Klasifikacija porijekla recepata na osnovu prisustva određenih sastojaka

Natalija Krsmanović, IN21/2018, krsmanovic.natalija99@gmail.com

I. Uvod

Tema izvještaja jeste analiza recepata i njihovih sastojaka i utvđivanje specifičnosti među sastojcima. Cilj je da se izvrši predikcija tj. kreira klasifikator koji će na osnovu zadatih sastojaka svrstati određeni recept u određenu kategoriju tj.odrediti iz koje zemlje on potiče.

II. Opis baze podataka

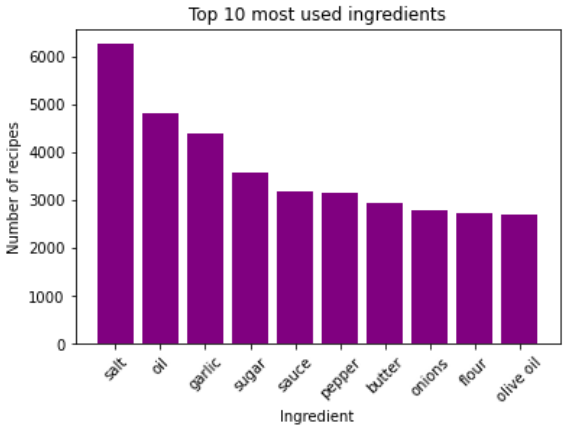
Baza sadrži podatke o 10566 uzoraka i 151 obilježju. Od prisuntih obilježja 150 predstavlja sastojke kao što su:brašno,šećer,ulje,so,voda i mnogi drugi,kao i dosta sastojaka koji se koriste izraženo u određenim kulturama, dok posljednje obilježje označava kategoriju odnosno zemlju porijekla. Jedan uzorak se odnosi na recept iz neke zemlje sa označenim sastojcima koji se u njemu koriste.

III. Analiza podataka

Analizom date baze podataka uočeno je da ne postoje nedostajuće vrijednosti, kao i da su podaci koji se odnose na sastojke predstavljeni u binarnom obliku (0 i 1). To znači da je za svaki sastojak prikazan podatak da li je on prisutan u receptu ili nije,a ne i sama količina upotrebljenog sastojka. Za posljednje obilježje tj. klasnu labelu koristi se 9 različitih oznaka odnosno u datoj bazi postoji 9 različitih zemalja porijekla datih recepata. To su: Velika Britanija, Kina, Francuska, Grčka, Italija, Japan, Meksiko, Jug SAD i Tajland. Utvrđen je sledeći broj recepata po klasama: Britanija (509), Kina (1291), Francuska (1565), Grčka (587), Italija (1670),Japan (755), Meksiko(1274), Jug SAD (2303), Tajland (612). Iz prikazanih podataka se vidi da postoji izražena nebalansiranost između klasa, gdje najviše recepata dolazi iz SAD a najmanje iz Britanije čiji je broj čak oko 4 puta manji. Ovu opasku treba uzeti u obzir prilikom daljne analize.

A. *Analiza pojavljivanja određenih sastojaka*

Analiza je izvršena tako što su se posmatrali najzastupljeniji sastojci. Na slici 1 su prikazani sastojci koji su najviše prisutni u odnosu na cijelu bazu podataka,a kada se posmatraju klase analizirani su oni sastojci koji su prisutni u više od 15% recepata za svaku klasu. Utvrđeno je da postoje sastojci koji se koriste u receptima iz svih klasa, I to su: so,ulje,šećer,voda,bijeli luk i drugi koji se podudaraju sa sastojcima na slici 1. Međutim iako se dati sastojci koriste u svakoj zemlji njihov procenat pojavljivanja u receptima varira odnosno nisu prsutni za sve zemlje u velikim količinama pa se ipak u bazi zadržavaju sva obilježja bez izbacivanja.



