Analiza podataka o receptima

Anastasija Popović, IN46/2018, popovicanastasija17@gmail.com

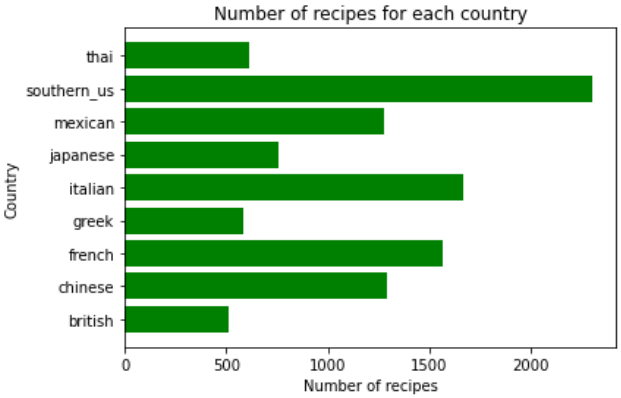
I. Uvod

U ovom izveštaju prikazaćemo analizu podataka korišćenja 150 različitih namirnica koje neke od država koriste kao sastojke u svojim receptima. Imamo ukupno 10566 recepata koji su nekog od sledećeg porekla: britanskog, kineskog, francuskog, grčkog, italijanskog, japanskog, meksičkog, južno-američkog kao i tajlandskog. Istraživanjem i analiziranjem podataka prikazaćemo neke od zaključaka koje smo kasnije potvrdili kNN i SVM klasifikatorima.

II. Baza podataka

Analiziranjem baze podataka zaključujemo da nema nedostajućih podataka, kao i nevalidnih i nelogičnih vrednosti. Baza sadrži kolone: „Unnamed: 0“, 150 različitih sastojaka i kolonu „country“. Kolonu „Unnamed: 0“ izbacujemo jer nam nije potrebna tokom analize i dodaćemo novu kolonu „recipes_no“ koja će prikazivati ukupan broj recepata po poreklu.

A. *Analiza recepata po poreklu*

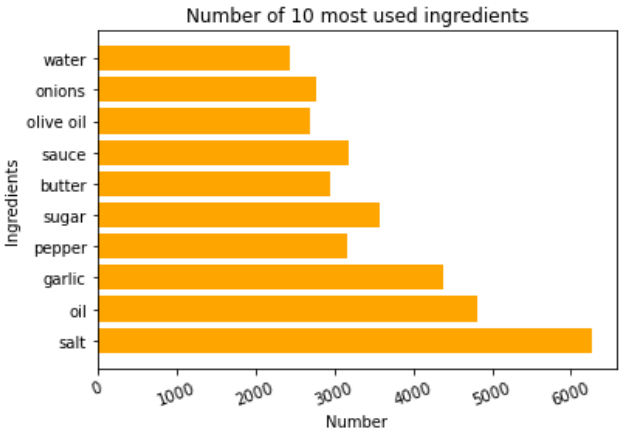


Slika 1. Prikaz recepata po poreklu

Sa slike 1 uočavamo da je najviše recepata južno-američkog porekla dok britanskih recepata ima najmanje. Primetimo da italijanska i francuska kuhinja imaju približan broj recepata.

B. *Analiza broja sastojaka po receptima*

Prikazan je broj recepata po poreklu te sada možemo da razmatramo broj sastojaka koji se koristi u tim receptima. Prikazaćemo 10 najčešće korišćenih sastojaka u receptima.



Slika 2. Prikaz 10 najčešće korišćenih sastojaka u receptima

Primećujemo na slici 2 da se je so kao začin hrani najzastupljeniji u receptima, zatim ulje. Primetimo da se voda nalazi na desetom mestu po najčešće korišćenim sastojcima te možemo zaključiti da se više recepata priprema na ulju nego na vodi.

