Analiza i modelovanje klasaifikatora za skup podataka recipes.csv

Aleksandra Borisavljević, IN17/2018, aleksandra.borisavljevic99@gmail.com

I. Uvod

Tema izveštaja je analiza i formiranje klasifikatora, čija uloga je da za dati recept odredi iz koje države potiče. Svaku državu karakterišu određeni začini. Neki začini su više zastupljeni dok su neki manje.Upravo klasifikator treba da nauči pravilnosti koje postoje i na osnovu njih da bude u mogućnosti da na pravilan način klasifikuje novodospeli uzorak, tj. recept u postojeću klasu, odnosno državu.

II. Baza podataka

Skup podataka koji se obrađuje je recipes.csv. Sadrži 10566 uzoraka i 150 obeležja. Jedan uzorak predstavlja recept dok obeležje predstavlja jedan od sastojaka koji je moguće da se pojavi u receptu. U skupu podataka se nalaze recepti iz 9 različitih država i to su: Kina, Japan, Tajland, Meksiko, Grčka, Italija, Britanija, Francuska i Južna Amerika.

III. Analiza podataka

U nastavku sledi analiza pojedinih začina i njihova prisutnost u određenim državama, kao i poređenje određenih država.

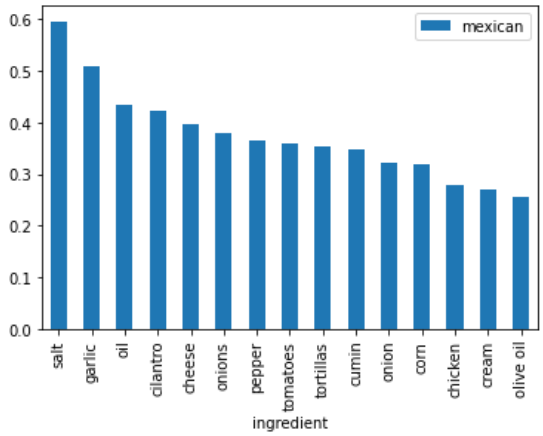
U skupu podataka ne postoji ni jedna null vrednost. Vrednost za obeležje može biti 0 ili 1, što nam ukazuje da li je sastojak prisutan ili nije, u datom receptu.

Japan, Tajland i Kina su države na Azijskom kontinentu i to je svakako razlog zašto su mnogi sastojci koji se koriste slični. Pirinač je jedan od njih i on čini oko 30% recepata iz Tajlanda i Kine. Origano se u ove tri države najmanje koristi. Takođe zanimljivo je primetiti da je korišćenje đumbira značajno izraženo u ove tri države. Isto važi i za susam i soja sos.

Poznato je da su Amerikanci veliki ljubitelji slanine, pa je i u osom skupu podataka Južna US zabeležena kao najveći konzumator slanine.

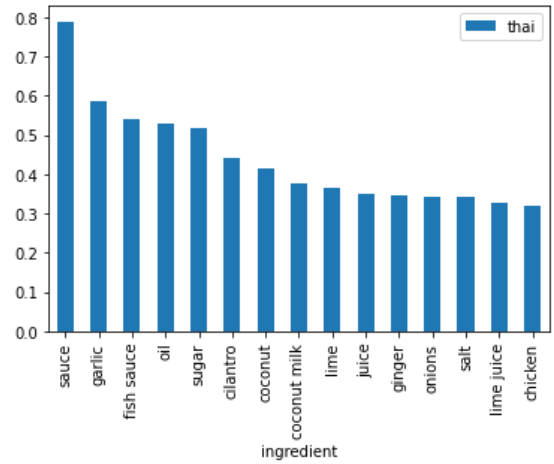
Dok je Meksiko naširoko poznat po svojim pikantnim jelima (Slika 1). Stoga kad se ispitiju zastupljenost čilija u prahu, najveća zastupljenost je u Meksiku, oko 21%, dok je u svim ostalim državama oko 1% ili manje. Zanimljivo je primetit da su na prvih četri mesta po korišćenju, odmah posle soli, beli luk, ulje persun i sir.

Što se tiče ulja, najzastupljenije je u jelima u Grčkoj, čak 67%, dok je zanimljivo primetiti da je najmnja zastupljeno u Britaniji i Južnoj US, oko 20%. Interesantno je primetiti da je zastupljenost šećera izrazito mala u Grčkoj, samo 9%, dok je u Tajlandu I Kini zastupljen u više od 50% jela.



