Klasifikacija recepata po državama na osnovu prisustva/odsustva određenih sastojaka

Ilija Rakočević, IN59/2018, rakocevicilija7@gmail.com

I. Uvod

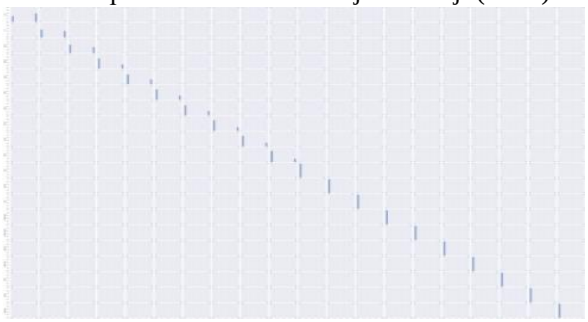
U izveštaju će biti opisano ispitivanje skupa podataka koji je ispunjen podacima o sastojcima iz određenih recepata. Pre nego što počnemo sa analizom treba skrenuti pažnju na algoritme koji će biti obrađeni i klasifikaciju uopšte. Klasifikacija se može definisati kao proces predviđanja pripadnosti određenoj klasi neke kategoričke promenljive. Ovo se radi izgradnjom modela zasnovanog na jednoj ili više numeričkih i/ili kategoričkih promenljivih (prediktori, obeležja..). Spada u grupu nadgledanog učenja i predstavlja pandan algoritmu nenadgledanog učenja pod nazivom „Klasterizacija“ (Clustering). Algoritmi klasifikacije koji će biti pokriveni su: KNN klasifikator i RF klasifikator.

II. Baza podataka

Skup je popunjen podacima o 10566 uzoraka i 152 obeležja pri čemu uzorci predstavljaju recepte, a obeležja sastojke. Vrednosti obeležja su binarnog tipa, osim prvo i za poslednje obeležje; tačnije ili imaju vrednost 0 ili vrednost 1. Vrednost 0 predstavlja odsustvo određenog sastojka iz datog recepta, dok vrednost 1 predstavlja njegovo prisustvo. Kada se uradi pregled jednog reda (uzorka) dobiju se svi sastojci koji su potrebni za recept. Poslednje obeležje, pod nazivom *„country“* je kategoričko, i njegove vrednosti su iz skupa (*'british', 'chinese', 'french', 'greek', 'italian', 'japanese', 'mexican', 'southern_us', 'thai'*). Prethodno definisani skup predstavlja državu kojoj recept pripada. Pored obeležja *„country“*, nebinarne i to numeričke vrednosti ima i prvo obeležje koje nema naziv. Kasnije kada je prvo obeležje prebačeno u pandas DataFrame dobija naziv „Unnamed: 0“. Skup nije balansiran što možemo videti sumiranjem svih uzoraka koji pripadaju istoj državi. Dobijamo sledeći rezultat: („southern_us“ : 2303; „french“ : 1565; „greek“ : 587; „mexican“ : 1274; „italian“ : 1670; „japanese“ : 755; „chinese“ : 1291; „thai“ : 612; „british“ : 509).

