Klasifikacija recepata na osnovu prisustva/odsustva određenih sastojaka

Nikola Savić, in35/2018, nikolaakv@gmail.com

I. Uvod

Izveštaj se bavi analizom sastojaka za recepte, kao i klasifikacijom odakle poticu recepti na osnovu sastojaka. Nakon što bi analizirali sastojake i utvrđivali uticaje različitih sastojaka na svoje poreklo, moći ćemo da formiramo dva klasifikatora koji bi na osnovu unetih sastojaka svrstao recept u odgovarajuću kategoriju tj. državu.

Klasifikator bi se mogao koristiti za automatizovano labeliranje recepata, nakon unešenih sastojaka.

II. Baza podataka

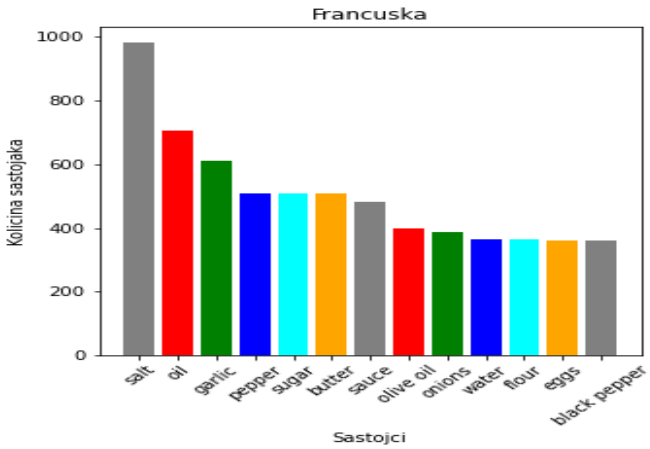
Trening baza sadrži podatke o 9509 uzoraka i 152 kategoričkih obeležja, od kojih jedno obeležje predstavlja klasu kojoj recept pripada i ona označava državu odakle taj recept potiče. Osim ovog obeležja imamo jos jedno kategoričko obeležje pod imenom „Unnamed: 0” koje označava redni broj recepta i kojeg smo izbacili iz našeg trening skupa.

Ostalih 150 obeležja predstavljaju sastojke. Za svako od obeležja koje predstavlja sastojak, moguće su dve vrednosti, 0 i 1, gde 0 predstavlja odsustvo, dok 1 predstavlja prisustvo određenog sastojka. Recept može da pripada jednoj od devet klasa recepata odakle potiču, a to su recept koji je iz Velike Britanije, Kine, Francuske, Grčke, Italije, Japana, Meksika, Juzne Amerike ili iz Tajlanda., recept za pecivo ili recept za picu. Test baza sadrži 1057 uzoraka.

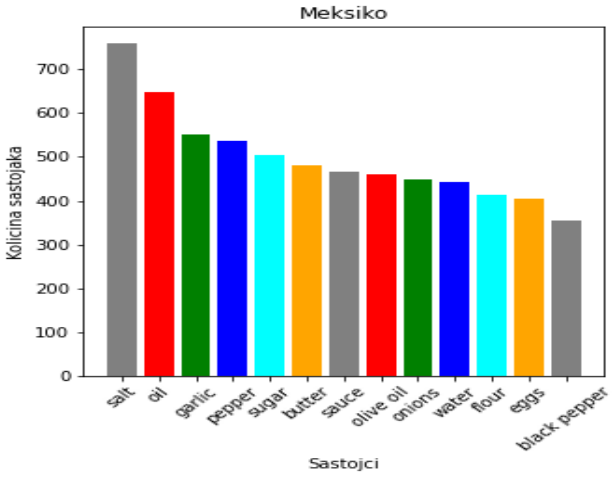
III. Analiza podataka

Analizom podataka trening i test baze utvrđeno je da ne postoje nedostajući podaci. Najveći broj recepata vidimo u klasi za državu Južne Amerike, njih 2303. Prati ga Italija sa 1670, Francuska sa 1565, za njom Kina sa 1291, pa Meksiko sa 1274 uzoraka. Dosta manji broj uzoraka vidimo kod država kao što su Japan sa 755, Tajland sa 612, zatim Grcka sa 587 i najmanje Velika Britanija sa 509 uzoraka.

