Određivanje recepata na osnovu sastojaka

Milica Milovanović, BI 8/2017, e-mail: milicaaaaa98@gmail.com

I. Uvod

Postoji dosta sastojaka koji su zajednički i za kolače i za peciva i za picu. Razlika u tome da li će se neka hrana svrstati u slanu ili slatku može da bude i samo jedan sastojak ili na primer značajno veća količina jednog sastojka u odnosu na druge.

Cilj ovog rada jeste se da na osnovu prisustva odnosno odsustva pojedinih sastojaka proceni da li se radi o receptu za kolače, pecivo ili picu. Značaj ovog rada je u tome da ljudi mogu sami da odrede na osnovu sastojaka koje imaju u svom domaćinstvu da li mogu da naprave neki kolač, pecivo ili možda picu. Takođe, jedna od primena ovog rada bi mogla da bude u razvrstavanju recepata iz neke kulinarske knjige po kategorijama pecivo, kolač i pica na osnovu spiska sastojaka.

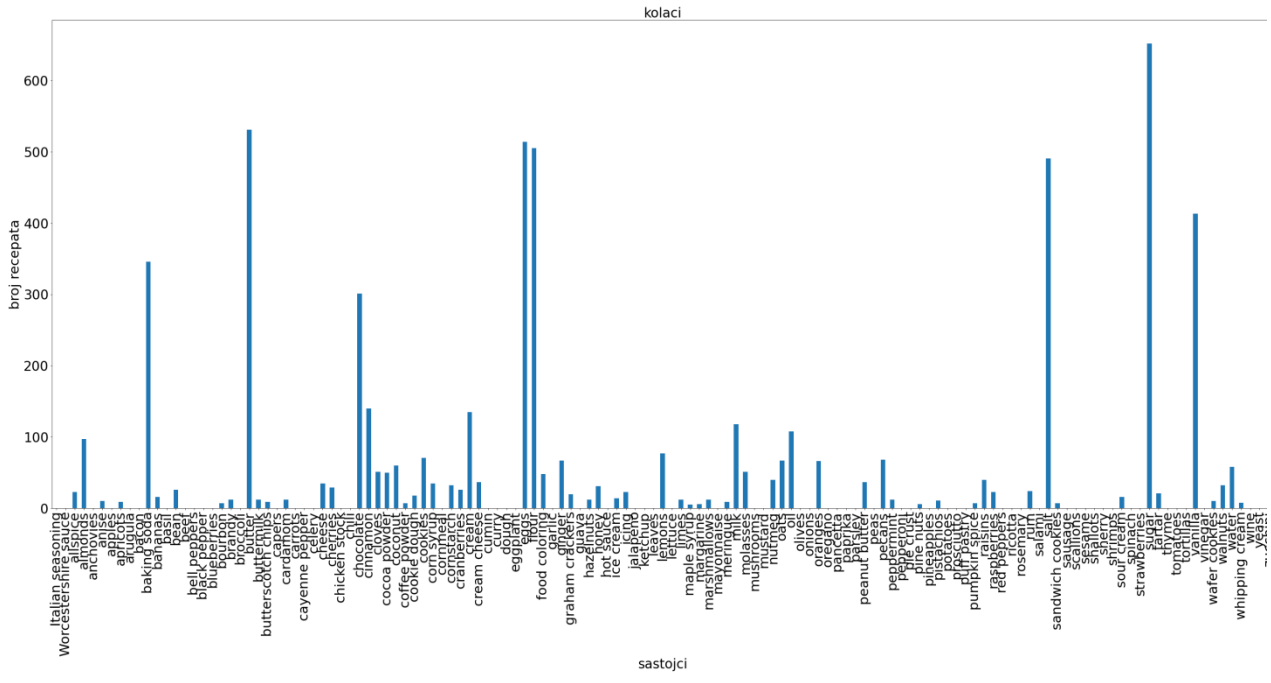
II. Baza podataka

Bazu podataka čine 1931 recept koji predstavljaju uzorke u ovoj bazi podataka i 133 sastojka koji su obeležja ove baze. Recepti su podeljeni tako da je u skupu za obuku 1738 recepata, a u test skupu 193. Obeležja u bazi su binarna, gde 0 predstavlja odsustvo nekog sastojka, a 1 njegovo prisustvo. Neki od sastojaka su ulje, brašno, šećer, so. Poslednju kolonu baze podataka čine klasne labele, a to su kolačići, peciva i pica. Cilj ovog rada je da se na osnovu prisustva odnosno odsustva sastojaka napravi klasifikator koji će moći da odredi da li je u pitanju recept za kolačić, pecivo ili picu.

