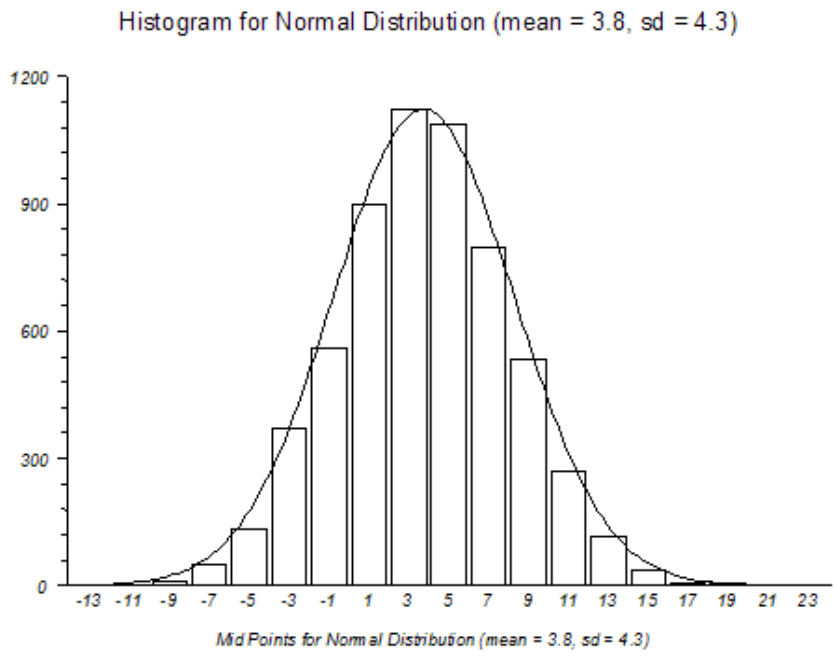
Algoritam za mašinsko učenje	Klase	Specifičnost	Senzitivnost	Preciznost predviđanja
	C4	87.22 %	92.31 %	
	C5	100 %	100 %	
Naivni Bejzov klasifikator	C1	100 %	100 %	84.71 %
	C2	91.46 %	100 %	
	C3	100 %	100 %	
	C4	100 %	70 %	
	C5	100 %	91 %	

Sličnu preciznost smo dobili kod modela koji primenjuje naivni Bejzov algoritam. Kod ovog modela je primećeno da je preciznost u zavisnosti od date kolekcije podataka za treniranje. Razlog za navedenu činjenicu je zato što ovaj model mašinskog učenja ne odgovara našem slučaju, a to je prvenstveno zbog distribucije vrednosti u atributima kolekcije podataka. Iako u našem slučaju postoji 34 atributa, ne postoji značajna razlika između distribucije atributa. Ilustracija na kojoj je prikazana distribucija jednog od atributa se nalazi na slici 3.



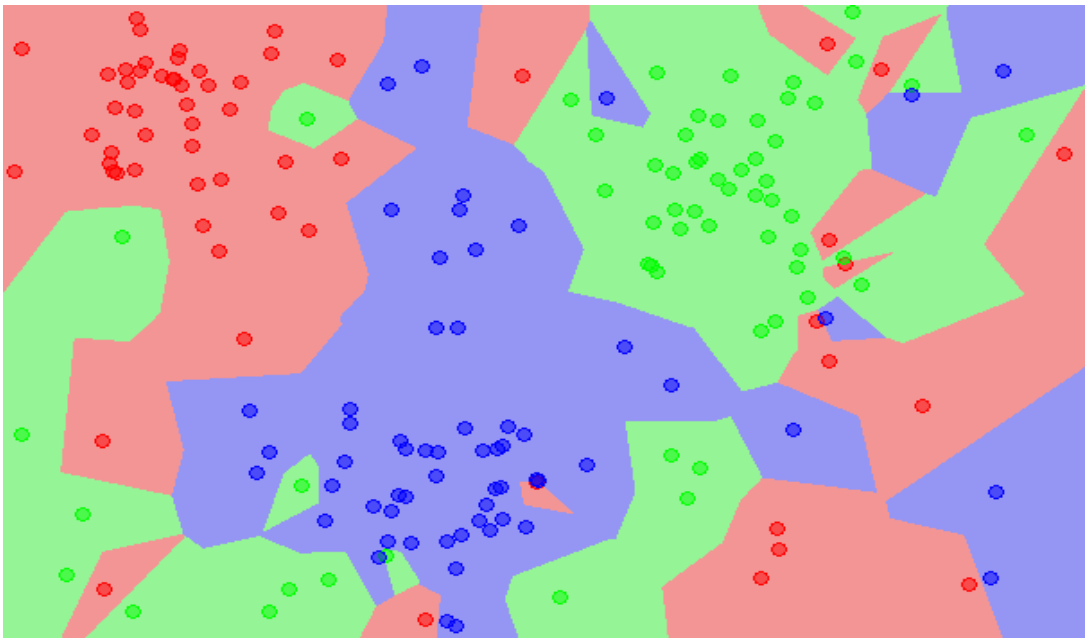
Slika 3: Distribucija jednog od atributa iz kolekcije podataka