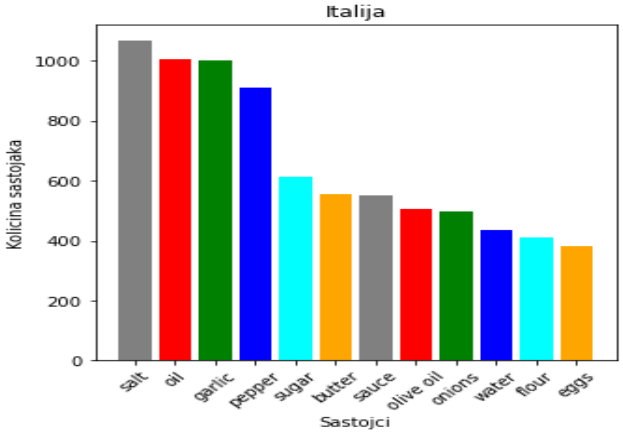
IV. Formatiranje



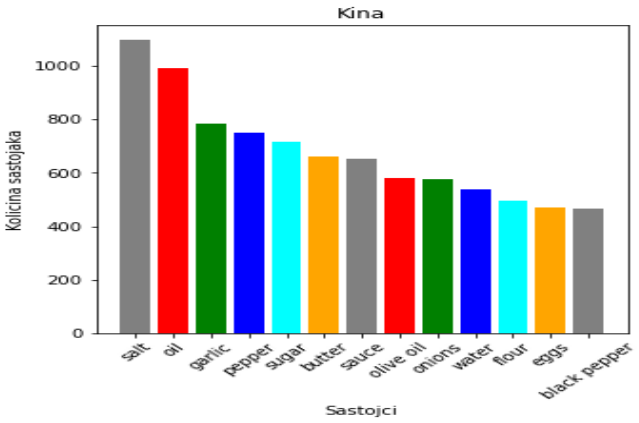
Sl. 1. Karakterističan obrazac sastojaka za Francusku.



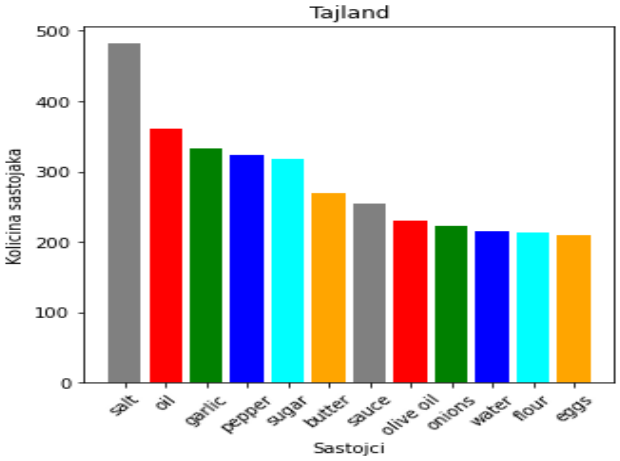
Sl. 2. Karakterističan obrazac sastojaka za Meksiko.



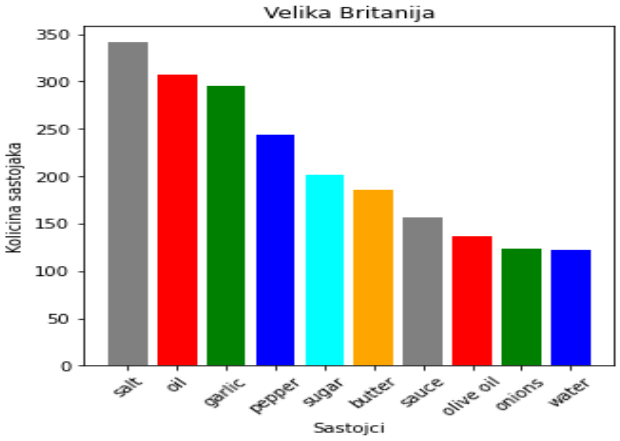
Sl. 3. Karakterističan obrazac sastojaka za Italiju



Sl. 4. Karakterističan obrazac sastojaka za Kinu.



