Klasifikacija recepata primenom kNN i SVM klasifikatora

Jovana Bogojević, IN32/2017, bogojevic.in32.2017@uns.ac.rs

I. UVOD

U ovom izveštaju predstavljeno je rešavanje klasifikacionog problema kao problema nadgledanog učenja, kod kog su izlazne promenljive za dostupne uzorke oznake za pripadnost uzorka određenoj klasi, odnosno, kategoriji.

Rezultat je fomiran model klasifikacije koji ima zadatak da utvrdi klasnu pripadnost za uzorke za koje ona nije poznata. U nastavku izveštaja biće navedene klasne labele problema kog se rešava.

II. OPIS BAZE PODATAKA

Baza obuhvata podatke koji predstavljaju sastojke za tri različite poslastice – kolačić, picu ili pecivo. Na raspolaganju je 1738 uzoraka iz skupa za obuku i 193 uzorka iz test skupa, gde svaki uzorak predstavlja jedan recept za jednu od prethodno navedenih poslastica. Dato je ukupno 133 različitih obeležja odnosno sastojaka. Vrednosti obeležja su neke od uobičajenih sastojaka poput: sećera, brašna, mlečnih proizvoda, raznih vrsta mesa, povrća kao i začina.

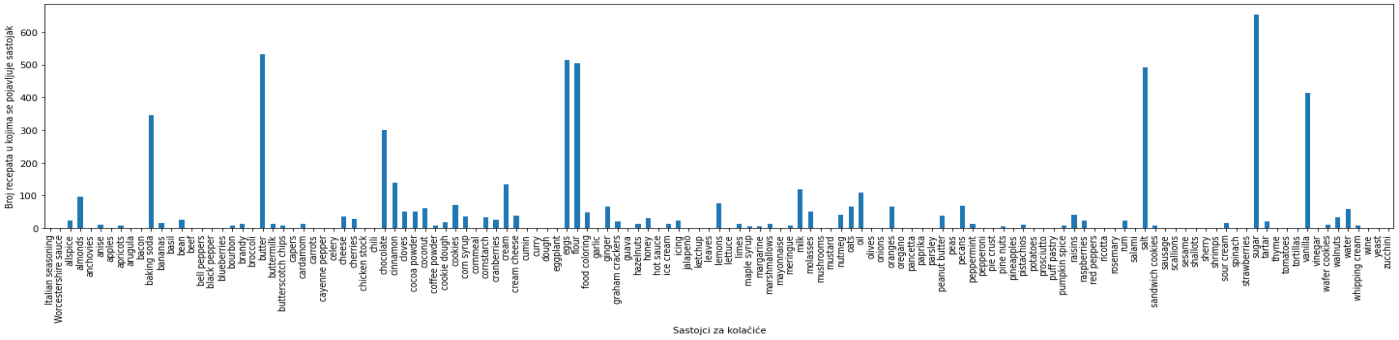
Ukoliko se neki od sastojaka pojavljuje u receptu, vrednost tog obeležja je jedan (1), u suprotnom nula (0).

U bazi ne postoje nedostajuće vrednosti. Sva obeležja imaju numeričke vrednosti, osim oznaka za klasne labele koje su kategoričke. Kategoričke vrednosti klasnih labela biće pretvorene u numeričke po principu:

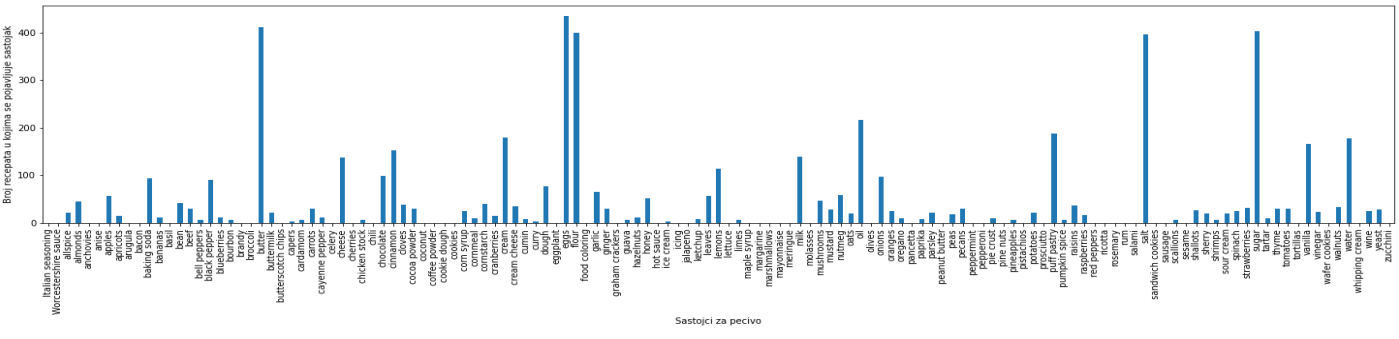
0	cookies (kolačić)
1	pastries(pecivo)
2	pizzas(pica)

