Klasifikacija recepata na osnovu odsustva/prisustva određenih sastojaka

Vladimir Trpka, IN41-2018, vtrpka@gmail.com

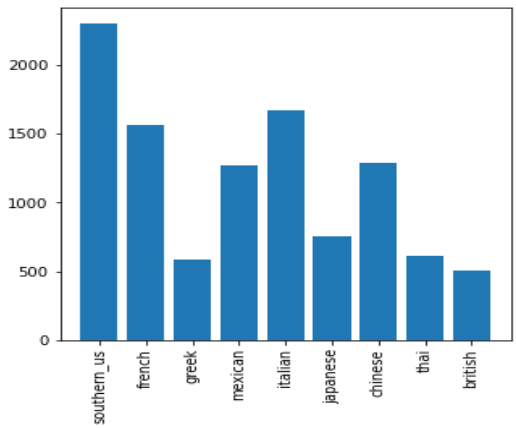
I. Uvod

Izveštaj se bavi analizom podataka o raznim receptima koji potiču iz nekoliko različitih zemalja. U drugom delu izveštaja vršena je klasifikacija uzoraka na osnovu sastojaka koje sadrže .

Baza podataka sadrži 152 obeležja i 10566 uzoraka. Prvo obeležje u bazi("Unnamed: 0") je numeričko sadrži različite numeričke vrednosti za svaki uzorak. Obeležja su razni sastojci i predstavljaju podatak o prisustvu(označeno sa 1) ili odsustvu(označeno sa 0) nekog sastojka. Poslednje obeležje u bazi("country") predstavlja klasu kojoj uzorak pripada. Baza je podeljena na trening I test skup. U trening skupu se nalazi 9509 uzoraka dok se kod test skupa nalazi 1057 uzoraka.

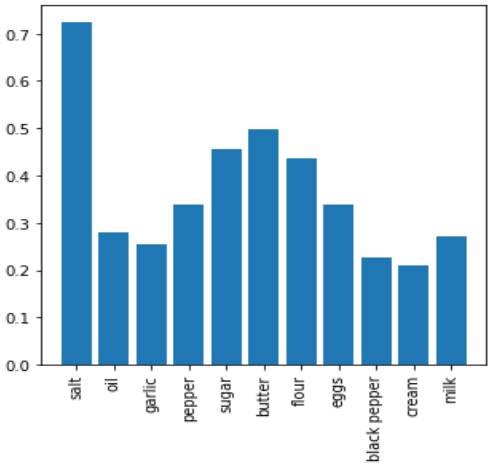
II. Analiza podataka

Obeležje "Unnamed: 0" je uklonjeno iz baze jer sadrži različite vrednosti za svaki uzorak u bazi i na osnovu toga nije relevantno za dalju analizu. Trening i test skup ne sadrže nedostajuće vrednosti. U trening skupu broj recepata po klasama je sledeći: southern_us(2073), french(1408), greek(528), mexican(1147), italian(1503), japanese(679), chinese(1162), thai(551), british(458). Dok je u test skupu : southern_us(230), french(157), greek(59), mexican(127), italian(167), japanese(76), chinese(129), thai(61), british(51).



Sl. 1. Broj recepata po državama iz originalnog skupa

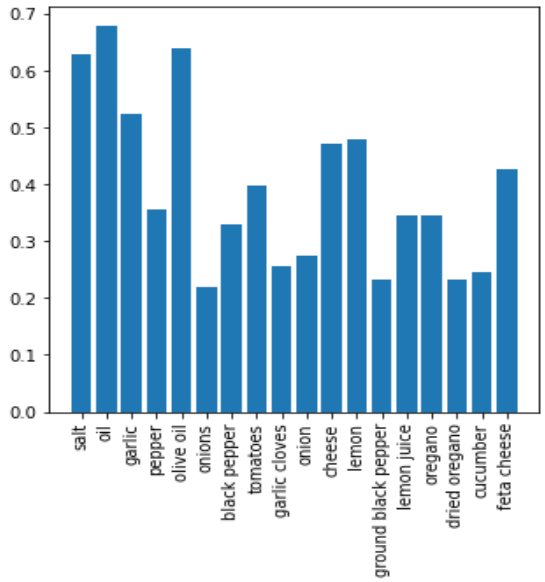
Takođe možemo da vidimo da su najzastupljeniji sastojci u Državama juga SAD-a(southern_us), so, ulje, šećer, brašno, mleko, jaja a ti sastojci su potrebni za pravljenje testa. Pa možemo zaključiti da se u toj državi ljudi hrane nezdravo i da ima puno gojaznih stanovnika.