Slika 1: 10 najčešće upotrebljenih sastojka

Također se može primjetiti da sastojci koji se koriste zavise i od geografskog položaja zemalja. Tako na primjer postoji velika sličnost kod recepata Velike Britanije i Francuske , Japana i Kine i Tajlanda ili Italije i Grčke. Satojci koje Britanija,Francuska i SAD koriste više od ostalih zemalja su puter, sećer, veće količine brašna.U britanskim receptima se pojavljuje više puta mlijeko kao i upotreba govedine. Za Italiju i Grčku se se izdvaja velika potrošnja maslinovog ulja, gdje Italija dodatno koristi više paradajiza i bosiljka,a Grčka maslina i feta sira. Uočava se odstupanje i kineskih i japanskih i tajlandskih recepata po pojavljivanju riže, sojinog sosa i sosa od povrća. Meksiko koristi dosta specifičnih sastojaka od kojih su neki: čili, tortilje,kim,..

IV. KNN klasifikator

Zbog upotrebe klasifikatora, baza podataka je podijeljena na 2 podskupa koji predstavljaju skup za trening i skup za test pri čemu test čini 10% ukupnih uzoraka.Prvi klasifikator koji se koristi jeste klasifikator metodom k najbližih susjeda (KNN).

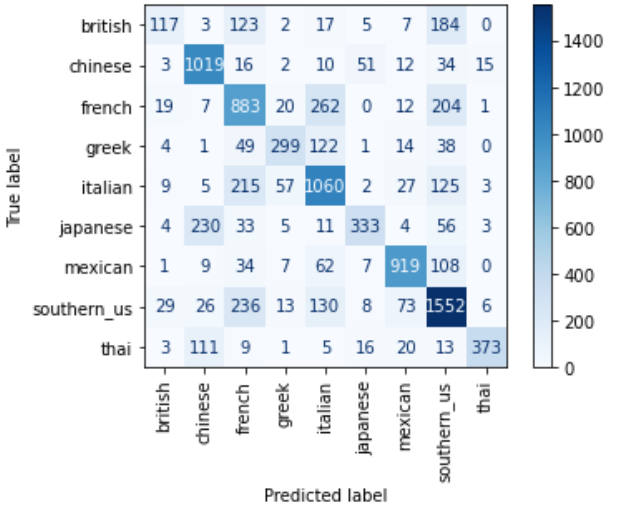
A. *Određivanje optimalnih parametara*

Parametri koji su potrebni za primjenu KNN metode pored labeliranih uzoraka su: broj najbližih susjeda koji se posmatraju pri odlučivanju (k) i metrika računanja rastojanja (m). Da bi se odredili optimalni parametri iskorištena su dva načina. Prvi je upotreba GridSearchCV funkcije koja za zadate moguće parametre ispituje koji daju najbolji rezultat. Navedene opcije za parametar k su:1,3,5,10,15 i 17, a za metrike su ‘Jaccard’, ‘Dice’ i

'Matching' koje se koriste za binarna obilježja. Mjera uspješnosti na osnovu koje su se poredili rezultati je mikro osjetljivost zbog nebalansiranosti između klasa. Kao rezultat je dobijeno da su optimalne vrijednosti za broj susjeda 17 i 'Jaccard' za metriku. Drugi način na koji se došlo do optimalnih parametara je ručno ispisana unakrsna validacija za kombiacije istih mogućih vrijednosti za k i za m. Prvi i drugi način su zapravo ekvivalentni tj. zasnivaju se na unakrsnoj validaciji pa se zbog toga drugi način koristio da bi se potvrdili rezultati iz GridSeerchCV i prikazale matrice konfuzije za sve kombinacije parametara. Kao rezultat ručnog načina također su dobijeni k=17 i m='Jaccard' sa mikroprosječnom osjetljivošću od 68.93%. Oba načina su za unakrsnu validaciju koristili StratifiedKFold funkciju sa 5 particija po kojima su se u jednakom odnosu raspoređivali uzorci po klasama.