Najviše uzoraka je iz klase *cookies*, ukupno 723, sledi 619 uzoraka iz klase *pastries*, a 369 uzoraka pripada klasi *pizzas*.

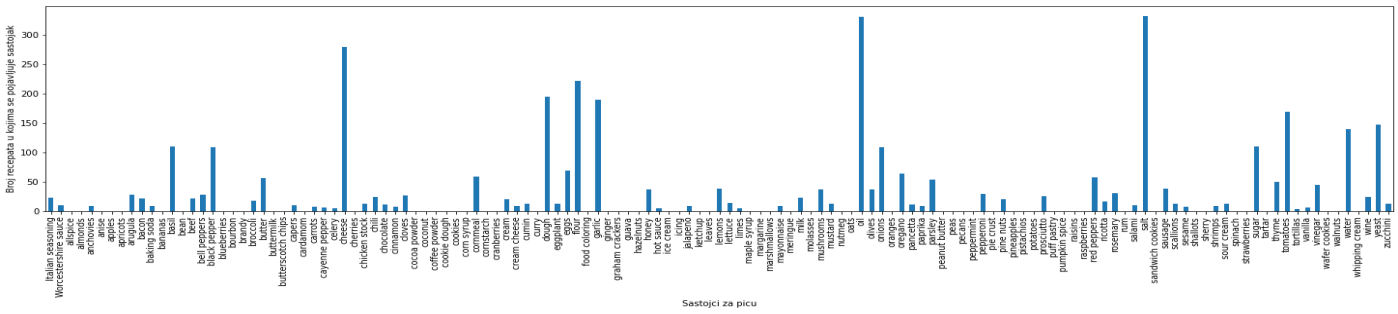
U cilju boljeg sagledavanja prisustva (1) odnosno odsustva (0) sastojaka za različite klase, formirani su karakteristični obrasci u vidu histograma pojavljivanja sastojaka iz kojih se mogu uvideti sličnosti ili razlike među klasama.



SLIKA 1: Histogram pojavljivanja sastojaka u klasi *cookies*



SLIKA 2: Histogram pojavljivanja sastojaka u klasi *pastries*



SLIKA 3: Histogram pojavljivanja sastojaka u klasi *pizzas*

Može se primetiti da su kod kolačića najčešći sastojci mlečni proizvodi, jaja, šećer, so, vanila i soda bikarbona. Na *Slici 3*. prikazan je histogram pojavljivanja sastojaka za picu odakle se može videti da su najučestaliji sastojci: brašno, jaja, ulje, so, sir kao i razno povrće poput maslina, crvene paprike, paradajza, belog luka…U receptima za picu se češće pojavljuje meso u raznim oblicima u odnosu na ostale klase.

Kada su u pitanju sastojci koji se pojavljuju u receptima za peciva, može se primetiti da se često pojavljuju sastojci koji su bili primetni i u receptima za picu i kolačiće poput brašna, jaja, ulja, soli, šećera…Ova činjenica bi mogla otežati klasifikaciju, gde bi se moglo desiti da se na osnovu sličnih sastojaka pogrešno klasifikuje recept za pecivo kao recept za kolačiće ili picu, i obrnuto.

III. kNN KLASIFIKATOR

Klasifikacija recepata upotrebom kNN (engl. *k-nearest neighbours*) algoritma podrazumeva da se odluka o pripadnosti određenog recepta nekoj od klasa donosi na osnovu klasne pripadnosti najbližih suseda iz skupa za obuku.