Ovaj podatak je za naivni Bejzov algoritam od velikog značaja, jer se sva izračunavanja vezana za ovaj algoritam svode na verovatnoću. Pored toga, Gausova verzija Bejzovog algoritma uzima za pretpostavku da je distribucija vrednosti atributa normalna (prikazano na slici 4).



Slika 4: Normalna distribucija

Međutim, najbolju preciznost smo dobili primenom modela k-najbližih suseda. Iako su algoritam klasterizacije metodom k-srednjih vrednosti i algoritam k-najbližih suseda slični po tome što koriste metriku razdaljine pojedinih instanci kao glavnu metriku pri određivanju predviđanja, ova dva algoritma imaju jednu bitnu razliku koja je ilustrovana grafovima na slikama 5 i 6.



Slika 5: Algoritam k-najbližih suseda – primer raspodele po klasama

Slika 6: Algoritam klasterizacije metodom k-srednjih vrednosti – primer raspodele po klasama

Kao što se može primetiti po grafovima na slikama 5 i 6, kod algoritma k-najbližih suseda klasifikacija je lokalna, dok je kod algoritma klasterizacije metodom k-srednjih vrednosti klasifikacija globalna zbog načina funkcionisanja algoritma koji je opisan u prethodnom poglavlju. Kod algoritma k-najbližih suseda može jedna klasa da bude na različitim delovima opsega atributa, dok kod algoritma klasterizacije metodom k-srednjih vrednosti sve instance jedne klase moraju biti striktno grupisane oko jednog centra (centroida). Zbog ove osobine algoritma klasterizacije metodom k-srednjih vrednosti se stvara razlika u preciznosti, jer su podaci sa kojima mi radimo pogodniji za lokalnu klasifikaciju.

Prilikom rada sa hibridnom metodom glasanja koja je pomenuta u poglavlju 3.2.4, primetili smo da dobijeni rezultati nisu doneli nikakva poboljšanja u odnosu na već definisani najbolje kvalifikovani algoritam.

U tabelama 3, 4 i 5 su prikazane matrice konfuzije za naša tri algoritma. U okviru njih se prikazuje odnos predviđenih i dobijenih rezultata za svih pet klasa oboljenja.

Tabela 3: Matrica konfuzije za algoritam klasterizacije metodom k-srednjih vrednosti

		Predviđeni				
		C1	C2	C3	C4	C5
Dobijeni	C1	32	0	0	0	0
	C2	0	10	0	0	0
	C3	0	0	19	0	0
	C4	0	10	0	4	0
	C5	0	0	0	0	10

Tabela 4: Matrica konfuzije za algoritam k-najbližih suseda

Dobijeni		Predviđeni				
		C1	C2	C3	C4	C5
	C1	32	0	0	0	0
	C2	0	9	0	1	0
	C3	0	0	19	0	0
	C4	0	2	0	12	0
	C5	0	0	0	0	10

Tabela 5: Matrica konfuzije za naivni Bejzov algoritam

Dobijeni		Predviđeni				
		C1	C2	C3	C4	C5
	C1	32	0	0	0	0
	C2	0	3	0	6	1
	C3	0	0	19	0	0
	C4	0	0	0	14	0
	C5	0	0	0	0	10

Iz tabela 3, 4 i 5 možemo da vršimo poređenja kod kojih klasa je svaki algoritam pravio greške. Očigledno je da svi algoritmi prave slične greške prilikom odabira predviđanja za oboljenja klasa 2 i 4 (seboroični dermatitis i *pityriasis rosea*). Zbog ove osobine se ovi algoritmi loše međusobno dopunjuju, jer postoji velika mogućnost da za instancu na kojoj je jedan klasifikator napravio grešku da i jedan od preostala dva napravi za istu instancu grešku.