C. *Analiza sastojaka po poreklu*

Tabelarno ćemo prikazati po poreklu koji sastojci se najmanje i najviše koriste što će nam pomoći da uočimo glavne sastojke za svaki predeo.

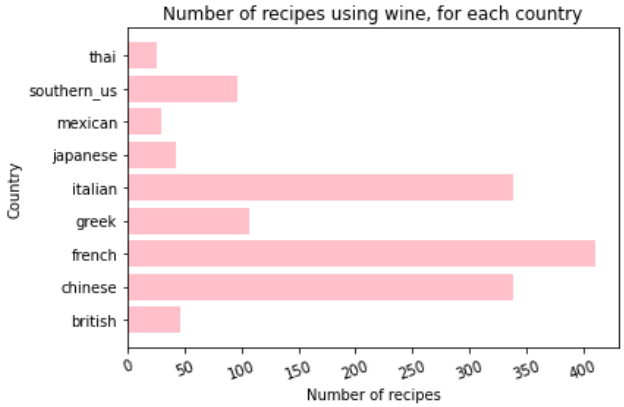
Tabela 1: Prikaz korišćenja sastojaka

Poreklo	***Često korišćeni***	***Retko korišćeni***
Britansko	Sos	Prašak za pecivo
Kinesko	So	Mirin
Francusko	Sos	Origano
Grčko	So	Riblje ulje
Italijansko	Ulje	Riblje ulje
Japansko	Sos	Origano
Meksičko	Ulje	Susamovo ulje
Južno-američko	So	Susam
Tajlandsko	So	Tortilja

Kao što je u prethodnoj analizi prikazano da je so najčešće korišćeni sastojak u receptima, iz tabele 1 možemo videti i koji predeli je koriste kao glavni sastojak. Primetićemo da su italijanska i grčka kuhinja slične mediteranske kuhinje koje retko koriste riblje ulje kao sastojak u receptima.

D. *Analiza recepata koji koriste vino kao sastojak*

Imamo 5 vrsta vina koja se koriste kao sastojci u reseptima, a to su: belo vino, suvo belo vino, crveno vino, vino i vinsko sirće. Procenat recepata koji koriste vino kao sastojak je 13.57%.



Slika 3. Prikaz recepata sa vinom po poreklu

Sa slike 3 vidimo da fransuska, kineska i italijanska kuhinja imaju najviše specijaliteta sa vinom. Kako su Francuska i Italija tradicionalne zemlje vina, ne čudi nas što je u receptima iz tih država vino najzastupljenije kao sastojak. Takođe primetimo da tajlandska i meksička kuhinja ne koriste ovaj dodatka u receptima.

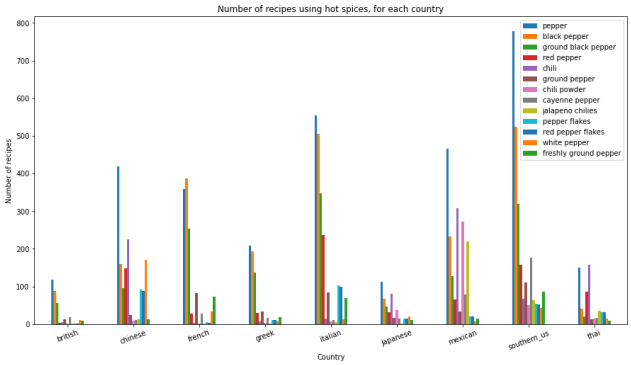
E. Analiza recepata koji koriste ljute začine u ishrani

Analizom smo uočili da svaki recept ima u sebi neki ljuti začin a u sledećoj tabeli 2 su prikazana 3 najčešća sastojka:

Tabela 2: Prikaz 3 najčešće korišćena ljuta začina

Ljuti začin	***Broj recepata***
Biber	3164
Crni biber	2199
Mleveni crni biber	1401

Međutim, imamo jos ljutih začina kao dodatke u ishrani. Prikazaćemo ih sve na slici 4 gde ćemo videti za svaki predeo koji je najzastupljeniji sastojak u ishrani.