Slika 1:Top 15 sastojaka u Meskiku

Takođe kokosovo mleko se u svim državama pojavljuje u 1% recepata ili manje, dok je u Tajlandu prisutan u 37% recepata. To nije neočekivan podataka, ako znamo činjenicu da je Tajland deveta država po proizvodnji kokosa. Takođe pored kokosovog mleka veoma je prisutam i kokos (Slika 2).



Slika 2: Top 15 sastojaka u Tajlandu

U Grčkoj je veoma često korišćen i limunov sok, kako je to mediteranska zemlja i mnogo jela koje se pripremaju su morsog porekla, receptima je uvek neophodan limun kao dodatak.

Neki od sastojaka koji su veoma korišćeni u svim državama su brašno i so.

Najmanje vode u svojim receptima imaju Grci, Amerikanci, Italijani i Meksikanci. Što može da ukaže na činjenicu da se u ovim državama manje pripremaju kuvana jela i da su mnogo češća pečena jela.

Ispitivanjem došlo se od zaključka da su odnosi broja uzoraka u klasama različiti (Tabela 1). Odnosno neravnomerna raspoređenost uzoraka po klasama je izražena i izazvaće probleme koje je neophodno rešiti primenom različitih metoda.

Tabela 1: Broj uzoraka po državi

Država	Broj uzoraka
British	509
Chinese	1291
French	1565
Greek	587
Italian	1670
Japanese	755
Mexican	1274
Southern_us	2303
Thai	612

Pre početka obuke modela neophodno je podeliti podatke na trening i test podatke. Trening skup podataka će biti korišten za dalju obuku modela dok će test podaci biti skriveni od modela i samog stvaraoca modela. Oni će biti primenjeni na samom kraju, kada budu dobijeni optimalni parametri za klasifikator. Na taj način sprečiće se naknadno podešavanje parametara spram uzoraka u test skupu, odnosno preobučavanja modela.

IV. KLASIFIKACIJA-NEURONSKE MREZE

A. Uvod

Za obuku biće korištena unakrsna validacija. Metod koji omogućava iskorišćenost svih uzoraka u testiranju. Takođe ovaj metod pordazmeva kreiranje više modela. Broj modela koji se kreira odgovara broju napravljanenih podskupova. U svakoji iteraciji koristi se n-1 podskup za obuku i 1 podskup za testiranje. Kako bi se napravio što bolji model, neophodno je obratit pažnju da pri podeli na podskupove budu održani klasni odnosi.

B. Procena parametara

Za prvi model korištena je neuronska mreža kao klasifikator. Iako je prethodnim ispitivanje utvrđeno da je prisutna neizbalansiranost između klasa, za početak ta činjenica će biti zanemarena, kako bi se bolje video napredak modela. Metodom unakrse validacije skup podataka je podeljen na 3 dela. Struktura neuronske mreže je podeljena na 3 sloja sa po 70 neurona. Model pravi velike greške, kao što se moglo očekivati. Nebalansiranost je velika, pa je kod manjih klasa napravljena veća greška.

Prosečna tačnost na nivou pojedinačnih klasa je velika zato što dominiraju pojedine klase. Stoga je bolja odluka ispitati osetljivost. Osetljivosti su mnogo manje vrednosti.

Postoje različite mere uspešnosti klasifikatora. Jedna od boljih odabira, kada je prisutan problem neravnomerne raspoređeneosti uzoraka po klasama, je osetljivost. Osetljivost predstavlja udeo ispravnio klasifikovanih uzoraka iz klase pozitiva.

Sada slede metode koje dovode do bolje izbalansiranosti između pojedinih klasa. To su metode undersamling i upsampling.

Undersampling je koristan kada postoji klasa koja je u velikoj meri izraženija od ostalih, stoga se omogućava odabiranje manje uzoraka koji pripradaju dominantnoj klasi. Mana ovog pristupa je što se na ovaj način može desiti da se odstrani velika količina podataka koja je prikupljena.