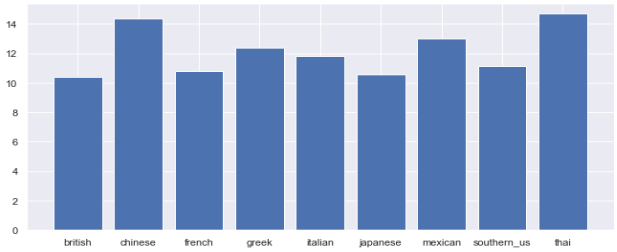
III. Analiza podataka

Analizu podataka započinjemo učitavanjem samog skupa podataka korišćenjem funkcije *„read_csv()“*. Pre nego što krenemo sa bilo kakvom dubljom analizom proveravamo da li ima nekih nedostajućih vrednosti. To uspevamo da vidimo korišćenjem funkcije *„isna()“*. Možemo primetiti da skup uopšte nema nedostajućih vrednosti što je zaista pogodno jer će nam svaka analiza i predikcija u budućnosti biti verodostojnije. Možemo primetiti da nam kolona *„Unnamed: 0“* (prethodno pomenuto prvo obeležje bez naziva) i nije od velikog značaja, jer za svaki uzorak ima drugačiju vrednost. Iz toga sledi da ga možemo obrisati jer neće doprinositi ni daljoj analizi ni klasifikatorima. Za brisanje smo iskoristili *„loc()“* i *„columns.str.match()“* na takav način da ostavljaju u skupu sve kolone osim one pod nazivom *„Unnamed: 0“*. Korišćenjem *„pairplot()“* funkcije iz *seaborn* paketa primećujemo da obeležja praktično uopšte nisu u međusobnoj korelaciji (Sl. 1.).



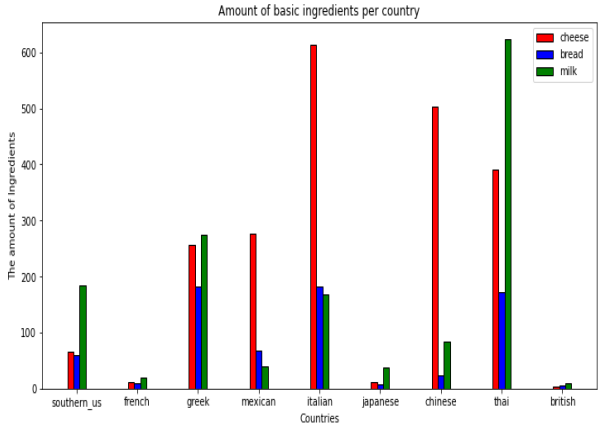
Sl. 1. Prikaz korelacije između obeležja

Na sledećoj slici (Sl. 2.) možemo videti grafik koji prikazuje prosečan broj sastojaka po receptu za svaku državu. Dolazimo do zaključka da države Kina, Tajland i Meksiko imaju najviše sastojaka po receptu. Mogli smo i pretpostaviti da će se Meksiko naći tu jer je njihova hrana zaista začinjena. da ržava koja koristi najviše začina u svojim receptima je Neke od analiza mogu biti ilustrovane korišćenjem slike ili tabele formatirane kao Tabela 1 i Sl. 1 u ovom tekstu. Dakle, svaka slika i tabela moraju imati i odgovarajući naziv koji jasno objašnjava šta se može videti na slici i/ili u tabeli.

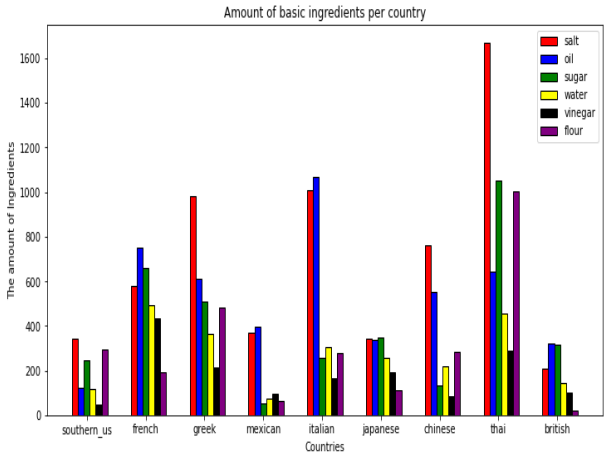


Sl. 2. Broj začina po receptu za sve države

Sada ćemo proverićemo da li ima razlike u konzumaciji nekoliko osnovnih namirnica za pripremanje hrane: *salt, oil, sugar, water, vinegar, flour*. Primećujemo da se osnovne namirnice u *„southern_us“* jako slabo konzumiraju. Moguć razlog je taj što je u Americi generalno zastupljena prerađena vestačka hrana, fastfood itd..., a pored toga retko ko ima vremena da sprema sve od početka. Samim tim će se kupovati delom gotovi proizvodi pa recept niko neće početi od osnovnih sastojaka. Italijanski recepti imaju mnogo veću upotrebu soli i ulja. Najverovatnije jer je u njihovoj kuhinji tradicionalno zastupljena testenina u svim oblicima koja zahteva ulje i so u pripremi. Vidimo da je obično belo brašno slabo kotirano na grafiku, a i ono se koristi u pripremi za testenine. Razlog za to je što italijani koriste specijalnu vrstu brašna za testenine, koje se razlikuje od obicnog belog brašna. (nazivaju ga brašno tipa *"00"* i primer je *AGUGIARO & FIGNA MOLINI* brašno). Japanska kihinja je uglavnom bazirana na morskim plodovima, pirinču itd... tako da ove namirnice nisu mnogo korišćene što i možemo primetiti na grafiku. U tajlandskim receptima su sve namirnice dosta zastupljene. Može biti razlog to što je Tajland jedna od siromašnijih zemalja i najjeftinije je priuštiti one osnovne namirnice od kojih će se proizvoditi hrana. Celokupnu iznad navedenu analizu možemo videti i
na grafiku (Sl. 3.).