Slika 4. Prikaz recepata za svaki ljuti začin po poreklu

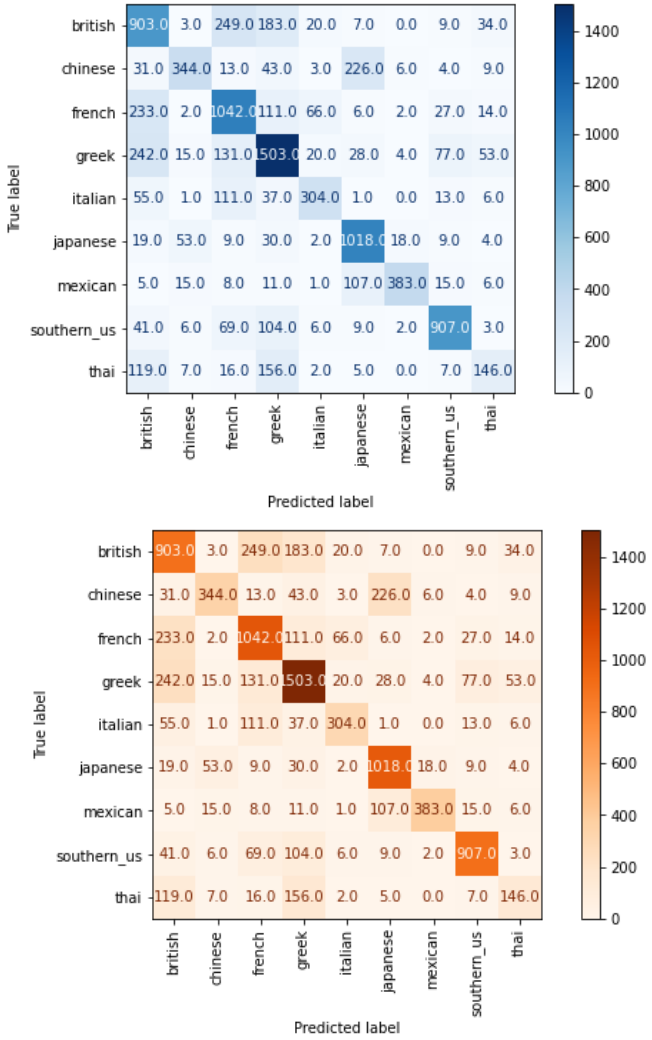
Možemo da uočimo da u južno-američkom i italijanskom predelu hrana je veoma začinjena kao i da se najmanje ljutih začina stavlja u britanskoj, japanskoj i tajlandskoj kuhinji.

III. KLASIFIKATOR METODOM K NAJBLIŽIH SUSEDA (KNN)

Podelili smo bazu podataka na dva podskupa. Jedan predstavlja skup za trening a drugi skup je za test. Uzimamo 10% uzoraka od ukupnih za test.

A. Određivanje optimalnih parametara i matrica konfuzije nakon unakrsne validacije

Potrebna su dva parametra za primenu kNN klasifikatora, to su k - broj najbližih suseda koji se posmatraju pri odlučivanju i m – metrika računanja rastojanja. Za parametar k korišćeni su 1,5 i 10 dok su za parametar m korišćeni jaccard i dice. Za rezultat dobijena je najbolja tačnost u četvrtoj iteraciji gde su parametri k=10 i m=jaccard. Za određivanje optimalnih parametara korišćena je funkcija StratifiedKFold sa pet particija. Korišćen je i drugi način za određivanje optimalnih parametara. Na ovaj način je isto korišćena funkcija StratifiedKFold sa pet particija, razlika je sto je na ovaj način ručno ispisana unakrsna validacija. Oba načina su dala isti rezultat da je najbolja tačnost u iteraciji broj 4.