Sa druge strane postoji upsampling koji je koristan kada postoji klasa koja je malobrojnija. Tada se radi češće odabiranje uzoraka iz malobrojnijih klasa. Ovim se sprečava problem koji može da nastane korišćenjem prethodno navedenog načina, međutim ovde nastaje problem postojanja kopija. Još jedna mana ovog pristupa je da se može desiti da se u trening skupu i u test skupu nađu isti uzorci. Međutim postoji način i da se ovo prevaziđe. Potrebno je da se trening skup duplira, a da test ostane neopromenjen.

Ima 5 država koje imaju veći broj uzoraka. Ne bi bio najbolji način da se uradi undersampling za svaku državu od tih 5, pošto ih je i više, time bi se izgubilo mnogo prikupljneih podataka. Neki od načina koji deluje kao najprikladniji jeste da se Južna US kao kategorija sa najviše uzoraka smanji za jednu trećinu, a da se pritom za ostale države, koje imaju izrazito manji broj uzoraka, poveća količina uzoraka primenjujući upsamling. Time se neće izbaciti velika količina podataka, a pritom se neće ubaciti veliki broj kopija.

Prvo je primenjen updsampling za 4 države sa manjim brojem uzoraka, gde se broj uzoraka povećao 2 puta. Kada se na takvim podacima ponovo izvršila klasifikacija metodom unakrsne validacije (Tabela 2), procenat tačno predviđenih se poboljšao sa 70% na 73%. Dok se prosečna osetljivost znatnije promenila, sa 67% na 75%.

Table 2: Promena nekih mera uspešnosti klasifikatora

	Pre upsamling-a	Posle upsampling-a
Procenat tačno predviđenih	70%	73%
Prosečna osetljivost	67%	75%

Kako bi se ispravnije evaluirao klasifikator, izvršen je undersampling Južne US (Tabela 3). Međutim kada je izvšena klasifikacija, evaluacije klasifikacije pokazivale su nepromenjene vrednosti. Stoga je neophodno na neki drugi način uticati na promenu modela.

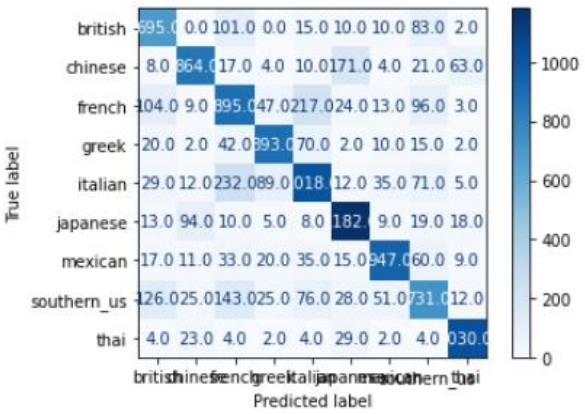
Tabela 3: Broj uzoraka po državi nakon undersampling-a i upsampling-a

Država	Broj uzoraka
British	916
Chinese	1162
French	1408
Greek	1056
Italian	1503
Japanese	1358
Mexican	1147
Southern_us	1217
Thai	1102

Sledeći korak je da se pokuša promena broja neruona po slojevima, kako bi se videlo da li će doći do značajnije izmene u rezultatima. Kada se smanji broj neurona po sloju sa 70 na 64, rezulati su ostali nepromenjeni. Međutim kada je broj neurona po sloju povećan na 100, prosečna osetljivost se povećala na 76%, dok je procenat tačno predviđenih porastao na 75%.

Stoga kao najbolji klasifikator odabran je onaj kada postoji 3 skrivena sloja sa po 100 neurona. Korištena je tangens hiperbolik funkcija aktivacije, koja je unela nelinearnost. Time se obezbeđuje da mreža ima više mogucnosti da se prilagodi podacima. Na matrici konfuzije (Slika 3) moguće je videti koliko je neuronska mreža sa odabranim parametrima dobra u kalsifikaciji uzoraka.

Na glavnoj dijagonali nalazi se broj koji govori koliko je uzoraka iz određene klase tačno klasifikovano. Može se uočiti da su najtamnija polja upravo na glavnoj dijagonali. Tamnija polja označavaju veće brojeve. Takođe se primećuje da je najtamnije polje u vrsti upravo ono koje pripada glavnoj dijagonali. Cilj pravljenja klasifikatora je da na glavnoj dijagonali budu sto veći brojevi, odnosno da se sto veći broj uzoraka klasifikuje tačano.