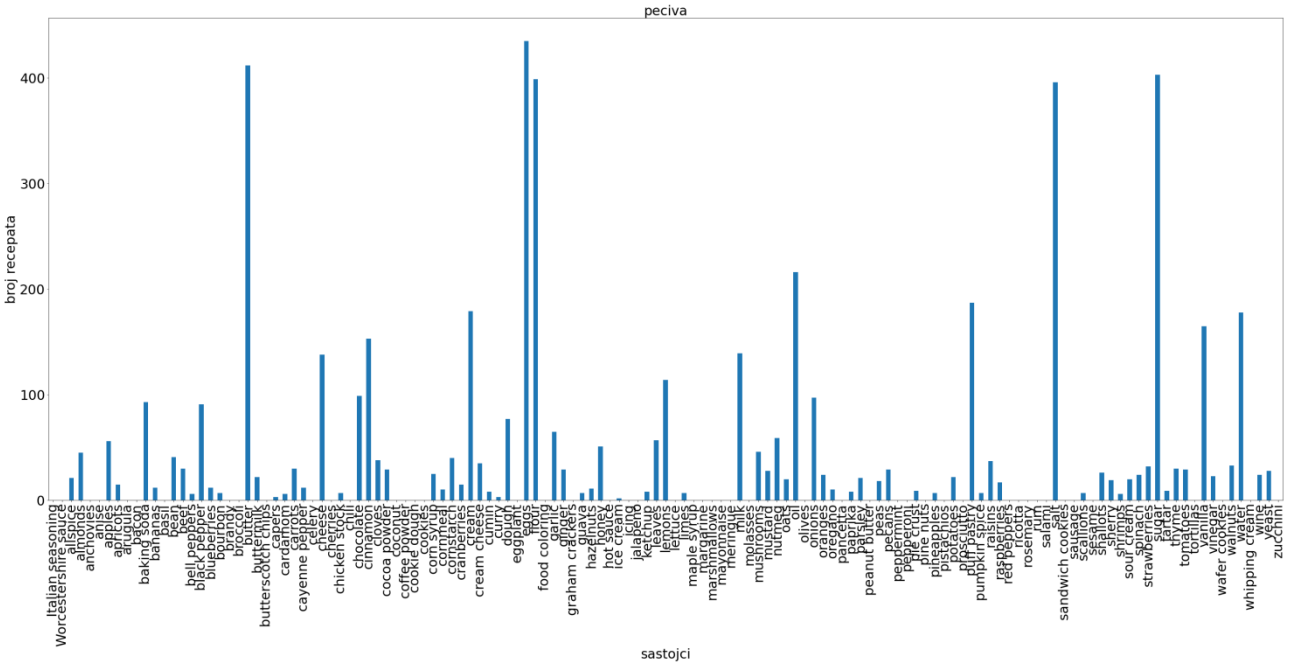
III. Određivanje klasifikatora

A. *Histogram pojavljivanja sastojaka za svaku klasu*

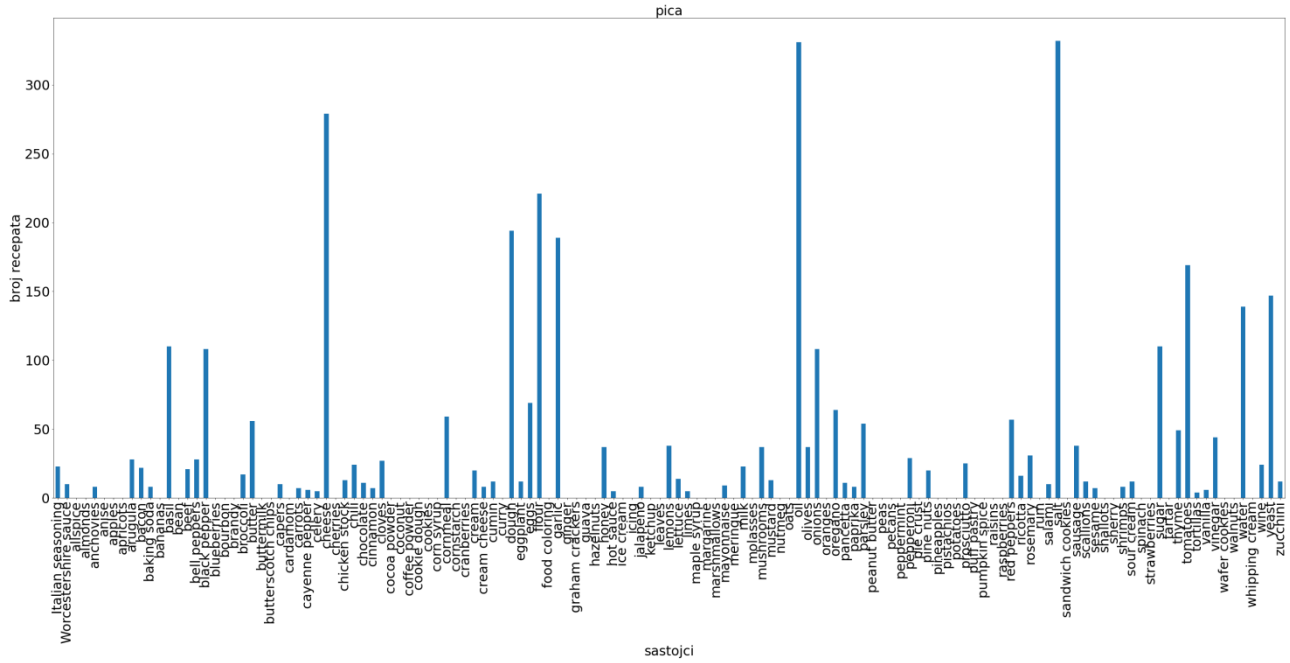
Za svaku od 3 klase nacrtani su histogrami koji prikazuju pojavljivanje sastojaka u receptima za kolače, peciva i picu. Na slici 1 prikazan je histogram za klasu kolača. Sa histograma se mogu videti sastojci koji se najviše koriste u receptima za kolače i to su: margarin, jaja, brašno, šećer, a izdvajaju se i soda bikarbona, vanila i čokolada, ali u nešto manjoj meri u odnosu na ove glavne sastojke. Kada se uporedi histogram pojavljivanja sastojaka za kolače sa histogramom za peciva, koji je prikazan na slici 2, može se zaključiti da se glavni sastojci za kolače pojavljuju kao glavni sastojci i u pecivima. Na slici 3, može se zapaziti da je brašno zastupljeno u velikom broju recepata za picu, te se može doneti zaključak da je brašno zajednički glavni sastojak koji se koristi i za pravljenje kolača i za pravljenje peciva i za pravljenje pica i to u velikoj meri. Razlog je verovatno to što se brašno koristi pri pravljenju testa, koje predstavlja osnovu nekog recepta u većini slučajeva. Interesantno je da se sir izdvaja kao jedan od glavnih sastojaka pice, a da nije toliko zastupljen u druge dve klase. Takođe, svaka od klasa ima i neke svoje sastojke koji su karakteristični samo za tu klasu, recimo u receptima za picu je to čili, koji se ne pojavljuje u druge dve klase, s druge strane rum, puter od kikirikija, jestive boje su specifični sastojci samo za kolačiće, a krompir je sastojak specifičan samo za peciva.