5. Zaključak

Izvršena analiza koja je predstavljena ovim dokumentom ima za cilj da prikaže koncept korišćenja algoritama mašinskog učenja zarad detekcije oboljenja iz grupe eritemo-skvamoznih oboljenja. Motivacija je bila da se odredi koji algoritam ima najveću preciznost predviđanja za datu kolekciju podataka. Odabrana je grupa oboljenja koja je slična po simptomima i koja se slično manifestuje u različitim stadijumima oboljenja. Zbog ovih razloga, otežano je davanje dijagnoze.

Prilikom implementacije je bilo neophodno da se izvrši podela procesa na tri faze, kako bi se izvršilo efikasnije predviđanje. U prvoj fazi je izvršeno „čišćenje“ date kolekcije podataka da bi mogla da se koristi kao ulaz naredne faze. Sledeća faza obuhvatala je implementaciju algoritama mašinskog učenja, odnosno algoritma klasterizacije metodom k-srednjih vrednosti, naivnog Bejzovog algoritma i algoritma k-najbližih suseda. Svaki algoritam je implementiran korišćenjem odgovarajućeg paketa iz programskog jezika *python*. U poslednjoj fazi je bila izvršena validacija dobijenih rezultata u odnosu na rezultate iz kolekcije podataka. Pored toga, izvršena je analiza u odnosu na eksterni rad i rezultat vezan za algoritam klasterizacije metodom k-srednjih vrednosti.

Rezultat je najefikasniji algoritam za datu kolekciju podataka, tj. najbolja preciznost koja je dobijena primenom modela k-najbližih suseda. Kod ovog algoritma je uočeno da je klasifikacija lokalna, usled čega je ovaj algoritam najefikasniji s obzirom da su korišćeni podaci pogodniji za lokalnu klasifikaciju.

Razmatrane su moguće promene nakon procesa implementacije koje bi dodatno unapredile korišćene algoritme. Pošto su razlike između klasa pod rednim brojem 2 i 4, tj. oboljenja seboroičnog dermatitisa i *pityriasis rosea* male, razvijena je ideja o kreiranju modela koji bi mogao da razlikuje ove klasifikacije. Ideja leži u formiranju hibridnog klasifikatora koji bi na osnovu predviđenog rezultata davao različite dinamičke kvote za svaki od svojih modela. Ukoliko su korišćena dva algoritma za mašinsko učenje i oba rade u isto vreme, može se u toku faze treninga utvrditi za koji slučaj je koji algoritam bolji, odnosno koji algoritam se kada primenjuje. Na osnovu toga se može za određene rezultate utvrditi koji se algoritam koristi za predviđanje.

Ovaj rad se može iskoristiti za dalje proučavanje i analizu upotrebe algoritama mašinskog učenja vezanih za eritemo-skvamozna i druga oboljenja, kao što je naša motivacija bila rad profesora Elifa Derija Ubejla i Erdogana Dogdua.

Literatura

[1] Übeyli, Elif Derya i Erdoğan Doğdu. „Automatic detection of erythemato-squamous diseases using k-means clustering“. Journal of medical systems 34.2 (2010): 179-184.

[2] Nilsel Ilter i H. Altay Guvenir. „Dermatology Database“. Dostupno na: http://archive.ics.uci.edu/ml/datasets/Dermatology. Poslednji put pristupano 08. januara, 2017.

[3] Radoš D. Zečević, Lidija Kandolf Sekulić, Željko P. Mijušković i Danilo Vojvodić. „Savremena terapija psorijaze“. Medija centar Odbrana, Beograd, Srbija, 2013.

[4] Mayo Clinic. „*Dermatitis*“. Dostupno na: http://www.mayoclinic.org/diseases-conditions/dermatitis-eczema/symptoms-causes/dxc-20204412. Poslednji put pristupano 29. decembra, 2016.

[5] Mayo Clinic. „*Lichen planus*“. Dostupno na: http://www.mayoclinic.org/diseases-conditions/lichen-planus/symptoms-causes/dxc-20188521. Poslednji put pristupano 29. decembra, 2016.

[6] Mayo Clinic. „*Pityriasis rosea*“. Dostupno na: http://www.mayoclinic.org/diseases-conditions/pityriasis-rosea/basics/definition/con-20028446. Poslednji put pristupano 29. decembra, 2016.

[7] Medscape. „*Pityriasis rubra pilaris*“. Dostupno na: http://reference.medscape.com/article/1107742-clinical. Poslednji put pristupano 29. decembra, 2016.

[8] Python Documentation. Dostupno na: https://www.python.org/doc/. Poslednji put pristupano 04. januara, 2017.