Slika 3: Matrica konfuzija za odabranu neuronsku mrežu

C. *Testiranje klasifikatora*

Klasifikator sa odabranim parametrima je obužen i testiran nad neviđenim test podacima. Dobijena je prosečna vrednost za osetljivost od 64%. Ova vrednost je manja od one koja je dobijena prilikom primene unakrsne validacije na trening podacima, koja je iznosila 76%. Međutim to nije čudno, zato što se u test skupu sačuvao odnos broja uzoraka u pojedinim klasma iz originalnog skupa.

V. KLASIFIKACIJA-KNN

A. *Uvod*

Za drugi model korišten je klasifikator k najbližih suseda. Parametri koji treba da se odrede kako bi klasifikator davao optimalna rešenja jesu: broj suseda koji se uzima u obzir i metrika koja se koristi za određivanje rastojanja.

B. *Procena parametara*

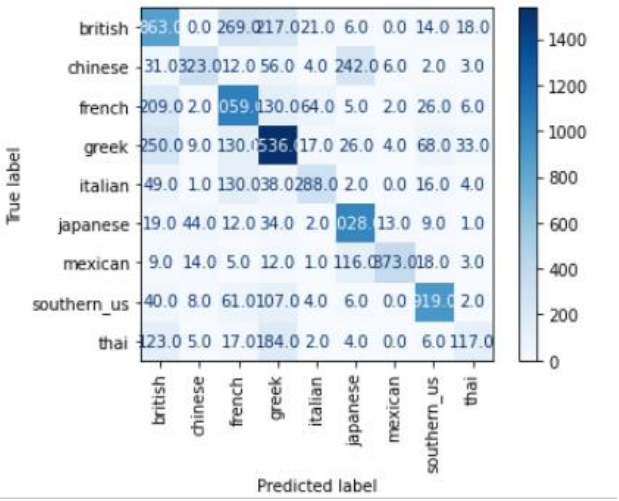
Evaluacija parametara je rađena uz pomoć metode unakrsne validacije. Metrike koje su iprobavane su metrike koje se koriste kada su date binarne vrednosti. Neke od njih su jaccard, matching, dice, kulsinski... Takođe testirane su različite vrednosti za broj suseda (Tabela 4). Može se uočiti da su vrednosti osetljivosti za matching metriku u svakoj iteraciji lošije. Stoga ona neće biti uzeta u obzir. Najbolja vrednost za osetljivost je postignuta u iteraciji sa 17 suseda i jaccard metrikom. Takođe isti postupak unakrsne validacije sproveden je nad metrikama dice i kulsinski. Vrednosti koje se dobiju su veoma slične kao prethodne. Dok je i u ovom sluaju najbolja vrednost za osetljivost ista, za parametre: 17 seseda i dice metrika.

Tabela 4: Osetljivost za promeljiv broj suseda i metrike

Broj suseda	jaccard	matching
3	63%	59%
7	67%	63%
9	67%	64%
11	67%	64%
15	68%	64%
17	68%	64%

Mera evaluacije koja je ovde izabrana za korišćenje je osetljivljivost kao mikromera. Mikromere se koriste kada je prisutan veći klasni dizbalans. Takođe osetljivost je pogodno za korišćenje zbog istog razloga. Upravo zato je u ovom skupu podataka prikladna ovakva mera evaluacije klasifikatora.

Matrica konfuzije (Slika 4), koja je dobijena akumulacijom matrica iz svake od iteracija unakrsne validacije, govori koliko je određeni klasifikator uspešan. Ovde se može uočiti da je posmatrajući svaku vrstu najtamnije polje ono koje se nalazi na glavnoj dijagonali,a elementi na glavnoj dijagonali govore broj uspesnih klasifikacija. Ova činejnića ukazuje da dobijeni klasifikator u nekoj meri uspećno radi posao.



Slika 4: Matrica konfuzija za odabrane parametre

Analizirane su i osetljivosti za svaku klasu ponaosob (Tabela 5). Može se primetiti da je najveća osetljivost za klasu Japanese i iznosi 88%. Ovo ukazuje na činjenicu da je posmatrajući 100 uzoraka koji su tačni, njih 88 klasifikovno kao tačni dok je 12 uzoraka pogrešno klasifikovano. Dok se može uočiti i jako mala vrednost osetljivosti za klasu Thai i iznosi 25%, odnosno samo 25 uzoraka od 100 tačnih, se kalsifikuje kao tačno, dok se 75 uzoraka netačno kalsifikuje. Međutim uglavno sve vrednosti za osetljivost su iznad 50% , a prosečna osetljivost je 63%, što ukazuje da ovaj klasifikator nema idelane karakteristike, ali da u određenoj meri uspešno radi klasifikaciju.