Slika 1. Histogram pojavljivanja sastojaka za kolače



Slika 2. Histogram pojavljivanja sastojaka za peciva



Slika 3. Histogram pojavljivanja sastojaka za picu

B. kNN klasifikacija

Jedna od dve korišćene klasifikacije u ovom radu je metoda k najbližih suseda. Ova metoda spada u grupu metoda kasnog učenja, koje podrazumevaju odlaganje obrade uzoraka za obuku do trenutka kada treba klasifikovati neobeleženi uzorak. Elementi potrebni za primenu kNN metode su: obeleženi skup uzoraka za obuku, odnosno da svaki uzorak ima dodeljenu klasnu labelu, celobrojni parametar k koji predstavlja broj najbližih suseda koji se uzima u obzir pri odlučivanju i metrika za utvrđivanje rastojanja između dva uzorka.

S obzirom da su u ovom radu obeležja binarna isprobane su različite metrike karakteristične za binarna obeležja kao što su Žakarova, Dajsova, Matching, Kulsinski i one su primenjivane za različiti broj suseda, tačnije za k=1, k=5 i k=10. Nakon što je primenjena unakrsna validacija sa 10 podskupova, pri čemu se vodilo računa o tome da u svakom podskupu bude dovoljan broj uzoraka svake klase, pokazalo se da je najbolja tačnost klasifikatora kada se koristi metrika Kulsinski sa jednim susedom. Za veći broj suseda tačnost opada.