Elementi koji su neophodni da bi se primenio kNN metod su skup uzoraka za obuku, celobrojni parametar k i metrika. Parametar k označava broj najbližih suseda koji se uzima u obzir pri odlučivanju, a na osnovu odabrane metrike meri se udaljenost između neobeleženog uzorka , za koji nije poznata klasna labela, i svakog uzorka iz skupa za obuku. Neophodno je da se odaberu optimalne vrednosti za metriku i parametar k kako bi se dobilo što bolje ponašanje klasifikatora.

Da bi se pronašle najbolje vrednosti za parametar k i metriku, neophodno je da se klasifikator obuči na trening skupu za različit broj suseda kao i za različite metrike, te da se izvrši procena koja kombinacija odabranih vrednosti daje najveću tačnost klasifikatora. S obzirom na to da su vrednosti obeležja binarne, u obzir se uzima Žakarovo („*jaccard"*) rastojanje, zatim Hamingovo („*hamming*") rastojanje koje je čest izbor za obeležja sa celobrojnim vrednostima kao i za binarna obeležja, kao i Eukldisko („*eucledean*") rastojanje.

Pored parametra k i metrike, posmatran je i parametar *weights* koji služi za dodeljivanje značaja svakom od k najbližih suseda u odlučivanju. Ukoliko svi susedi imaju jednak značaj, onda se vrednost ovog parametra postavlja na „*uniform",* a ukoliko je značaj suseda obrnuto proporcionalan njihovoj udaljenosti od neobeleženog uzorka, onda se vrednost parametra postavlja na „*distance*".

Da bi se prevazišao problem korištenja nekih uzoraka samo u svrhu testiranja, koristi se metod unakrsne validacije. Dostupan skup uzoraka za obuku podeljen je na 10 podskupova. Metod unakrsne validacije podrazumevaće da se proces obuke i testiranja kNN klasifikatora ponavljata 10 puta tako što će u svakoj iteraciji 9 podskupova koristiti za obuku modela, a testiranje se vrši na preostalom podskupu, odnosnom na validacionom skupu.

Primenom unakrsne validacije za svaku kombinaciju vrednosti parametara *k*, *weights* i *metric*, najbolje performanse kNN klasifikatora, posmatrajući meru tačnosti klasifikatora, postižu se Hamnigovo rastojanje, parametar k = 8 i za parametar *weights* postavljen na „*distance*".

Da bi se procenile performanse klasifikatora na skupu za obuku neophodno je uporediti predviđene i klasne labele, a u tu svrhu koristi se matrica konfuzije. Na osnovu finalne matrice konfuzije, koja se dobija akumulacijom matrica iz svake od 10 iteracija unakrsne validacije, računaju se mere uspešnosti klasifikatora odnosno tačnost, preciznost, osetljivost, specifičnost i F-mera. S obzirom na to da u bazi postoje tri klase recepata, neophodno je odrediti mere uspešnosti za svaku klasu ponaosob i odrediti prosečnu tačnost klasifikatora.

Najbolje performanse klasifikatora su za klasu recepata za picu gde tačnost, odnosno udeo ispravno klasifikovanih uzoraka, iznosi 0,981. Performanse klasifikatora za klasu kolačića su takođe korektne. Najslabije mere uspešnosti klasifikatora su za klasu recepata za peciva, gde se pravi najviše grešaka – 79 recepata iz klase peciva pogrešno je klasifikovano kao recept za kolačiće. Greške prilikom klasifikovanja su verovatno posledica prisustva istih sastojaka u receptima i za peciva i za kolačiće. Kao konačna mera uspešnosti klasifikatora računa se prosečna tačnost, kao prosek tačnosti po klasama. U skladu sa tim, izračunata

prosečna tačnost klasifikatora iznosi 0.947, što predstavlja dobar rezultat.