B. Analiza matrice konfuzije dobijene kao rezultat unakrsne validacije

Matrica konfuzije za dobijene optimalne parametre nakon pomenute StratifiedKFold unakrsne validacije je prikazana na slici 2. Kao što se može primjetiti veliki broj uzoraka nije pravilno klasifikovan. To je najviše izraženo kod recepata iz Velike Briatanije gdje je tačno klasifikovano 117 recepata ali čak 184 britanskih recepta je klasifikovano kao američki i 123 kao francuski recepti. To se dešava zbog toga što su sastojci u velikoj mjeri slični za navedene države kao što je prethodno i pomenuto u analizi pojavljivanja sastojaka.Također se uočava i velika greška kod klasifikacije japanskih recepata. Njih 333 je tačno klasifikovano ali 230 je proglašeno kao kineski recept. Razlog je isti kao i u prethodnom slučaju zbog sličnih sastojaka, ali ono što također utiče je i broj uzoraka.Za kineske recepte je dostupno skoro duplo više uzoraka, pa se zato dosta japanskih kategoriše kao kineski, dok obrnuto i nije slučaj.

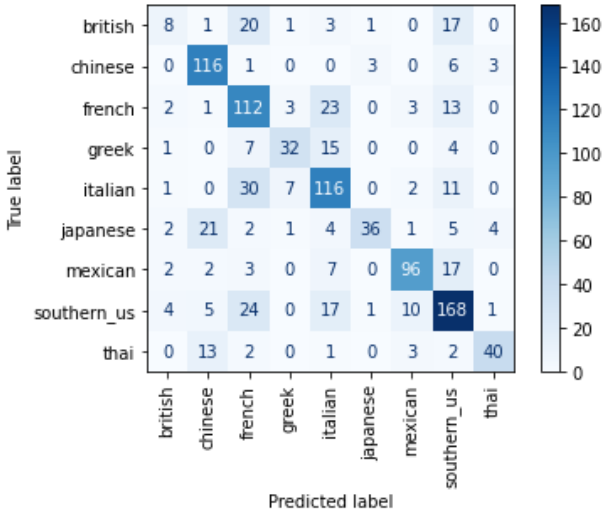


Slika 2:Matrica konfuzije nakon unakrsne validacije

Bolja analiza date matrice konfuzije može da se odredi na osnovu mjere osjetljivosti. U tabeli 1 je prikazana prosječna osjetljivost mikro tipa za datu matricu konfuzije kao i osjetljivost za svaku od klasa u koloni za unakrsnu vaidaciju (druga kolona). Od prikazanih najveću osjetljivost ima Kina sa 87.6% što znači da od 100 recepata koji dolaze iz Kine njih 87 je klasifikovano tačno. Kao što je i očekivano klasifikator radi najboljle za klase koje su imale dostupan veliki broj uzoraka, kao što su Italija,Francusa,Meksiko,SAD. Najmanja osjetljivost je izračunata za Britaniju koja ujedno ima i najmanji broj dostupnih recepata. Iako američkih recepata ima najviše, meksikčk recepti su bolje klasifikovani jer sadrže više sastojaka koji su jedinstveni baš za tu državu.

C. Analiza matrice konfuzije na test skupu

Nakon unakrsne validacije slijedi obučavanje modela nad cijelim trening skupom i predikcija nad test skupom koji se u prethodnim koracima nije koristio. Za finalni KNN klasifikator se također koriste kao parametri broj susjeda koji je 17 i 'Jaccord' metrika. Matrica konfuzije koja je dobijena kao finalni rezultat je prikazana na slici 3. Kao i nakon unakrsne validacije najviše grešaka je pri klasifikaciji recepata koji bi trebali biti iz Britanije ali kao što se vidi naspram 8 tačnih, 20 i 17 je klasifikovano sa francuskim i američkim porijeklom redom. I ostale vrijednosti su slične kao i kod prethodne matrice konfuzije što se može vidjeti i po bojama datih ćeija.