Sl. 3. Zastupljenost osnovnih namirnica u receptima



Sl. 4. Zastupljenost nekih namirnica u receptima

Na slici iznad (Sl. 4.) možemo primetiti da je sir kao namirnica izuzetno korišćena u receptima svih zemalja.

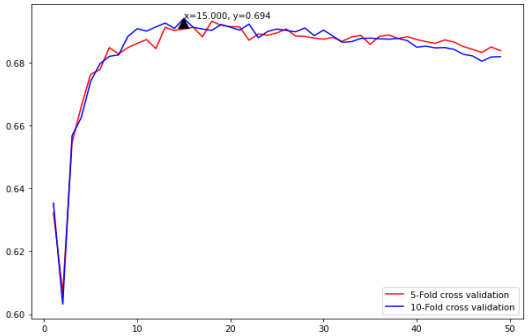
IV. FUNKCIJA EVALUACIJE

Funkcija evaluacije je pod nazivom *„evaluation_classif“* i ona računa binarnu matricu konfuzije za svaku klasu pojedinačno na sledeći način:
TP je vrednost gde se ista klasa preklapa u redu i koloni, odnosno na glavnoj dijagonali. *FP* predstavlja vrednost koja se dobija deljenjem *TP* sa vrednostima iz kolone u kojoj se *TP* nalazi. *FN:* predstavlja vrednost koja se dobija deljenjem *TP* sa vrednostima iz reda u kojem se TP nalazi. *TN:* je vrednost koja se dobija sumom svih vrednosti koje nisu u koloni i redu u kojoj je *TP*. Za računanje mera uspešnosti klasifikatora su korišćene još tri dodatne funkcije koje u sebi poziva funkcija *„evaluation_classif“*. Nazivi funkcija su *„TN_calculation“*, *„FN_calculation“* i *„FP_calculation“*. Sve tri kao parametre primaju matricu konfuzje i poziciju na osnovu kojih i računaju ono što je potrebno. Skup smo podelili na trening i test pomoću funkcije *„train_test_split()“* u kojoj smo odredili da *10%* uzoraka ide u test skup. Promenljive u kojima će biti smešteni trening i test skupovi su *X_train, Y_train, X_test* i *Y_test.*

V. KLASIFIKACIJE SA UKARSNOM VALIDACIJOM

A. *Odabir optimalnih parametara za KNN*

Što se tiče KNN klasifikatora, prvo ćemo pronaći optimalne parametre koje treba ubaciti u algoritam kako bi dao optimalno rešenje. Nećemo tražiti optimalnu metriku („*metric*“) jer vrednosti obeležja nisu realni brojevi, pa poređenje ne možemo vršiti klasičnim operacijama. Iz tog razloga se koristi *„jaccard“* metrika koja je pogodna kada su vrednosti binarne. Dodatni razlog zbog kog je uzet *„jaccard“* je taj što se dobijaju iste vrednosti i sa *„dice“* metrikom, tako da je bilo svejedno šta uzeti. Kako bi našli parametre korišćena je funkcija *„cross_val_score()“*. Funkcija kao parametar prima „*estimator“*, koji je u našem slučaju „*knn“*, podatke koji treba da se obuče („*X_train“*), podatke koje treba da prediktuje („*Y_train“*), parametar *cv* koji predstavlja na koliko foldova deli skup i „*scoring“* parametar koji predstavlja strategiju za procenu performanse cross validiranog modela na testnom skupu. Drugim rečima ako je vrednost *„scoring“* parametra postavljena na *„accuracy“* on će davati vrednost **tačnosti** za svaki test skup koji je testiran u cross validaciji. Na kraju, pomoću finkcije *„mean()“*, dobijemo srednju vrednost svih tačnosti i time smo dobili vrednost za **srednju tačnost** što nam je konačna mera uspešnosti klasifikatora sa više klasa.
Na kraju su optimalni parametri predstavljeni vizuelno na grafiku (Sl. 5.).