Sl. 5. Karakterističan obrazac sastojaka za Italiju

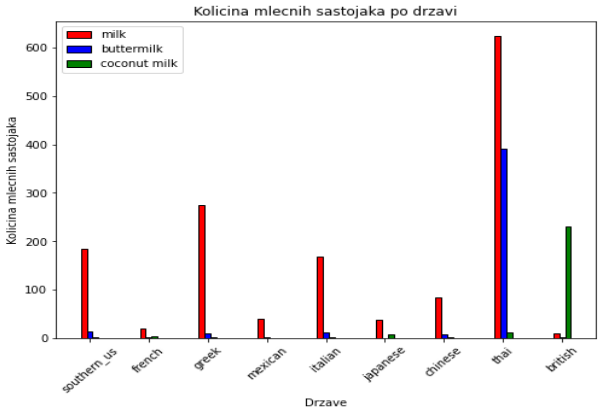


Sl. 6. Karakterističan obrazac sastojaka za Veliku Britaniju

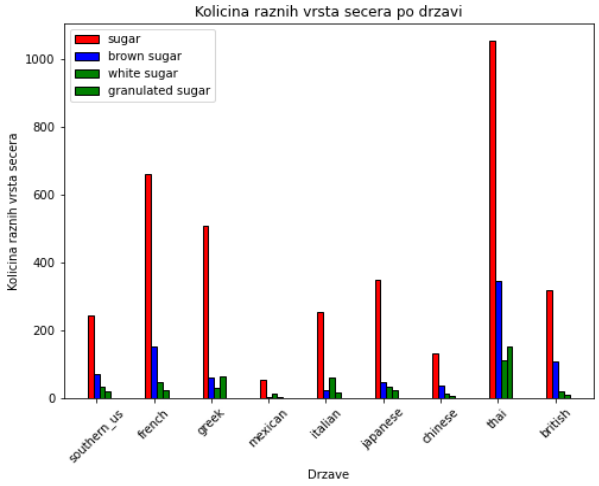
Iz karakterističnih obrazaca različitih klasa recepata, možemo primetiti da su određeni sastojci zastupljeni u većini recepata. Kao što vidimo sa slika iznad osnovni sastojci su brašno, ulje, luk, sećer, puter, voda, dok su neki određeni sastojci karakteristični samo za pojedine države.

U slučaju recepata u Velikoj Britaniji, Francuskoj i Meksiku, karakteristični sastojak koji se razlikuje u odnosu na ostale države je crni biber.

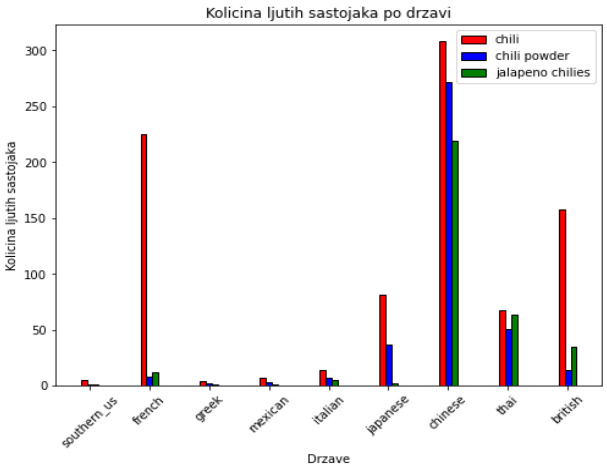
V. DODATNE ANALIZE



Sl. 7. Količina mlečnih proizvoda po državi.



Sl. 8. Količina raznih vrsta secera po državi.



Sl. 9. Količina ljutih sastojaka po državi.

Sa slike (Sl 7.) vidimo da je država u kojoj je najzastupljenije mleko u receptima u Tajlandu, takodje vidimo da u istoj državi dominira puterovo mleko. Kokosovo mleko se jedino od ovih 9 država koristi u sastojcima iz Velike Britanije, dok vidimo da u ovoj državi da skoro i nema korisćenja obicnog mleka u sastojcima u našem setu podataka.

Što se tiče raznih vrsta šećera u receptima, sa slike (Sl. 8.) vidimo da ponovo Tajland dominira što se tiće običnog, a i braon sećera. Granulisani šećer najviše je zastupljen u Tajlandu, dok ga prati Grcka.

