Klasifikacija recepata

Vanja Babić, IN14/2017, babicvanja11@gmail.com

I. Uvod

Kolačići, peciva i pice svakako jesu neizbežan deo svakaog događaja i hrana koja se često priprema. Ukoliko bismo potražili recepte, naišli bismo na bezbroj različitih sastojaka i njihovih kombinacija koje se koriste. Kako bismo se bolje upoznali sa tim koji sastojci se češće koriste u pripremi koje hrane i kako po receptu prepoznati o kojoj od ove tri vrste hrane je reč, u nastavku izveštaja ćemo se baviti različitim receptima i analizom njihovih sastojaka.

II. Baza podataka

Baza podataka korištena za izradu ovog izveštaja sadrži 1931 uzorak, od kojih se 1738 koristi za trening i kreiranje modela, a 193 za testiranje modela. Svaki uzorak predstavlja recept i govori nam da li se u tom receptu pojavljuje određeni sastojak ili ne. Ukoliko se sastojak u receptu pojavljuje, vredost za taj sastojak u uzorku je 1, a ukoliko se ne pojavljuje 0. U bazi se nalaze 133 sastojka, a neki od njih su: brašno, ulje, jaja, mleko, šećer, puter, jabuke, cimet, sir, itd. Baza sadrži obeležje *class* koje može da ima jednu od tri vrednosti: 'Cookies', 'Pizzas' i 'Pastries' i ono nam govori da li se radi o receptu za keksić, picu ili pecivo. U bazi nema nedostajućih podataka. Među trening uzorcima se nalaze 723 uzorka klase keksići, 619 uzoraka klase pevica i 396 uzoraka klase pica, dok u testnom skupu imamo 80 uzoraka klase keksići, 69 uzoraka klase peciva i 44 uzorka koja pripadaju klasi pica.

III. Analiza podataka

Sastojci koji se najčešće pojavljuju u receptima za pripemu pice su so, ulje, sir, brašno, paradajz, testo. Za pripremu peciva najčešći sastojci su puter, jaja, brašno, so i šećer, a za pripremu kolačića najčešće se koriste šećer, puter, jaja, brašno, čokolada, vanila. Peciva su ta kod kojih se generalno koristi najveći broj datih sastojaka, što je razumljivo obzirom da ona mogu da budu i slana i slatka i da je veliki broj namirnica koje mogu da se koriste u njihovoj pripremi. Takođe, u pripremi kolačića se nikada ne pojavljuju sastojci kao što su luk, ljuta paprika, slanina, kečap ili tikvica, dok se u receptima za picu nikada neće naći kako, kokos, jagoda, sladoled, itd. što je takođe očekivano i intuitivno.

IV. KNN klasifikator

Metoda k najbližih suseda (KNN – *k nearest neighbors*) predstavlja neparametarsku metodu klasifikacije i direktno koristi trening skup za klasifikaciju. Za svaki novi uzorak se pronalazi njegovih k najbližih suseda iz trening skupa, a zatim se taj novi uzorak svrstava u onu klasu kojoj pripada najveći broj njegovih suseda, jer se pretpostavlja da je neviđeni uzorak sličniji onim uzorcima za trening koji su mu bliži u prostoru obeležja. Kao mera udaljenosti koristi se metrika. Neke od najčešće korištenih su Euklidska, Menhetn, Hemingova, Žakarova, itd. Odabir metrike zavisi od tipa obeležja, od toga da li obeležja imaju realne ili celobrojne vrednosti, binarne, i sl.

Klasifikator će se obučavati metodom unakrsne validacije sa 10 podskupova. Dakle, trening skup se deli u 10 podskupova i u svakoj od 10 iteracija se jedan od podskupova izdvaja kao test skup, a ostalih 9 se koristi za obuku. Na ovaj način će se svaki uzorak u jednom momentu naći u test skupu i tako izbegavamo natprilagođenje modela nekom podskupu. Procena greške za model sa određenim vrednostima hiperparametara dobija se kao srednja vrednost procena dobijenih prilikom testiranja 10 treniranih modela u postupku unakrsne validacije.