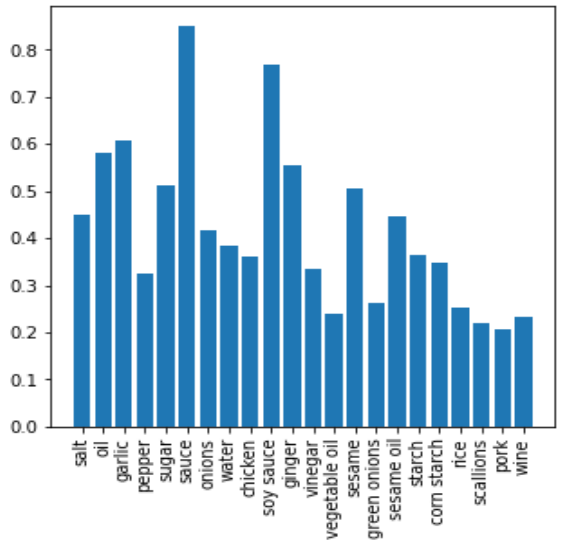
Sl. 2. Najzastupljeniji sastojci u Državama juga SAD-a (southern_us)

Najdominantniji sastojci država Evrope(greek, french, british, italian) i Amerike(southern_us, mexican) su: so, ulje, biber.

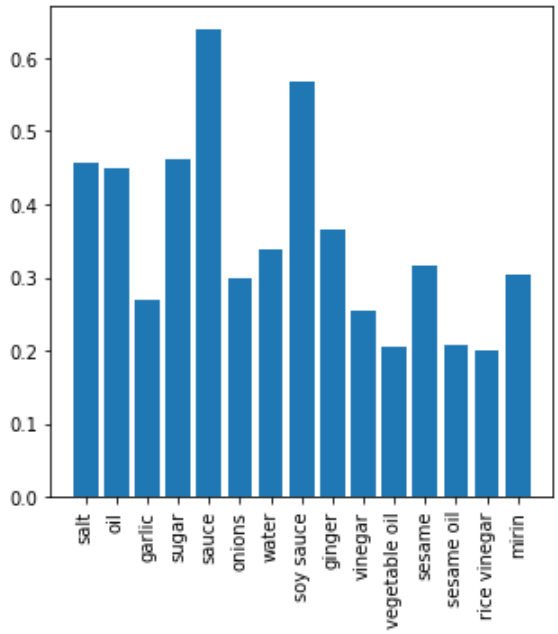
Dok su u zemljama Azije(thai, chinese, japanese) najdominantniji sastojici sos i soja sos, dok su sastojci so i ulje takođe zastupljeni ali nisu najdominantniji.



Sl. 3. Najzastupljeniji sastojci u Grčkoj(greek)



Sl. 4. Najzastupljeniji sastojci u Kini(chinese)



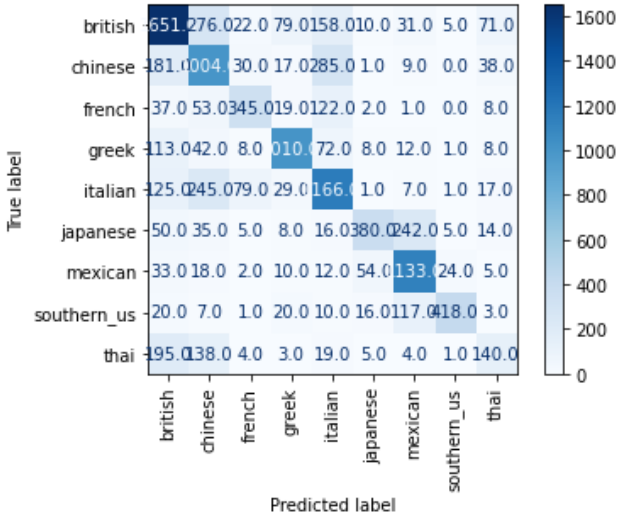
Sl. 5. Najzastupljeniji sastojci u Japanu(japanese)

III. KLASIFIKATORI

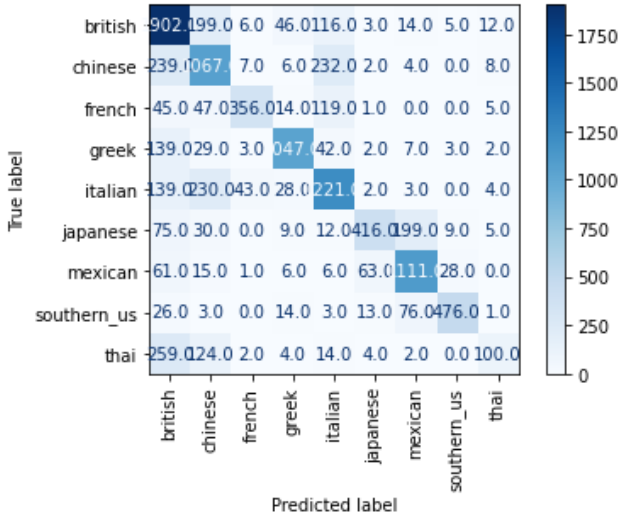
A. *Određivanje optimalnih parametara*

Unakrsnom validacijom sa podelom na tri podskupa i očuvanjem odnosa zastupljenosti svake klase podataka u svakom od tri podskupa, dobijeno je da su optimalni parametri za kNN klasifikator sledeći: celobrojni parametar k tj. broj najbližih suseda je 10, metrika je jaccard. Prosečna osetljivost klasifikatora na osnovu izračunate matrice konfuzije, dobijene akumulacijom matrica iz svake od tri iteracije unakrsne validacije je 64%. Za klasu southern_us osetljivost je: 71.6%,za klasu french: 64.1%, za klasu greek :58.7%, za klasu mexican: 79.2% ,za klasu italian: 69.8%, za klasu japanese :50.3%, za klasu chinese: 87.7%, za klasu thai : 68.3%, za klasu british : 27.5%.