Slika 3: Matrica konfuzije nad test skupom

U tabeli 1 u posljednjoj koloni su prikazane prosječna i osjetljivost za svaku klasu. Prosječna mikro osjetljivost se smanjila na konačnom modelu sa 68.9% na 68.4%. Najveće odstupanje za klase je kod Velike Britanije, gdje je osjetljivost pala sa 25.5% na 15.6% . I nakon testiranja za kineske recepte je najbolja klasifikacija gdje je od 129 uzoraka njih 116 proglašeno tačnim odnosno 89.9%.

Tabela 1 :Prosječna osetljivost i osjetljivosti po državama nakon testiranja nad test skupom

Država	**Osjetljivost-Unakrsna validacija**	**Osjetljivost-Finalni model**
Velika Britanija	0.255	0.156
Kina	0.876	0.899
Francuska	0.627	0.713

Grčka	0.566	0.542
Italija	0.705	0.694
Japan	0.490	0.473
Meksiko	0.801	0.755
Jug SAD	0.748	0.730
Tajland	0.676	0.655
Sve zajedno	**Mikroprosječn a osjetljivost**	**Mikroprosječna osjetljivost**
	0.689	0.684

V. SVM KLASIFIKATOR

Drugi klasifikator koji je upotrebljen za klasifikaciju recepata jeste klasifikator na bazi vektora nosača.

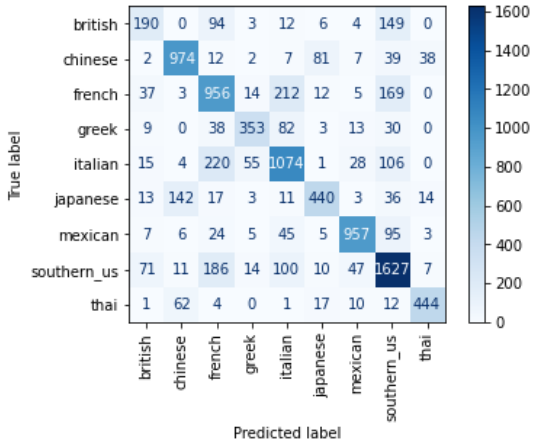
A. Određivanje optimalnih parametara

Parametri koje je potrebno odrediti su: regularizacioni parametar koji određuje toleranciju na pogrešnu klasifikaciju (C), funkcija kernela kojom se povećava dimenzionalnost zarad bolje razdvojivosti (kernel), način donošenja odluke kod višeklasnih problema (decision_function_shape) i težinski faktor (class_weight). Kao i kod KNN klasifikatora korištena je GridSeerchCV funkcija i ručni pristup prolaženja kroz moguće opcije parametara da bi se dobili oni optimalni. Kao rezultat za oba pristupa su sledeći vrijdnosti za parameter: C=5, kerel='rbf', decision_function_shape='one vs rest', class_weight=none. S obzirom da je izabrani kernel rbf tipa, potrebno je podesiti tj. odrediti i dodatni parametar gamma. Nakon isprobanih kombinacija, gama je postavljen na vrijednost 'scale'.

B. Analiza matrice konfuzije dobijene kao rezultat unakrsne validacije

Na slici 3 je prikazana matrica konfuzije SVM modela nakon unakrsne validacije sa prethodno izabranim optmalnim parametrima. Za razliku od KNN modela, SVM model daje bolje rezultate što se može vidjeti sa prikazane matrice. Iako se najčešće greške klsifikacije javljaju kod britanskih recepata sada je taj broj manji, odnosno 190 recepata je klasifikovano tačno, dok je od onih koji trebaju da pripadaju klasi Britanija njih 94 u klasi Francuska a 149 u američkoj klasi. Kao interesantan detalj može se izdvojiti to što za recepte sa Tajlanda ne postoji veliki promašaj kada su u pitanju recepti koji su svrstani u tajlandske a ne pripadaju toj klasi.