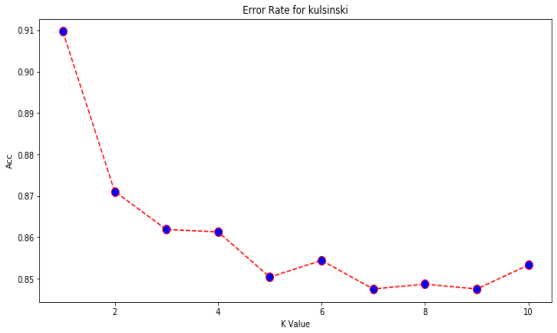
U datoj bazi podataka vrednosti svih obeležja su ili 0 ili 1. Samim tim, normalizacija vrednosti obeležja nije neophodna. U skladu sa tim što su sve vrednosti u bazi binarne, bira se i metrika. Metrike koje se najčešće koriste u slučaju binarnih vrednosti obeležja su Žakarova, *Dice*, Kulsinski, metrika podudaranja. Žakarova metrika računa rastojanje kao broj mesta na kojima se uzorci razlikuju kroz ukupan broj nenultih dimenzija. *Dice* metrika računa rastojanje kao količnik broja mesta na kojima se uzorci razlikuju i sume nenultih dimezija i onih u kojima oba uzorka imaju vrednost 1. Metrika podudaranja vrednosti rastojanja računa kao broj mesta na kojima se uzorci razlikuju kroz ukupan broj uzoraka.

Za vredost parametra k, tj. broj suseda koji će se uzeti u obzir, isprobaćemo vrednosti od 1 do 10, a konačan model će se trenirati i testirati za ono k i onu metriku koja tokom obuke da najbolje rezultate.

V. Rezultati klasifikacije metodom knn

Nakon obuke klasifikatora, najbolju tačnost je dao klasifikator obučen za k = 1 i sa metrikom Kulsinski. Udeo uspešno klasifikovanih uzoraka je 91%. Na slici 1

može da se vidi kako se menjala uspešnost klasifikatora sa metrikom Kulsinski za različite vrednosti parametra k. Vidimo da je ubedljivo najveći broj uspešno klasifikovanih uzoraka za k=1, dok se za sve ostale vrednosti k, tj. od 2 do 10, udeo uspešno klasifikovanih uzoraka kreće približno između 85% i 87%.



Slika 1. Grafik uspešnosti kNN klasifikacije sa metrikom Kulsinski nakon 10 iteracija unakrsne validacije za svako k

Na slici 2 prikazana je matrica konfuzije za odabrani klasifikator nakon 10 iteracija unakrsne validacije. U vrstama su prikazane prave vrednosti, a u kolonama predviđene, dok je redosled klasa: kolačići, peciva, pice. Vidimo da se 56 puta desilo da kolačić bude klasifikovan kao pecivo, a 54 puta da pecivo bude klasifikovano kao kolačić. Dakle, najčešće se greška dešava prilikom klasifikacije uzoraka ove dve klase, što je donekle i očekivano, obzirom da iskustveno znamo da za pripremu peciva i kolačića koristimo slične sastojke. Peciva su takođe 20 puta klasifikovana kao pica, što isto možemo da razumemo, jer za pripremu slanih peciva koristimo slične sastojke kao kada pravimo picu.

Tačnost za klasu keksići iznosi 93%, za klasu peciva 91% i za klasu pice 97%. Prosečna tačnost klasifikatora iznosi 93%.

664	56	3
54	545	20
6	18	372

Slika 2. Matrica konfuzije nakon 10 iteracija unakrsne validacije za parametre k = 1 i metriku Kulsinski

Nakon što su utvrđeni parametri koji su dali najbolje rezultate na trening uzorku, klasifikator je konačno obučen na celokupnom trening skupu i testiran na test skupu koji je na samom početku izdvojen. Dobijena je matrica konfuzije prikazana na slici 3. Tačnost za klasu keksići iznosi 95%, za klasu peciva 93% i za klasu pice 97%. Prosečna tačnost klasifikatora iznosi 95%.

77	3	0
5	63	1
1	3	40

Slika 3. Matrica konfuzije dobijena na osnovu rezultata za test skup za KNN klasifikator

Vidimo da se na ovom test skupu ređe dešava pogrešna klasifikacija keksića u peciva i obrnuto, a samo se jednom desilo da su peciva klasifikovana kao pica. Takođe, pogrešna klasifikacija između keksića i pice se desila samo jednom, ali se ovaj slučaj retko dešavao i tokom unakrsne validacije. Dakle, najveće greške se dešavaju prilikom klasifikacije peciva i to definitivno jeste očekivano, obzirom da, kao što je već rečeno, veliki broj sastojaka možemo da koristimo tokom njihove pripreme i zasigurno je da peciva imaju najveći broj različitih recepata sa različitim sastojcima.

VI. SVM KLASIFIKATOR

Mašina na bazi vektora nosača ili SVM (*Support Vector Machine*) je klasifikator koji se zasniva na klasifikatoru maksimalne margine, čiji je cilj da podeli prostor na dva dela tako da se u jednom delu nađu samo uzorci jedne klase, a u drugom delu uzorci iz druge klase, pomoću hiperravni koja treba da se odredi. Da bi ovo bilo moguće, klase treba da budu linearno separabilne, a to je u praksi redak slučaj. Zbog toga se dozvoljava da uzorci jedne klase pređu na stranu uzoraka druge klase, tj. na da budu na pogrešnoj strani hiperravni razdvajanja i tada dobijamo klasifikator meke margine.