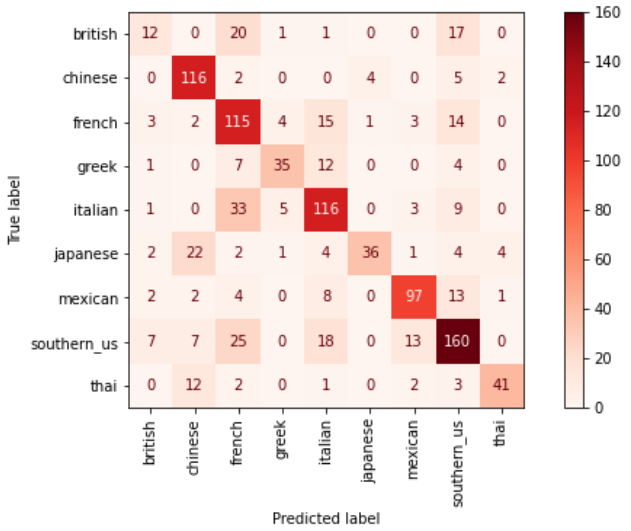
Slika 5. Matrice konfuzije nakon unakrsne validacije

Na slici 5 vidimo dve matrice konfuzije, plava matrica je dobijena na prvi način a narandžasta je dobijena na drugi način. Obe matrice su rezultati dobijeni u četvrtoj iteraciji a načini dobijanja su prethodno objašnjeni. Primetićemo da je rezultat isti kao i da veliki broj uzoraka nije ispravno klasifikovan. Nepravilnost uočavamo kod tajlandske kuhinje koja ima 146 recepata dok je 156 tajlandskih recepata klasifikovano kao grčkih recepata. Iz prethodne analize zaključili smo da je ukupan broj recepata u ove dve države približan, kao najčešći sastojak

im je so i u slaboj meri obe kuhinje koriste ljute začine. Ono što još možemo da zaključimo jeste da 344 recepta potiče iz kineske kuhinje a da je 226 recepata japanske kuhinje. Razlozi su ne toliko sličnost među sastojcime već skoro duplo više recepata ima kineska kuhinja od japanske.

B. *Matrica konfuzije nad test skupom*

Sledi nam obučavanje modela nad celim skupom za trening i predikcija test skupa. Koristićemo funkciju KneighborsClassifier sa parametrima koje su bili u iteraciji broj četri sa najboljom tačnošću, a to su k=10 i m=jaccard.



Slika 6. Matrice konfuzije nad test skupom

Na slici 6 možemo videti konačnu matricu konfuzije u kojoj uočavamo greške prilikom klasifikacije britanskih recepata kojih ima 12 dok je 20 britanskih recepata klasifikovano kao francuskih i 17 kao južno-američkih. Ono što primećujemo jeste da za svaki predeo možemo uočiti da nema velikih promašaja jer smo dobili većinom 0 za uzorke.

Tabela 3: Tačnost po poreklu posle testiranja test skupa

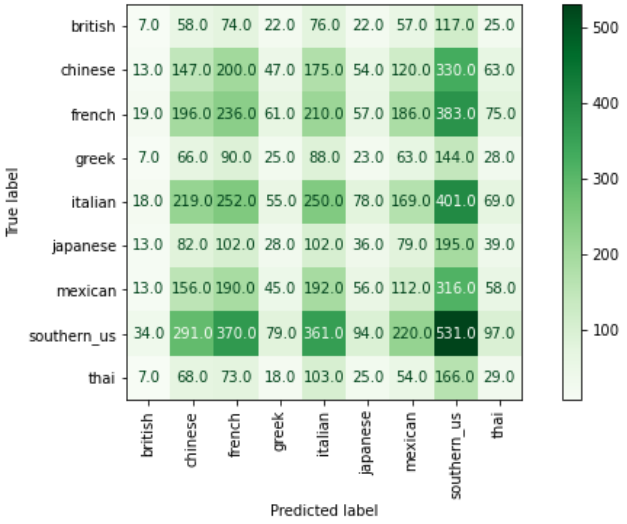
Poreklo	***Tačnost po klasi***	***Tačnost po klasi – finalni model***
Britansko	0.953	0.896
Kinesko	0.943	0.874
Francusko	0.868	0.907
Grčko	0.963	0.730
Italijansko	0.887	0.734
Japansko	0.954	0.782
Meksičko	0.957	0.656
Južno-američko	0.869	0.899
Tajlandsko	0.978	0.789