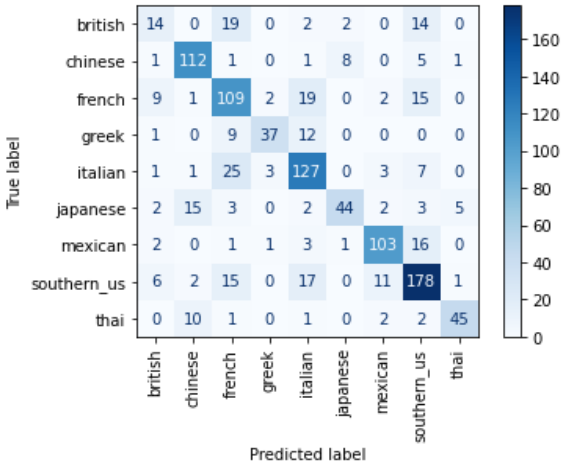
Mjera osjetljivosti je izračunata i prikazana na slici 4, Kina i dalje predstavlja zemlju sa najbolje klasifikovanim vrijednsotima. Poslije dolazi Tajland sa 80% tačno klasifikovnaih uzoraka naspram pozitivne kalse. To se također može objasniti prisustvom pojedinih sastojaka koji nisu u velikoj mjeri zastupljeni u ostalim državama što pomaže prilikom klasifikacije. Japanski recepti se i kod ovog klasifikatora u velikom broju klasifikuju kao kineski.



Slika 4: Matrica konfuzije nakon unakrsne validacije

C. Analiza matrice konfuzije nad test skupom

Za obučavanje konačnog modela iskorišten je cijeli trening skup,a testiranje je vršeno na prethodnom izdvojenom test skupu. Za konačni model upotrebljeni su optimlani parametri dobijeni nakon unakrsne validacije. Matrica konfuzije koja prikazuje rezultate klasifikacije nad test skupom je na slici 5.



Slika 5 : Finalna matrica konfuzije dobijena nad test skupom

Da bi se bolje uočila razlika u odnosu na rezultate unakrsne valiacije posmatraju se vrijednosti mjere osjetljivosti iz tabele 2. Nakon testiranja na pravom test skupu prosječna osjetljivost je pala sa 73.7% na 72.7% , što i nije velika razlika pa se može reći da je model podjednako vršio klasifikaciju i prilikom unakrsne validacije kao i na test skupu, odnosno nije došlo do preobučavanja skupa. Klase kod kojih se može izdvojiti veća razlika je Velika Britanija čija je osjetljivost nad test skupom lošija za 14%. Uočava se i da je prilikom klasifikacije recepata iz Britanije kod unakrsne validacije mnogo veći broj klasifikovan kao američki, dok je na test skupu, najveći broj klasifikovan kao francuski tj. klasifikator se drugačije ponaša za klase sa manjim brojem uzoraka. Ostale klase u prosjeku razlikuju za otprilike 5 %.

Tabela 2: Prosječna osjetljivost i osjetljivosti po državama nakon testiranja nad test skupom

Država	**Osjetljivost-Unakrsna validacija**	**Osjetljivost-Finalni model**
Velika Britanija	0.414	0.274
Kina	0.838	0.868
Francuska	0.678	0.694
Grčka	0.714	0.627
Italija	0.714	0.760
Japan	0.473	0.578
Meksiko	0.755	0.811
Jug SAD	0.730	0.774
Tajland	0.805	0.737
Sve zajendo	**Mikroprosječna osjetljivost**	**Mikroprosječna osjetljivost**
	0.737	0.727

VI. POREĐENJE REZULTATA KNN I SVM KLASIFIKATORA

Na osnovu podataka iz tabele 3 koji predstavljaju rezultate za oba klasifikatora nad konačnim modelom može se utvrditi da za sve klase blago veće vrijednosti daje SVM klasifikator. Mjera koja najviše varira za klase u slučaju oba klasifikatora jeste osjetljivost. Za Kinu i Francusku je blago viša za KNN klasifikator ali u ostalim slučajevima SVM klasifikator se pokazao bolji. I za KNN i SVN Kina ima najveći udio tačno klasifikovanih racapata od klase pozitiva. Što se tiče specifičnosti koja predstavlja udio ispavno klasifikovanih recepata iz klase negativa za sve klase za oba klasifikatora prelazi 90%. To se može objasniti time što je broj klasa veći i zato je veliki broj uzoraka koji su ispravno klasifikovani u klasi negativa, zato data mjera nije najbolja za upoređivaje usjpješnosti klasifikatora u ovom slučaju.