Za klasifikator SVM optimalni parametri su : regularizacioni parametar 1, vrsta kernela „rbf“, sa „ovo“ pristupom. Njegova prosečna osetljivost je: 68.2% što je za 4% bolje u odnosu na kNN klasifikator. Prosečna osetljivost po klasama je: za klasu southern_us osetljivost je: 72.3%, za klasu french je: 63%, za klasu greek je:64.5%, za klasu mexican je: 82.8%, za klasu italian: 67.9%, za klasu japanese je: 60.7%, za klasu chinese je: 82.1%, za klasu thai je: 79.2%, za klasu british je: 41.4%



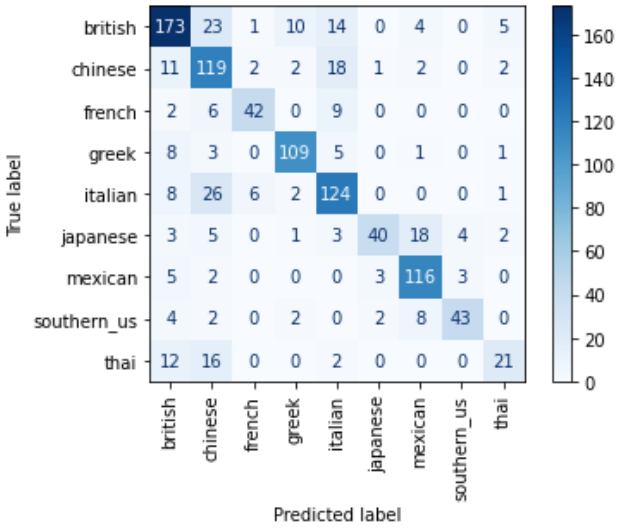
Sl. 6. Matrica konfuzije nakon unakrsne validacije kod kNN klasifikatora



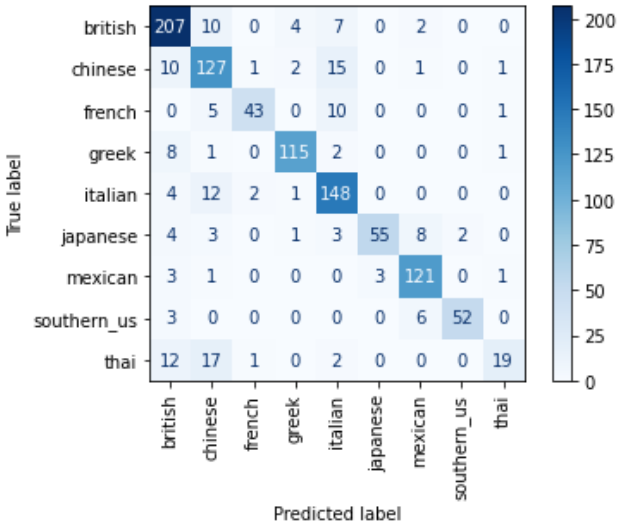
Sl. 7. Matrica konfuzije nakon unakrsne validacije kod SVM klasifikatora

B. *Rezultati testiranja test skupa*

Klasifikatori kNN i SVM sa prethodno izabranim parametrima obučeni su na trening skupu i onda testirani na test skupu. Bolju osetljivost je postigao SVM klasifikator, osetljivost od 83.9% dok je kNN klasifikator postigao osetljivost od 74.4%.



Sl. 8. Matrica konfuzije na test skupu kod kNN klasifikatora



Sl. 9. Matrica konfuzije na test skupu kod SVM klasifikatora

C. Upoređivanje klasifikatora

Procenat pogodjenih uzoraka	0.74
preciznost mikro	0.74
preciznost makro	0.76
osetljivost mikro	0.74
osetljivost makro	0.70
f mera mikro	0.74
f mera makro	0.72

Tabela 2: Performanse SVM klasifikatora

Procenat pogodjenih uzoraka	0.83
preciznost mikro	0.83
preciznost makro	0.86
osetljivost mikro	0.83
osetljivost makro	0.79
f mera mikro	0.83
f mera makro	0.81

Tabela 2: Performanse SVM klasifikatora

Mikro osetljivost za SVM klasifikator je 83% dok je za kNN klasifikator mikro osetljivost manja za 9%.

Makro osetljivost za oba klasifikatora je za 4 % manja od njihove mikro osetljivosti.