U tabeli 3 je prikazana tačnost po klasi za svaki predeo za matricu konfuzije i finalni model. Primećujemo da Grčka ima najveću razliku u tačnosti te da je tačnost opala sa 96.3% na 73%. Najveću tačnost ima Francuska sa 86.8%. Procenat pogođenih uzoraka je 86.8%.

IV. MAŠINE NA BAZI VEKTORA NOSAČA (SVM)

A. *Određivanje optimalnih parametara i matrica konfuzije nakon unakrsne validacije*

SVM algoritam se koristiti za probleme klasifikacije i regresije i ima za cilj da stvori najbolju granicu odlučivanja. Za ovaj algoritam određujemo regularizacioni parametar C koji određuje toleranciju na pogrešnu klasifikaciju, kernel koji povećava dimenzionalnost zbog bolje razdvojivosti i parametar decision_function_shape koji predstavlja način donošenja odluke kod višeklasnih problema. StratifiedKFold je funkcija koju koristimo, isto kao i kod kNN klasifikatora. Kao rezultat funkcije dobijamo da je najbolja tačnost u iteraciji 2 gde je parametar C=1, kernel=rbf i decision_function_shape=ovr.

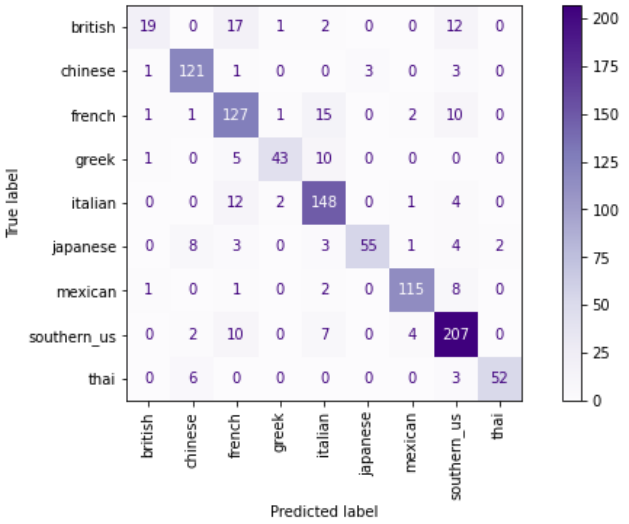


Slika 7. Matrice konfuzije nakon unakrsne validacije

Sa slike 7 uočavamo da tajlandska kuhinja sada ima 68 recepata od kojih je sada samo 18 grčkih. Kao što nam boje na matrici prikazuju imamo velika odstupanja, svaka od kuhinja ima veliki broj recepata neke druge kuhinje, neke čak i znatno veći broj nego svojih recepata. Ovde možemo da izdvojimo samo južno-američke recepte kojih ima 531, vidimo da je 370 ovih recepata u francuskoj kuhinji i 361 u italijanskoj kuhinji.

B. *Matrica konfuzije nad test skupom*

Prilikom obučavanja modela ovde je iskorišćen ceo trening skup. Parametri koji su korišćeni su oni koji su davali najbolju tačnost u iteraciji 2, što je već pomenuto. Konačna matrica prikazana na narednoj slici 8 nam govori da je naš model dobro obučen, pošto su većinom nule u matrici. Možemo izdvojiti britanske recepte kojih ima 19 od kojih je 17 u francuskoj kuhinji i 12 u južno-američkoj.