Preciznost je udio ispravno klasifikovanih uzoraka naspram svih predviđenih kao pozitivni. Za oba klasifikatora su približno jednake vrijdnosti, gdje je za oba najmanja zabilježena kod Velike Britanije. Najbolje klasifikovani recepti prema preciznosti za KNN je Japan(87.8%) što znači da od 100 recepata koji su predviđeni kao japanski njih oko 88 stvarno potiče iz Japana. Za SVM po preciznosti su najbolje klasifikovani recepti iz Tajlanda.

Tabela 3: Mjere uspješnosti po klasama za oba klasifikatora

Država	**Osjetljivost**	**Specifičnost**	**Preciznost**
Velika Britanija (KNN)	0.157	0.988	0.400
Velika Britanija (SVM)	0.274	0.978	0.388
Kina(KNN)	0.899	0.953	0.729
Kina(SVM)	0.868	0.968	0.794
Francuska(KNN)	0.713	0.901	0.557
Francuska(SVM)	0.694	0.917	0.595
Grčka(KNN)	0.542	0.987	0.727
Grčka(SVM)	0.627	0.994	0.860
Italija(KNN)	0.694	0.921	0.623
Italija(SVM)	0.760	0.935	0.690
Japan(KNN)	0.473	0.990	0.878
Japan(SVM)	0.578	0.988	0.800
Meksiko(KNN)	0.755	0.979	0.834
Meksiko(SVM)	0.811	0.978	0.837
Jug SAD(KNN)	0.730	0.909	0.691
Jug SAD(SVM)	0.774	0.925	0.742
Tajland(KNN)	0.655	0.992	0.833
Tajland(SVM)	0.737	0.993	0.965

Kada se upoređuju vrijednosti mjera na globalnom nivou koje su prikazane u tabeli 4. Razlika između mikro i makro mjera je to što se za mikro mjere prvo sumiraju sve komponente po klasama (TP,TN,FN,TN) a zatim se na njih primjenjuju formule za mjere dok se za makro mjere računaju za svaku klasu pa se traži njihov prosjek. Nova mjera koja je također prikaza je F mjera koja označava harmonijsku sredinu između preciznosti i osjetljivosti. Ono što je zanimljivo je da su sve mjere podjednake za svaki od modela odnosno iako postoji nebalansiranost između klasa to nije uticalo na mjere na nivou čitave matrice konfuzije. Za sve mjere SVM klasifikator ima veće vrijednosti od KNN za oko 4%.

S obzirom na date rezultate kao i rezultate po pojedinačnim klasama može se zaključiti da je za dati problem na ovom setu podataka bolji izbor SVM klasifikator. Zajedničko za oba modela je to što su pravili greške nad klasama koje su imale manji broj dostupnih uzoraka, a ono što je također uticalo na predikciju je i prisutnost specifičnih sastojaka za svaku zemlju. Prednost KNN modela je ta što nema proces obuke modela ali ipak klasifikacija može duže da traje ukoliko je set podataka veliki jer se računa distanca za sve uzorke, dok SVM vrši bolju predikciju ukliko je broj klasa veći, također je otporniji na autlajere dok bi se za KNN trebala izvršiti standardizacija prije njegove upotrebe.

Tabela 4 :Mjere uspješnosti za oba klasifikatora

Mjera uspješnosti	**KNN**	**SVM**
Proječna tačnost	0.685	0.727
Preciznost mikro	0.685	0.727
Preciznost makro	0.697	0.730
Osjetljivost mikro	0.685	0.727
Osjetljivost makro	0.625	0.680
F mjera mikro	0.685	0.727
F mjera makro	0.643	0.699