Jedan od parametara koji se podešava je parametar C i on predstavlja regularizacioni parametar i određuje kolika će biti tolerancija greške. Kako bi se utvrila najbolja vrednost parametra C za datu bazu, isprobaćemo vrednosti 10,20,30,40 i 50.

Obzirom da postoje problemi kod kojih linearna separabilnos nije moguća, tada se primenjuje *kernel trik*, tj. uzorci se preslikavaju u višedimenzioni prostor u kom jesu linearno separabilni, a tome u linearnom prostoru odgovara nelinearna granica odlučivanja. Za dati problem biće isprobani linearni, radijalni i polinomijalni kernel.

Iako je rečeno da SVM deli prostor na dva dela, odnosno da radi binarnu klasifikaciju, ovaj algoritam je moguće koristiti i kada imamo problem sa više klasa. Postoje dva pristupa koja se primenjuju: svaki protiv svakog - OVO (one vs one) i jedan protiv svih – OVR (one vs rest) i biće isprobane obe opcije kako bismo utvrdili koja daje bolje rezultate za dati problem.

Pri kreiranju ovog modela takođe se koristi unakrsna validacija sa 10 podskupova.

VII. REZULTATI KLASIFIKACIJE METODOM SVM

Najbolje rezultate dao je SVM klasifikator sa parametrima C = 30, za radijalni kernel, a OVO i OVR pristupi daju identične rezultate tako da je moguće koristiti bilo koji od njih. Udeo uspešno klasifikovanih uzoraka je 92%.

Matrica konfuzije za SVM klasifikator nakon 10 iteracija unakrsne validacije prikazana je na slici 4. Kao i kod KNN klasifikatora, vidimo da se najčešće mešaju klase kolačići i peciva, tačnije, 52 puta je kolačić klasifikovan kao pecivo, a 43 puta je pecivo klasifikovano

kao kolačić. Pecivo je 10 puta klasifikovano kao pica, što je duplo manje nego kod KNN klasifikatora nakon unakrsne validacije. Pica je 10 puta pogrešno klasifikovana kao kolačić i 10 puta kao pecivo.

Tačnost za klasu keksići iznosi 93%, za klasu peciva 93% i za klasu pice 98%. Prosečna tačnost klasifikatora iznosi 94%.

669	52	2
43	565	11
10	10	376

Slika 4. Matrica konfuzije nakon 10 iteracija unakrsne validacije za parametre C = 30, radijalni kernel i OVR pristum

Nakon što su utvrđeni najbolji parametri, SVM klasifikator je obučen na celokupnom trening skupu, a testiran na unapred odvojenom test skupu. Dobijena matrica konfuzije prikazana je na slici 5.

Kolačići su samo tri puta pogrešno klasifikovani u klasu peciva a ni jednom se nije desilo da kolačić bude klasifikovan kao pica. Pecivo takođe ni jednom nije bilo pogrešno klasifikovano kao pica, ali jeste 3 puta kao kolačić. Pica ni jednom nije klasifikovana kao pecivo, ali se tri puta desilo da bude pogrešno klasifikovana kao kolačić i ovo je dosta neočekivan rezultat, obzirom da se do sada retko dešavalo da pice budu klasifikovane kao kolačići i obrnuto, dok se nešto češće javljala greška između klasa peciva i pice.

Tačnost za klasu keksići iznosi 95%, za klasu peciva 96% i za klasu pice 98%. Prosečna tačnost klasifikatora iznosi 96%.

77	3	0
3	66	0
3	0	41

Slika 5. Matrica konfuzije dobijena na osnovu rezultata za test skup za SVM klasifikator

VIII. Poređenje rezultata KNN i SVM klasifikatora

Kod KNN klasifikatora je tačnost klase keksići 95%, klase peciva 93% i klase pica 97%. SVM klasifikator je za klasu keksići dao tačnost od 95%, za klasu peciva 96% a za klasu pica 98%. Dakle, vidimo da SVM klasifikator daje nešto bolje rezultate u odnosu na KNN, mada razlika nije velika. Najveća razlika je svakako za klasu peciva. Prosečna tačnost KNN klasifikatora je 95%, a SVM 96%.

Mikro i makro mere se kod oba klasifikatora ne razlikuju značajno. Mikro i makro preciznost kod KNN klasifikatora iznose oko 93%, a kod SVM klasifikatora oko 96%, tako da opet prednost po dobijenim rezultatima za ovaj problem ima SVM klasifikator. Mikro i makro osetljivost i F mera KNN klasifikatora iznose oko 93%, a kod SVM klasifikatora oko 95%. Stopa greške KNN klasifikatora je oko 13%, a SVM klasifikatora oko 9%.