Slika 8. Matrice konfuzije nad test skupom

V. POREĐENJE REZULTATA KLASIFIKATORA

Nakon analize kNN i SVM klasifikatora možemo da uporedimo dobijene rezultate mera uspešnosti.

Tabela 4: Mere uspešnosti kNN i SVM klasifikatora po klasama

Poreklo	*kNN/SVM*	*Tačno.*	*Osetlj.*	*Prec.*
Britansko	kNN	0.896	0.064	0.067
Britansko	SVM	0.989	0.852	0.962
Kinesko	kNN	0.874	0.068	0.075
Kinesko	SVM	0.977	0.723	0.948
Francusko	kNN	0.907	0.050	0.049
Francusko	SVM	0.965	0.372	0.826
Grčko	kNN	0.730	0.154	0.153
Grčko	SVM	0.945	0.886	0.791
Italijansko	kNN	0.734	0.157	0.144
Italijansko	SVM	0.925	0.808	0.721
Japansko	kNN	0.782	0.122	0.117
Japansko	SVM	0.976	0.937	0.876
Meksičko	kNN	0.656	0.212	0.213
Meksičko	SVM	0.936	0.900	0.824
Južno-američko	kNN	0.899	0.058	0.063
Južno-američko	SVM	0.981	0.728	0.914
Tajlandsko	kNN	0.789	0.100	0.104
Tajlandsko	SVM	0.981	0.905	0.934

Iz tabele primetimo da SVM klasifikator ima veće vrednosti za svaku državu za tačnost, osetljivost i preciznost. U oba klasifikatora preciznost ima najveću razliku izmedju ova dva klasifikatora. Najmanja preciznost za kNN klasifikator je zabeležena za Francusku 4.9% a najveću preciznost primetimo kod britanske kuhinje za SVM klasifikator i iznosi 96.2%. Najbolje klasifikovane recepte za kNN prema preciznosti ima meksička kuhinja što znači da od 100 predviđenih meksičkih recepata otprilike 21 i potiče stvarno iz Meksika.

U narednoj tabeli 5 ćemo prikazati mere uspešnosti klasifikatora na globalnom nivou.

Tabela 5: Mere uspešnosti kNN i SVM klasifikatora

Mera uspešnosti	*kNN*	*SVM*
Procenat pogođenih uzoraka	0.688	0.839
Preciznost mikro	0.688	0.839
Preciznost makro	0.707	0.866
Osetljivost mikro	0.688	0.839
Osetljivost makro	0.640	0.790
F mera mikro	0.688	0.839
F mera makro	0.658	0.815

Primetimo mikro i makro mere u tabeli, razlika je što se za računanje mikro mere prvo sumiraju sve komponente po TP,TN,FP,FN pa se na njih primenjuju formule za mere a makro mere se računaju za svaku klasu pa se traži njihov prosek. Pored ove dve mere imamo i F meru koja predstavlja harmonijsku sredinu između preciznosti i osetljivosti.

Kao i u prethodnoj tabeli vidimo da SVM ima veće vrednosti za svaku od mera uspešnosti u tabeli. Sada na osnovu ovih analiza zaključujemo da je za ovaj skup podataka recepata bolje izabrati SVM klasifikator. Kako SVM ima bolju predikciju podataka nad većim brojem klasa otporniji je i na autlajere.

Prednost kNN modela je to što nema proces obuke modela, međutim klasifikacija je trajala dugo zbog velikog broja podataka u skupu podataka.

Rezultati su nam pokazali da je procenat pogođenih uzoraka najveći kod SVM klasifikatora. Značajne razlike između ova dva klasifikatora smo uočili u matricama konfuzije a to je posledica različitih pristupa koje koriste algoritmi.