Za uporedni prikaz usklađenosti stvarnih i predviđenih

labela koristi se matrica konfuzije. Dobijena matrica konfuzije za kNN klasifikator izgleda ovako:

$$\begin{bmatrix} 663 & 58 & 2 \\ 55 & 543 & 21 \\ 7 & 18 & 371 \end{bmatrix}$$

Vrste matrice konfuzije odgovaraju stvarnim vrednostima labela, a kolone predviđenim. Na glavnoj dijagonali matrice konfuzije nalaze se ispravno klasifikovani uzorci odnosno recepti. Na osnovu ove matrice konfuzije ispravno klasifikovani recepti su 663 recepata za kolačiće, 543 recepata za peciva i 371 recept za picu. Iz matrice konfuzije jasno se vidi koje klase klasifikator meša, a u ovom slučaju najviše meša recepte za kolačiće i peciva. Od 723 recepata za kolače, klasifikator je 58 recepata dodelio pecivima, a 2 recepta je dodelio picama. Zatim je 55 recepata za pecivo pomešao sa kolačima, a 21 recept za pecivo sa picom. Kod pica je najmanje grešio, ali treba uzeti u obzir da je i broj recepata za picu u skupu za obuku manji, 7 recepata je svrstao u kolače, a 18 u pecivo, iako originalno pripadaju receptu za picu. Iz matrice konfuzije se vidi da klasifikator najviše problema ima sa klasom peciva, a razlog za to bi mogli biti neki sastojci koji su karakteristični ili samo za kolače ili samo za picu, ali se ipak javljaju i u pecivima.

Na osnovu matrice konfuzije računaju se mere uspešnosti klasifikatora, kao što su: tačnost, preciznost, osetljivost (odziv), specifičnost i F-mera. U tabeli 1 su prikazane izračunate vrednosti za prosečnu tačnost klasa, kao prosek tačnosti po klasama i tačnost za svaku od klasa pojedinačno za unakrsnu validaciju i test skup.

.

Mere uspešnosti	Unakrsna validacija	Test skup
Prosečna tačnost	0,90173	0,93264
Tačnost za kolače	0,92980	0,95337
Tačnost za peciva	0,91254	0,93782
Tačnost za picu	0,97238	0,97409

Tabela 1: Mere uspešnosti kNN klasifikatora

Na osnovu izračunatih mera uspešnosti kNN klasifikatora za unakrsnu validaciju može se primetiti da se očekuju velike vrednosti, preko 90% za prosečnu tačnost, ali i tačnosti za kolače, peciva i picu, odnosno očekuje se veoma uspešan model. To se pokazalo i na nepoznatim uzorcima, jer su dobijene vrednosti čak veće od onih dobijenih unakrsnom validacijom.

C. SVM klasifikacija

SVM klasifikator se koristi za rešavanje problema binarne klasifikacije tako što identifikuje hiperpovrš koja treba da razdvoji uzorke u prostoru obeležja tako da između oblasti popunjenih uzorcima različitih klasa postoji što širi prostor. Ovaj algoritam se zasniva na klasifikatoru maksimalne margine, čiji je cilj da odredi hiperravan koja će na optimalan način podeliti prostor na dva dela tako da se u jednom delu nađu uzorci iz jedne klase, a u drugom delu uzorci iz druge. Da bi ovo bilo moguće, klase moraju biti linearno separabilne, a da bi hiperravan razdvajanja bila optimalna, ona mora obezbediti najveću moguću marginu, odnosno, najveće rastojanje od hiperravni do najbližeg uzorka iz obe klase. Ovako definisana hiperravan omogućava bolju generalizaciju, što znači da se očekuje tačnija klasifikacija novih uzoraka u odnosu na bilo koju drugu hiperravan koja bi takođe razdvojila uzorke iz skupa za obuku.

Unakrsnom validacijom se određuju parametri SVM klasifikatora, a to su regularizacioni parametar C koji određuje ukupnu toleranciju na grešku klasifikacije, kernel i pristup. Moguća su dva pristupa: svaki protiv svakog (*one vs. one*) i jedan protiv svih (*one vs. all* ili *one vs. rest*). Unakrsna validacija je u ovom radu pokazala da bi trebalo koristiti *one vs. one* pristup, gde se uzorak dodeljuje onoj klasi koja je najveći broj puta bila uspešnija u nadmetanju klasifikatora po parovima. Takođe se unakrsnom validacijom došlo do zaključka da bi trebalo parametar C postaviti na 10, a za kernel odabrati radijalni kernel.