Kao što svi znamo, zemlja sa najviše ljutih i dinamičnih recepata je Kina. Sa grafika sa slike (Sl. 9.) vidimo da se

intenzivno koriste ljuti sastojci, ljute paprike i jalapeno ljuti sastojak. Takodje vidimo da se u državama kao što su Grcka, Meksiko, Italija i Juzna Amerika skoro pa uopste ne nalaze recepti koji sadrze ljute sastojke.

VI. KLASIFIKATORI

A. *Odabir optimalnih parametra*

Testirajuci KNN klasifikator za razne vrednosti k, od 1 do 20 i podešavajuci parametar koji govori na koliko cemo delova da podelimo naš trening skup, na 5 ili 10, dolazimo do zakljucka da najbolje rezultate dobijamo korišćenjem podelom k najbližih suseda sa podelom trening skupa na 10 delova.

S obzirom da nam klasifikacija daje najbolje vrednosti sa podelom našeg trening skupa na deset podskupova, preostaje nam samo da vidimo koju vrednost ćemo izabrati za k najbližih suseda da bi klasifikacija bila optimalna u odosu na Dzakardovu i Dajsovu metriku.

Testirali smo Dzakard i Dajsovu metriku jer su nam vrednosti obeležja brojevi 0 i 1, tj vrednosti nisu realni brojevi pa metriku ne možemo raditi standardnim metrikama poput Euklidske. Takodje razlog iz kojeg koristimo Đakardovu metriku jeste da je on pogodan za vrednosti koje su boolean.

Primenom funkcije za pronalaženje optimalnih parametra dobijeno je da najbolje rezultate, tj za najveću tačnost dobijamo za k=16, gde je tačnost 69,4%. Isti rezultat je dobijen i sa Dajsovom i sa Dzakardovom metrikom.

	Britanija	Kina	Francuska	Grcka	Italija	Japan	Meksiko	JuznaA.	Tajland
Britanija	1633	179	17	71	125	7	14	4	23
Kina	245	862	22	9	250	0	7	0	13
Francuska	43	42	310	15	115	1	0	0	2
Grcka	140	31	6	997	60	4	7	1	2
Italija	161	190	52	27	1059	3	3	1	7
Japan	65	26	6	3	13	350	203	7	6
Meksiko	48	15	3	9	7	53	1010	15	2
Juzna A.	19	6	1	16	8	13	114	373	1
Tajland	198	121	3	5	19	7	3	0	102

Tabela 1: Matrica konfuzije nakon unakrsne validacije kod KNN klasifikatora za 9 država

Sa druge strane, klasifikatorom logističke regresije, deleći naš skup na 5 jednakih delova treba da proverimo za koju iteraciju i za koji solver nam daje najbolje rezultate. Korišćenjem funkcije za pronalaženje optimalnih parametara za logisticku regresiju dobijamo da najbolje rezultate dobijamo korišćenjem broja iteracija koji je jednak 500 u slučaju "Newton-cg" solvera. U tom slucaju tačnost je 71,4%, mikro osetljivost 71,4% i makro osetljivost je 67,9%

	Britanija	Kina	Francuska	Grcka	Italija	Japan	Meksiko	JuznaA.	Tajland
Britanija	1635	183	18	55	106	9	13	10	44
Kina	245	905	12	8	207	5	6	1	19
Francuska	30	44	341	18	85	2	1	1	6
Grcka	102	22	5	960	39	4	5	3	7
Italija	131	227	43	30	1056	6	3	1	6
Japan	53	28	3	14	12	368	176	13	12
Meksiko	51	18	2	6	10	67	975	30	3
Juzna A.	19	2	2	10	2	16	59	439	2
Tajland	208	109	4	4	18	8	1	1	105

Tabela 2: Matrica konfuzije nakon unakrsne validacije kod KNN klasifikatora za 9 država

Najveće greške možemo videti iz tabela (Tabela 1 i Tabela 2) gde klasifikator pravi između klasa recepata za državu Kinu. U slučaju KNN klasifikatora, 245 uzoraka koji priadaju klasi recepata za državu Kinu je proglasio da je iz države Britanije i 250 da je iz Italije. Takodje vidimo da je čak 198 recepata koji pripadaju klasi recepata iz Tajlanda proglasio da je iz Velike Britanije. 203 recepata iz Japana je proglasio da su iz Meksika.