	Kolačići	Pecivo	Pica
Preciznost	0.889	0.934	0.969
Tačnost	0.935	0.927	0.981
Senzitivnost	0.889	0.935	0.968
Specifičnost	0.914	0.967	0.991
F-mera	0.889	0.935	0.969

SLIKA 4: Mere uspešnosti kNN klasifikatora po klasama na trening skupu

S obzirom na to da se klasifikator do sada obučavao na skupu za obuku, koji je podeljen na trening skup i validacioni skup, neophodno je obučiti klasifikator sa konačno odabranim parametrima na čitavom skupu za obuku, zatim testirati na test skupu.

	Kolačići	Pecivo	Pica
Preciznost	0.879	0.983	1
Tačnost	0.943	0.953	0.979
Senzitivnost	0.879	0.983	1
Specifičnost	0.902	0.991	1
F-mera	0.879	0.938	1

SLIKA 5: Mere uspešnosti kNN klasifikatora za svaku klasu koji je testiran na skupu za testiranje

Prosečna tačnost klasifikatora koji je obučen na celokupnom skupu za obuku, a testiran na test skupu iznosi 0.958. Uočeno je poboljšanje performansi klasifikatora za klasu peciva, gde je dosta smanjen broj pogrešno klasifikovanih uzoraka, ali mere uspešnosti su i dalje najbolje za klasu recepata za picu.

IV. SVM KLASIFIKATOR

SVM (engl. *Support Vector Machine*) klasifikator, odnosno klasifikator „mašina na bazi vektora nosača", prvenstveno se koristi za rešavanje problema binarne klasifikacije pronalaženjem ravni koja razdvaja podatke koji pripadaju različitim klasama u prostoru obeležja. Mašina na bazi vektora nosača zasniva se na *klasifikatoru maksimalne margine*, čiji je cilj da odredi ravan koja će na optimalan način podeliti prostor obeležja na dva dela tako da se u jednom delu nađu uzorci iz jedne, a u drugom delu uzorci iz druge klase pod uslovom da su klase linearno separabilne.

Kod velikog broja klasifikacionih problema uzorci nisu linearno separabilni i u tom slučaju se dozvoljava da pojedini uzorci za obuku budu sa pogrešne strane ravni koja razdvaja uzorke različitih klasa.

Da bi se rešio problem klasifikacije recepata primenom SVM klasifikatora, neophodno je odrediti optimalne vrednosti parametara klasifikatora, a to su: regularizacioni parametar *C*, kernel i odgovarajući pristup SVM klasifikatora, a nakon toga izvršiti obuku.

Parametar C je regulariozacioni parametar čije veće vrednosti ukazuju da se manji broj uzoraka nalazi sa pogrešne strane hiperravni i unutar margine, koja predstavlja rastojanje hipperravni i najbližeg uzorka iz obe klase. Manje vrednosti paramatra C ukazuju na suprotno.

Kada uzorci nisu linearno razdvojivi koristi se tzv. *kernel trik* koji podrazumeva preslikavanje uzoraka u višedimenzioni prostor obeležja u kojem su uzorci linearno razdvojivi, da bi se u tom prostoru pronašla optimalna hiperravan razdvajanja.

Postoje dva različita pristupa SVM klasifikatora u slučaju kada se primenjuje pri razdvajanju K klasa odnosno za K > 2: svaki protiv svakog (engl. *one vs. one*) i jedan protiv svih (*one vs. rest*). Oba pristupa podrazumevaju podelu skupa podataka više klasa u višestruke binarne probleme klasifikacije.

ODABIR OPTIMALNIH PARAMETARA

Da bi se pronašli optimalni parametri SVM klasifikatora i izvršila obuka, primenjena je metoda unakrsne validacije sa 10 podskupova za različite vrednosti parametra *C*, kernela i pristupa SVM klasifikatora. Oslanjajući se na mere tačnosti klasifikatora, kao optimalni parametri odabrani su: $C = 50$, linearni kernel i *one vs. rest* pristup.

Analizom matrice konfuzije, koja je dobijena akumulacijom matrica iz svake od 10 iteracija unakrsne iteracije, dobijene su mere uspešnosti SVM klasifikatora po klasama, na osnovu kojih se može uvideti da je najveća tačnost klasifikatora za klasu *pizzas* – 0.976, dok je tačnost za klasu *pastries* najmanja, odnosno, 74 recepta za peciva pogrešno je klasifikovano kao recept za kolačiće. Prosečna tačnost klasifikatora iznosi 0.939.