Tabela 5: Osetljivost po klasama

Država	Osetljivost (unakrsna vlidacija)	Osetljivost (finalna obuka)
British	61%	15%
Chinese	47%	89%
French	70%	71%
Greek	74%	54%
Italian	54%	69%
Japanese	88%	47%
Mexican	67%	75%
Southern_us	80%	73%
Thai	25%	65%
Prosečna osetljivost	63%	62%

C. Testiranje klasifikatora

Ispitivanjem najbolje mere evaluacije klasifikatora, dobijeni su parametri: 17 suseseda i jaccard metrika. Sledi obuka i testiranje finalnog modela. Testiranje je izvršeno na neviđenim podacima. Dobijena je osetljivost od 68% i ona se podudara sa osetljivošću dobijenom prilikom unakrsne validacije. Međutim može se videti da se osetljivosti za pojedinačne klase razlikuju. Kako sam klasifkator prilikom metode unakrsne validacije nije davao velike vrednosti za mere evaluacije, nije se moglo očekivati da se nad test skupom dobiju zavidni rezultati. Stoga nije začuđujuće što se rezultati razlikuju. Dok se prosečna osetljivost nije značajnije promenila. Sto se poklapa sa činjenicom da je klasni odnos u tening i test skupu ostao isti kao i u originalu.

VI. POREĐENJE KLASIFIKATORA

Sledi poređenje dobijenih klasifikatora.

Postoji razlika između makro i mikro mere. Makromera će izračunati meru nezavisno za svaku klasu, i zatim uzeti prosek. Odnosno tretiraće svaku kalsu podjednako. Dok će mikromera agregirati doprinose svih klasa za izračunavanje prosečne mere. Ukoliko se radi sa viseklasnim problemima, mikromere su korisne ako se sumnja na nebalansiranost klasa.

Kod oba klasifikatora može se primetiti (Tabela 6) da je za osetljivost i za F meru vrednost za mikromeru veća od makromere. To je u skladu sa činjenicom da je u skupu podataka prisutna neizbalansiranost uzoraka u klasama. Razlog zašto su ove makromere manje, može biti postojanje jedne ili više klasa koje imaju malu meru za osetljivost, pa su te male vrednosti povukle prosečnu. Kada se posmatraju vrednosti za osetljivost za pojedinačne klase može se uočiti da je kod KNN-a za British klasu, vrednosti samo 15%. Dok kada se posmatra kod neurosnke mreže vrednost za British iznosi 17%, dok su za sve ostale klase vrednosti mnogo veće.

Tabela 6: Poređenje mera uspečnosti klasifikatora

		Preciznost	Osetljivost	F mera
KNN	Micro	68%	68%	68%
	Macro	69%	62%	64%
Neuronska mreža	Micro	70%	70%	70%
	Macro	70%	64%	66%

Takođe kod preciznosti za KNN se može primetiti da je vrendosti za makromeru malo veća od mikromere, što ukazuje na postojanje klase koja je imala nešto veću vrednost za preciznost.

Može se primetiti da su vrednosti kod neuronske mreže nešto više od KNN. Ovo može da ukazuje da je neruonska mreža malo kvalitetniji način kalsifikacije. Kod neuronskih mreža postoji mogućnost slojevite građe, tj dubine. Složenost modela se ostavaruje dubinom. Takođe neuronske mreže koriste funkcije aktivacije koje mogu da unesu nelinearnost i time se obezbeđuje da mreža ima više mogućnosti što se tiče prilagođavanja podacima. Mana neuronske mreže je potreba da se definiše veliki broj hiperparametara i nemanje pravila kako najbolje odrediti neke od parametara kao što su broj slojeva ili neurona po sloju. Već je neophodno eksperimentalnim putem utvrđivati. Dok je kod KNN-a neophodno samo navesti broj suseda i metriku.

Kod KNN-a prilikom određivanja odgovarajuće klase, za novopristigli uzorak, posmatraju se samo određeni uzorci, odnosno k najblizih suseda, za koje će se računati metrika. Ovim se postiže značajna usteda vremena.