Sl. 5. Optimalni parametri za KNN klasifikator sa CV

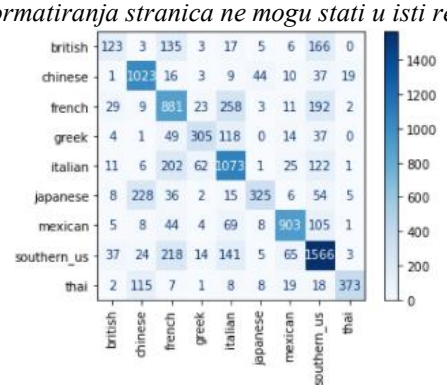
Optimalni parametri, prikazani na grafiku (Sl. 5.) su: Vrednost za k najbližih suseda je 15, u pitanju je 10-Fold unakrsna validacija pri čemu je početni međusobni odnos klasa očuvan, jer funkcija *„cross_val_score“* podrazumevano koristi „*StratifiedKFold“*. Metrika je kao što smo već rekli *„jaccard“*.

B. Odabir optimalnih parametara za RF

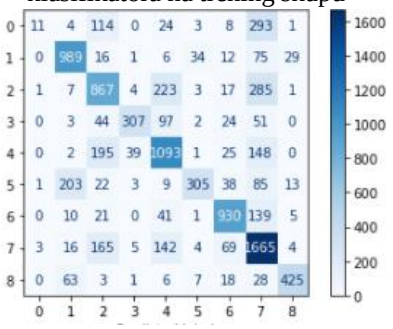
Što se tiče RF klasifikatora, optimalni parametri su pronađeni korišćenjem *„GridSearchCV()“* funkcije za 5-Fold i 10-Fold unakrsnu validaciju. Za parametre su postavljeni: broj stabala koji će se generisati, kriterijum podele, maximalna dubina do koje ide stablo, najmanji broj uzoraka da bi izvršio deljenje u čvoru, težine klasa... Kao optimalne parametre smo dobili: *„bootstrap“: „False“, „class_weight“: „None“, „criterion“: „gini“, „max_depth“: „50“, „min_samples_split“: „0.01“, „n_estimators“: „100“*.

C. Poređenje klasifikatora posle obuke na train setu

Ubacivanjem ovih vrednosti u algoritam KNN klasifikatora dobijamo prosečnu tačnost na osnovu izračunate finalne matrice konfuzije, dobijene akumulacijom ukupno deset matrica iz deset iteracija unakrsne validacije. Ta vrednost je *69.1%*. Istim procesom se kod RF klasifikatora stiglo do rezultata prosečne tačnosti od *~69.32%*. Dakle postoji blaga razlika od *0.22%* u tačnosti na trening skupu između klasifikatora, ali ipak na ogrmonim skupovima podataka i ovako mali procenat igra veliku ulogu. Matrice konfuzije možemo videti na slikama ispod (Sl. 6. i Sl. 7.). Svi rezultati mera za KNN i RF na trening skupu po klasama su u tabelama ispod (Tabele 1, 2, 3, 4). *Klase su podeljene u dve tabele jer zbog načina formatiranja stranica ne mogu stati u isti red.*



Sl. 6. Matrica konfuzije nakon unakrsne validacije KNN klasifikatora na trening skupu



Sl. 7. Matrica konfuzije nakon unakrsne validacije RF klasifikatora na trening skupu

Tabela 1: Mere uspešnosti po klasama KNN (trening)

	British	Chinese	French	Greek	Italian
Precision	0.5590	0.7219	0.5547	0.7314	0.6282
Accuracy	0.9545	0.9439	0.8702	0.9647	0.8880
Sensitivity	0.2685	0.8803	0.6257	0.5776	0.7139
Specificity	0.9892	0.9527	0.9127	0.9875	0.9206
F_score	0.3628	0.7933	0.5881	0.6455	0.6683