	Kolačići	Pecivo	Pica
Preciznost	0.892	0.893	0.961
Tačnost	0.926	0.913	0.977
Senzitivnost	0.893	0.893	0.961
Specifičnost	0.92	0.943	0.989
F-mera	0.893	0.893	0.961

SLIKA 6: Mere uspešnosti SVM klasifikatora po klasama na trening skupu

Dobijene mere su performanse SVM klasifikatora koji je obučen na trening skupu, a testiran na validacionom skupu. Neophodno je obučiti klasifikator na celokupnom skupu za trening, zatim testirati na test skupu čije podake klasifikator prvi put testira.

	Kolačići	Pecivo	Pica
Preciznost	0.923	0.927	0.977
Tačnost	0.943	0.948	0.985
Senzitivnost	0.923	0.923	0.977
Specifičnost	0.947	0.959	0.993
F-mera	0.923	0.927	0.977

SLIKA 7. Mere uspešnosti SVM klasifikatora za svaku klasu koji je testiran na skupu za testiranje

Može se uočiti da mera tačnosti klasifikatora za svaku klasu je veća od 0.9, što ukazuje na manji broj grešaka koje su napravljene prilikom klasifikacije recepata. Najbolje performanse pokazuje klasa *pizzas*, a najviše grešaka napravljeno je prilikom klasifikacije recepata za kolačiće i peciva, gde su 4 recepta za pecivo pogrešno klasifikovana kao recepti za kolačiće, a 5 recepata kolačića je klasifikovano kao recept za pecivo. Prosečna tačnost SVM klasifikatora, koja se dobije kao prosek mera tačnosti po klasama, iznosi 0,958.

V. ANALIZA PERFORMANSI kNN i SVM KLASIFIKATORA

kNN klasifikator koji je obučen na trening, a testiran na validacionom skupu, najviše grešaka je pravio prilikom klasifikacije uzoraka iz klase peciva kao uzorak iz klase kolačića. Istu osobinu pokazao je kNN klasifikator testiran na skupu za test, gde je takođe došlo do pogrešnog klasifikovanja recepta za pecivo kao recepta za kolačić, ali za manji broj uzoraka, što predstavlja poboljšanje. Može se doneti zaključak da postoje poteškoće prilikom klasifikovanja recepata u ove dve klase, što može biti posledica postojanja istih sastojaka u obe klase.

SVM klasifikator obučen na trening, a testiran na validacionom skupu takođe najveći broj grešaka je pravio prilikom klasifikacije recepata za peciva kao recepata iz klase kolačića, a isto ponašanje pokazuje i SVM klasifikator testiran na test skupu.

Može se doneti zaključak da oba klasifikatora greše prilikom klasifikacije recepata iz klase kolačića kao recept iz klase peciva i obrnuto, dok se najveća tačnost postiže za recepte iz klase pica u slučaju oba klasifikatora.

UPOREDBA MERA USPEŠNOSTI

	kNN	SVM
Preciznost μ	0.938	0.937
Preciznost M	0.954	0.943
Osetljivost μ	0.938	0.938
Osetljivost M	0.931	0.939
F-mera μ	0.938	0.938
F-mera M	0.939	0.941

SLIKA 8: Mikroprosečne i makroprosečne mere

Kada se uporede makroprosečne i mikroprosečne mere uspešnosti klasifikatora, oba klasifikatora imaju slične performanse sa vrednostima mera iznad 0.9, što predstavlja dobre rezultate.

Oba klasifikatora prave iste greške koje podrazumevaju klasifikaciju recepata za kolačiće kao recepte za peciva i obrnuto. Performanse klasifikatora bi se mogle poboljšati kada bi u bazi pored prisustva postojala i informacija o procentu prisustva pojedinih sastojaka u receptima što bi na primer olakšalo klasifikaciju recepata za kolačiće i peciva, jer recimo sastojak šećer je prisutan u oba recepta, međutim sigurno je količina šećera u receptu za kolačić veća od količine u receptu za pecivo te bi ta informacija bila od značaja prilikom klasifikacije.