Ukoliko posmatramo matricu konfuzije kod klasifikatora na bazi logističke regresije, u poređenju sa KNN klasifikatorom, vidimo da je znatno smanjio grešku klasifikovanja recepta iz Kine kao recepta iz Italije, dok je na primer povećao grešku klasifikovanja recepta iz Italije kao recepta iz Kine.

B. *Rezultati testiranja na test skupu*

Oba klasifikatora, KNN klasifikator i klasifikator logističke regresije je sa predhodno izabranim parametrima obučen na celokupnom trening skupu i nakon toga testiran na test skupu. Veću tačnost postigao je klasifikator logističke regresije od KNN klasifikatora, gde je tačnost logističke regresije 72,3%, dok je KNN klasifikator postigao nešto nižu tačnost od 70,8%.

Preciznost(%)	Britanija	Kina	Francuska	Grcka	Italija	Japan	Meksiko	JuznaA	Tajland
KNN	65	66	73	86	64	82.6	71	91.1	68.4
Logistička	67	62.2	68.5	86.7	74.2	73.8	78.6	85	50

Tabela 3: Preciznost kod KNN i klasifikatora logističke regresije svih za 9 država

Kao što vidimo iz tabele (Tabela 3) KNN klasifikator je bolje prediktovao vrednosti za Tajland 68.4%, dok je logistička dosta manje, samo 50%. U slučaju Grčke imamo jednako dobre rezultate, oko 86%. Slučaj gde logistička regresija daje bolje rezultate je država Italija sa 74.2%.

	Britanija	Kina	Francuska	Grcka	Italija	Japan	Meksiko	JuznaA.	Tajland
Britanija	182	16	0	6	16	2	5	1	2
Kina	27	100	0	2	25	0	0	1	2
Francuska	10	5	27	0	16	0	1	0	0
Grcka	12	0	0	110	4	0	1	0	0
Italija	18	15	10	3	119	0	1	0	1
Japan	3	3	0	3	2	38	26	0	1
Meksiko	0	0	0	2	2	5	118	2	0
Juzna A.	1	1	0	2	1	1	14	41	0
Tajland	26	11	0	0	1	0	0	0	13

Tabela 4: Matrica konfuzije nakon unakrsne validacije kod na test skupu kod KNN klasifikatora za 9 država

	Britanija	Kina	Francuska	Grcka	Italija	Japan	Meksiko	JuznaA.	Tajland
Britanija	180	17	3	9	12	2	3	0	4
Kina	30	99	4	1	17	2	0	0	4
Francuska	4	6	37	0	10	0	0	0	2
Grcka	14	1	1	111	0	0	0	0	0
Italija	12	19	9	3	121	1	0	1	1
Japan	4	4	0	2	0	48	17	0	1
Meksiko	0	0	0	1	1	10	110	7	0
Juzna A.	2	0	0	0	0	2	10	47	0
Tajland	23	13	0	1	2	0	0	0	12

Tabela 5: Matrica konfuzije nakon unakrsne validacije kod na test skupu kod klasifikatora logističke regresije za 9 država

Iz ove matrice konfuzije iz tabele (Tabela 4) vidimo da najveće odstupanje kod KNN-a ima situacija gde KNN klasifikator iz klase za Japan svrstava 26 recepta za Meksiko. Najveće odstupanje za logističku regresiju vidimo iz tabele (Tabela 5) gde za 30 recepata iz Kine svrstava u Veliku Britaniju.

C. Upoređivanje rezultata unakrsne validacije i rezultata dobijenih na test skupu

U slučaju klasifikacije KNN aloritmom unakrsnom validaciom vidimo da je procenat pogodjenih uzoraka 69,36%, dok je procenat pogodjenih uzoraka nad test skupom jednak 70,08%. Iz ovoga vidimo da je pokazuju maltene iste rezultate. Što se tiče klasifikatora logističke regresije vidimo takodje priblizne rezultate 71,3% kod unakrsne i 72,4% kod testiranja nad test skupom podataka.

Što se tiče uporedjivanja preciznosti izmedju svake klase, tj drzave možemo videti da većina drzava ima sličnu preciznost u rezultatima kod unakrsne i kod rezultata nad test skupom. Drzava u kojoj se rezultati razlikuju jeste Kina gde je preciznost redom kod prve 58.5%, dok je kod druge klasifikacije 66,2%.