Tabela 2: Mere uspešnosti po klasama KNN (trening)

	Japanese	Mexican	Southern_us	Thai
Precision	0.8145	0.8526	0.6817	0.9232
Accuracy	0.9549	0.9579	0.8698	0.9780
Sensitivity	0.4786	0.7872	0.7554	0.6769
Specificity	0.9916	0.9813	0.9016	0.9965
F_score	0.6029	0.8186	0.7167	0.7811

Tabela 3: Mere uspešnosti po klasama RF (trening)

	British	Chinese	French	Greek	Italian
Precision	0.6875	0.7625	0.5991	0.8527	0.6660
Accuracy	0.9524	0.9494	0.8821	0.9711	0.8992
Sensitivity	0.0240	0.8511	0.6157	0.5814	0.7272
Specificity	0.9994	0.9631	0.9284	0.9940	0.9315
F_score	0.0464	0.8043	0.6073	0.6914	0.6952

Tabela 4: Mere uspešnosti po klasama RF (trening)

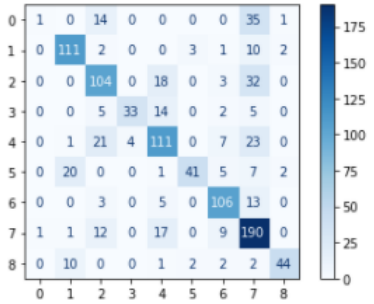
	Japanese	Mexican	Southern_us	Thai
Precision	0.8472	0.8150	0.6013	0.8891
Accuracy	0.9548	0.9549	0.8409	0.9811
Sensitivity	0.4491	0.8108	0.8031	0.7713
Specificity	0.9937	0.9747	0.8515	0.9940
F_score	0.5871	0.8129	0.6877	0.8260

Iz priloženih matrica konfuzija dolazimo do sledećih zaključaka: RF klasifikator je 114 uzoraka koji pripadaju klasi *„british“* proglasio za *„french“*, a KNN 135. Sa druge strane, RF skoro duplo više greši kada uzorke iz klase *„british“* proglasi za *„southern_us“*. Ista situacija je i sa klasom *„french“*. Manje greši kada proglašava uzorke klase *„french“* za *„italian“*, ali zato više promašuje

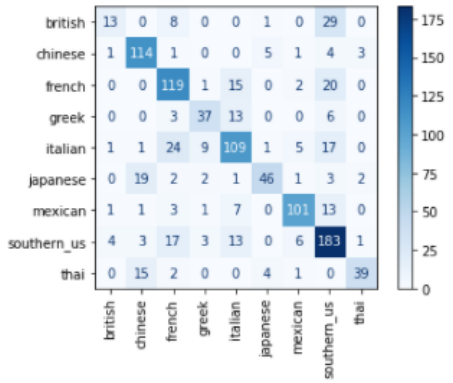
između *„french“* i *„southern_us“*. Za utorke klase *„chinese“* KNN definitivno nudi manje greške predikcije. Medjutim, RF bolje prediktuje kada su u pitanju uzorci iz klasa *„southern_us“* i *„thai“*. Primećujemo da oba klasifikatora prave ne malu grešku kod uzoraka koji pripadaju klasi *„chinese“*, a prediktovana je klasa *„japanese“*. Razlog toga može biti, što i sami znamo, da je ishrana i jednih i drugih državljana identična, sa istim specijalitetima, korišćenjem istih začina u jelima (morski plodovi, suši id...). To isto važi i za uzorke klase *„thai“*, koji su predvidjeni kao *„chinese“*, s tim što je RF klasifikator dosta bolje predvideo, ali opet ostaje ne mali broj losih predikcija.

VI. PROVERA NA FINALNOM TEST SKUPU

Oba klasifikatora su sa predhodno izabranim optimalnim parametrima obučeni na celokupnom trening skupu i nakon toga testirani na finalnom test skupu (onih izdvojenih 10 % na početku). Procenat pogođenih uzoraka KNN klasifikacije na test skupu je *72%*, dok RF klasifikator ima nešto niži procenat pogodaka koji iznosi *~70%*.