Matrica konfuzije dobijena nakon 10 iteracija unakrsne validacije za SVM klasifikaciju izgleda ovako:

$$\begin{bmatrix} 669 & 54 & 0 \\ 43 & 561 & 15 \\ 7 & 15 & 374 \end{bmatrix}$$

Za razliku od matrice konfuzije kod kNN klasifikatora, broj tačno klasifikovanih recepata je veći, 669 recepata za kolače, 561 recept za pecivo i 374 recepta za picu su dobro klasifikovani. Zatim, iz matrice konfuzije se vidi da je 54 recepta dodelio receptima za pecivo umesto za kolače, ali nijedan recept za kolač nije pogrešno dodelio receptu za picu, što znači da recept za kolač ne meša sa picom, ali recept za picu ipak meša sa kolačem i to 7 puta, dok je 15 recepata za picu dodelio pecivima. Što se tiče recepata za pecivo, njih i dalje najviše meša sa kolačima, ali manje nego kNN metoda, 43 recepta za pecivo je dodelio kolačima, a 15 recepata picama.

U tabeli 2 su date izračunate vrednosti tačnosti za sve klase zajedno i pojedinačno na skupu za obuku i test skupu.

Mere uspešnosti	Unakrsna validacija	Test skup
Prosečna tačnost	0,91907	0,96373
Tačnost za kolače	0,93802	0,96373
Tačnost za peciva	0,92693	0,97293
Tačnost za picu	0,97871	0,98445

Tabela 2: Mere uspešnosti SVM klasifikatora

Udeo ispravno klasifikovanih uzoraka odnosno recepata u celoj populaciji na test skupu daje veće mere uspešnosti nego unakrsnom validacijom, što znači da je očekivan veliki procenat tačno klasifikovanih uzoraka, što se i potvrdilo na test skupu. Ispostavlja se da SVM klasifikacija daje bolje performanse, jer su vrednosti veće nego kod kNN klasifikacije, bar kada se radi o tačnosti klasifikatora.

D. Odabir klasifikatora

Nakon što su dobijene matrice konfuzije i za kNN i za SVM klasifikator na test skupu, određene su i druge mere uspešnosti, pored već pominjane tačnosti, a to su odziv, specifičnost i preciznost, za svaku klasu posebno, kao i mikro i makro preciznost i mikro i makro F-mera. Mikroprosečne mere svode se na sumiranje TP, FP i FN za sve klase, nakon čega se računaju mere uspešnosti na standardan način, dok makroprosečne mere predstavljaju proseke mera dobijenih po klasi. Ove vrednosti su date u tabeli 3 za oba klasifikatora.

Mere uspešnosti	kNN	SVM
Prosečna tačnost	0,93264	0,96373
Tačnost za kolače	0,95337	0,96373
Tačnost za peciva	0,93782	0,97927
Tačnost za picu	0,97409	0,98445
Odziv za kolače	0,9625	0,975
Odziv za peciva	0,91304	0,97101
Odziv za picu	0,90909	0,93182
Specifičnost za kolače	0,94690	0,95575
Specifičnost za peciva	0,95161	0,98387
Specifičnost za picu	0,99329	1,0
Preciznost za kolače	0,92771	0,93976
Preciznost za peciva	0,91304	0,97101
Preciznost za picu	0,97561	1,0
Mikro preciznost	0,93264	0,96373
Makro preciznost	0,93879	0,97026
F-mera mikro	0,93264	0,96373
F-mera makro	0,93302	0,96426

Tabela 3: Ostale mere uspešnosti oba klasifikatora

U tabeli 3 se ispostavilo da su i ostale mere uspešnosti klasifikatora veće za SVM klasifikator nego za kNN. Kao što je već rečeno, veće vrednosti mera ukazuju na bolje performanse modela, tako da je na kraju izabran SVM klasifikator.

IV. ZAKLJUČAK

Poređenjem kNN i SVM klasifikatora koji su analizirani u ovom radu došlo se do zaključka da treba izabrati SVM klasifikator, zbog najboljih performansi modela.

V. LITERATURA

[1] „Praktikum iz mašinskog učenja“, Tijana Nosek, Branko Brkljač, Danica Despotović, Milan Sečujski, Tatjana Lončar-Turukalo