D. Upoređivanje klasifikatora

Procenat pogodjenih uzoraka	0.7077
Mikro preciznost	0.7077
Makro preciznost	0.7417
Osetljivost mikro	0.7077
Osetljivost makro	0.6451
F mera mikro	0.7077
F mera makro	0.6702

Tabela 6: Performanse KNN klasifikatora

Procenat pogodjenih uzoraka	0.7237
Mikro preciznost	0.7237
Makro preciznost	0.7183
Osetljivost mikro	0.7237
Osetljivost makro	0.6810
F mera mikro	0.7237
F mera makro	0.6929

Tabela 7: Performanse klasifikatora logističke regresije

KNN	Britanija	Kina	Francuska	Grčka	Italija	Japan	Meksiko	JuznaA	Tajland
Preciznost	0.64	0.59	0.74	0.85	0.64	0.80	0.74	0.93	0.65
Tačnost	0.86	0.89	0.97	0.96	0.89	0.96	0.95	0.98	0.96
Osetljivost	0.78	0.61	0.59	0.79	0.70	0.52	0.87	0.68	0.22
Specificnost	0.89	0.92	0.99	0.98	0.93	0.99	0.96	0.997	0.99
F-mera	0.70	0.60	0.65	0.82	0.67	0.63	0.80	0.78	0.33

Tabela 8: Performanse klasifikatora na bazi KNN-a po klasama

Logistička Regresija	Britanija	Kina	Francuska	Grčka	Italija	Japan	Meksiko	JuznaA	Tajland
Preciznost	0.67	0.62	0.69	0.867	0.74	0.74	0.79	0.86	0.5
Tačnost	0.88	0.89	0.96	0.97	0.97	0.96	0.96	0.98	0.95
Osetljivost	0.78	0.63	0.63	0.87	0.72	0.63	0.85	0.78	0.26
Specificnost	0.89	0.93	0.98	0.98	0.95	0.98	0.97	0.992	0.99
F-mera	0.72	0.63	0.65	0.87	0.73	0.68	0.82	0.81	0.32

Tabela 9: Performanse klasifikatora na bazi logističke regresije po klasama

Kada pogledamo tabele (Tabela 8 i 9) možemo zaključiti da klasifikator KNN ima najveći udeo ispravno klasifikovanih recepata iz Južne Amerike u odnosu na sve uzorke koji su predviđeni da pripadaju klasi recepata iz Južne Amerike sa čak 93%. Sa druge strane vidimo da je u tom slučaju za logističku regresiju nešto manji procenat udela ispravno klasifikovanih uzoraka i to 86%.

Gledajući tabelu (Tabela 11) za udeo ispravno klasifikovanih recepata iz Tajlanda za logističku regresiju žaključujemo da je neipravno klasifikovano čak 50 % recepata iz Tajlanda u ondosu na sve uzorke koji su predviđeni da su u klasi recepata iz Tajlanda. U slučaju KNN-a za državu Tajland imamo malo veći procenat uspešno klasifikovanih recepata koji dolaze iz Tajlanda i to 65 %.

Generalno, kada bi poredili ova dva algoritma po udelu o ispravno klasifikovanih recepata vidimo da prvi klasifikator u slučaju nekih država bolji od drugog, takodje i da je drugi klasifikator u nekim slučajevima bolji od prvog. S pogledom na tu situaciju vidimo da je ukupni procenat udela pogodjenih kod KNN i logističke regresije približno jednak, redom 70,08% i 72,3%.

Iz tabele (Tabela 6 i Tabela 7) Harmonijska sredina mikro preciznosti i mikro osetljivosti je za KNN klasifikator je 70,08%, dok je u slučaju klasifikator logističke regresije 72,3%%. Sa druge strane, vidimo da je harmonijska sredina makro preciznosti za KNN klasifikator nesto veća od klasifikatora logističke regresije za skoro 3%. Što se tiče makro osetljivosti vidimo da nesto bolje rezultate daje klasifikator logističke regresije sa 68% u odnosu na KNN klasifikator sa 64,5%.

Iz ovoga smo videli da klasifikator za malo bolje funkcioniše sa velikim skupom podataka u odnosu na KNN klasifikator.