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



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

Tabela 5: Mere uspešnosti po klasama KNN (test)

	British	Chinese	French	Greek	Italian
Precision	0.6500	0.7450	0.6648	0.6981	0.6898
Accuracy	0.9574	0.9489	0.9072	0.9640	0.8987
Sensitivity	0.9574	0.8837	0.7579	0.6271	0.6526
Specificity	0.9930	0.9579	0.9333	0.9839	0.9449
F_score	0.3661	0.8085	0.7083	0.6607	0.6707

Tabela 6: Mere uspešnosti po klasama RF (test)

	British	Chinese	French	Greek	Italian
Precision	0.5	0.7762	0.6459	0.8918	0.6646
Accuracy	0.9517	0.9526	0.8959	0.9716	0.8940
Sensitivity	0.0196	0.8604	0.6624	0.5593	0.6646
Specificity	0.9990	0.9655	0.9366	0.9959	0.6646
F_score	0.0377	0.8161	0.6540	0.6874	0.6646

Tabela 7: Mere uspešnosti po klasama KNN (test)

	Japanese	Mexican	Southern_us	Thai
Precision	0.8070	0.8632	0.6654	0.8666
Accuracy	0.9612	0.9602	0.8684	0.9735
Sensitivity	0.6052	0.7952	0.7956	0.6393
Specificity	0.9887	0.9827	0.8887	0.9939
F_score	0.6917	0.8278	0.8887	0.7358

Tabela 8: Mere uspešnosti po klasama RF (test)

	Japanese	Mexican	Southern_us	Thai
Precision	0.8913	0.7851	0.5993	0.8979
Accuracy	0.9621	0.9526	0.8420	0.9791
Sensitivity	0.5394	0.8346	0.8260	0.7213
Specificity	0.9949	0.9688	0.8464	0.9949
F_score	0.6721	0.8091	0.6946	0.8000

Radi preglednosti, tabele koje prikazuju iste klase su stavljene jedna ispod druge. Tačnosti po klasama su uglavnom iste kod oba klasifikatora. Malo veća razlika je možda kod klase „french“ za koju KNN predviđa tačno sa verovatnoćom od *90.72%*, dok RF sa *89.59%* i kod klase „southern_us“, za koju KNN ima tačnost *86.84%*, a RF *84.20%*. Posmatranjem matrica konfuzije za test skup primećujemo sledeće: RF uopšte nije mešao druge klase sa klasom *„british“* tokom predikcije, tačnije, pogrešno je predvideo samo jedan uzorak koji pripada klasi *„southern_us“* i proglasio ga *„british“* uzorkom. Kod KNN klasifikatora nije takav slučaj. Pogrešno je predivdeo 4 uzorka iz *„southern_us“* klase i još jedan iz *„chinese“* klase i proglasio da pripadaju klasi *„british“*.

VII. CROSS VALIDACIJA VS FINALNI TEST

Kod KNN klasifikatora prosečna tačnost na test skupu je za 3% veća nego nakon unakrsne validacije. Što se tiče pojedinačnih klasa, tačnost nakon unakrsne validacije i tačnost na test skupu su približne (+- 0.5). Najveća razlika jeste u tačnosti kod klase *„french“*, u kojoj je tačnost na test skupu veća za *3 %*. Kada je reč o RF klasifikatoru, prosečna tačnost je veća na test skupu za 1%. Vrednosti u pojedinačnim klasama su približne (+- 0.5).

VIII. POREĐENJE PERFORMANSI KLASIFIKATORA

Tabela 9: Performanse KNN klasifikatora

Procenat pogodjenih uzoraka	0.70104
preciznost mikro:	0.7010
preciznost makro:	0.7391
osetljivost mikro:	0.7010
osetljivost makro:	0.6320
f mera mikro:	0.7010
f mera mikro:	0.6484

Tabela 10: Performanse FR klasifikatora

Procenat pogodjenih uzoraka	0.70104
preciznost mikro:	0.7010
preciznost makro:	0.7391
osetljivost mikro:	0.7010
osetljivost makro:	0.6320
f mera mikro:	0.7010
f mera mikro:	0.6484