Klasifikator recepata

Ksenija Stanojević, BI 32/2017, ksenijastanojevic@uns.ac.rs

I. Uvod

Cilj ovog zadatka je utvrđivanje recepata iz zadate baze i određivanje pripadnosti recepata jednoj od tri date klase - kolačići, peciva i pice, na osnovu sastojaka ponuđenih u bazi. Manipulacijom podataka iz baze mogu se izdvojiti recepti koji ne sadrže alergene ili određene toksične supstance, kako bi se lakše regulisala ishrana ljudi. Ovo je značajno za ljude koji imaju alergiju ili intoleranciju na određene sastojke, kao i za ljude koji žele na taj način da regulišu svoju ishranu.

II. Baza podataka

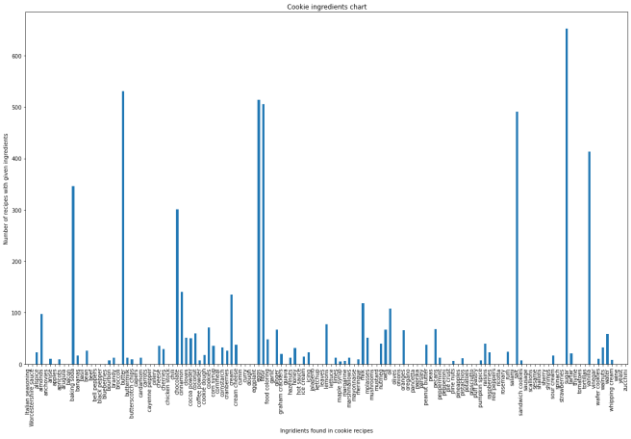
Ova baza sadrži 723 recepta iz klase kolačića, 619 recepata iz klase peciva, kao i 396 iz klase pica. Podaci su podeljeni u tri klase – kolačići, peciva i pice.

Obučavanje klasifikatora je izvršeno na bazi koja sadrži 1738 uzoraka (recepata). Ponuđeno je 133 sastojka, a za svaki od recepata je označeno prisustvo ili odsustvo određenog sastojka. Prisustvo i odsustvo određenog sastojka, označeno je brojevima 0 ili 1. Na osnovu grupe sastojaka koji se nalaze u receptu, određeno je kojoj klasi pripada recept.

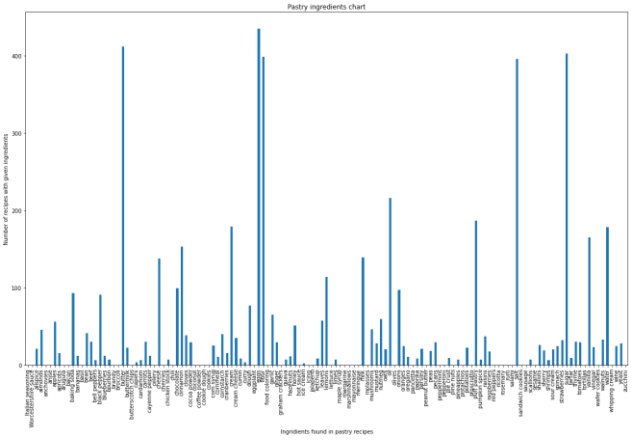
Na Slikama 1, 2 i 3, prikazani su histogrami pojavljivanja određenog sastojka, u određenoj klasi recepata. Na osnovu analize baze podataka, može se primetiti da se puter, prašak za pecivo, jaja, brašno, so i šećer, nalaze u većini recepata za kolače (Slika 1.) i peciva (Slika 2.). Primećujemo da ove dve klase u bazi imaju određeni broj istih sastojaka.

U klasi kolačića, nalaze se sastojci - vanila, ulje, mleko, bademi, limun i cimet, a u klasi peciva – sir, čokolada, cimet, voda, med, luk.

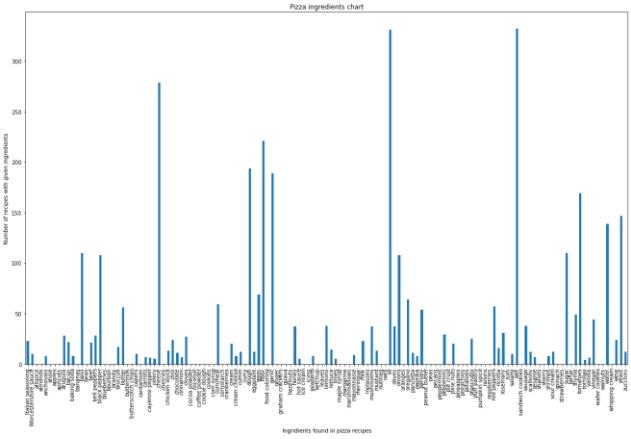
Analizom većeg broja sastojaka, dolazimo do preciznijih razlika među klasama. Sa druge strane, dominantni sastojci u receptima za picu su: sir, brašno, ulje i so (Slika 3.). Takođe, u klasi pica se nalaze puter, jaja i šećer, ali sa manjom učestanošću pojavljivanja, nego što je to slučaj u klasi kolačića i klasi peciva.



Slika 1. Histogram sastojaka koji se nalaze u kolačićima



Slika 2. Histogram sastojaka koji se nalaze u pecivima



Slika 3. Histogram sastojaka koji se nalaze u picama

III. ANALIZA PODATAKA

U ovom radu, za klasifikaciju sastojaka, korišćen je klasifikator pod nazivom mašina na bazi vektora nosača, kao i neuronske mreže. Učitani su podaci iz trening i test baza, na kojima je klasifikator obučen.

1) Neuronske mreže - NN

Urađena je unakrsna validacija sa 10 podskupova za prvi klasifikator – neuronske mreže, na trening podacima. U ovom klasifikatoru imamo 133 neurona u ulaznom sloju (vrednosti ulaznih obeležja) i 3 neurona u izlaznom sloju (vrerovatnoća pripadnosti određenoj klasi).

Implementacija je odrađena u okviru biblioteke Sckit-learn, u modulu neural_network, kroz klasu MLPClassifier. Klasa MLPC je korišćena, jer je namenjena za primenu u klasifikacionim problemima, što je slučaj u ovom radu.

Nakon implementacije, dobijena je konačna matrica konfuzije i procenat tačno predviđenih uzoraka. Iz matrice konfuzije možemo videti broj tačno predviđenih, na glavnoj dijagonali, dok ostali elementi matrice predstavljaju pogrešno klasifikovane uzorke. Procenat tačno predviđenih uzoraka iznosi 91,48% i može se bolje videti iz toplotne mape (Slika 4.).

Određeni su parametri – tačnost i osetljivost (Tabela 1.), za svaku od 3 klase. Prosečna tačnost iznosi 94,32%, a prosečna osetljivost 91, 93%.

	Tačnost [%]	*Osetljivost [%]*
Kolačići	92.69	91.98
Testa	91.89	88.37
Pice	94.32	91.93

Tabela 1. Tačnost i osetljivost

Zatim je ponovo urađena unakrsna validacija sa 10 podskupova za prvi klasifikator, ali na modifikovanim podacima, kako bismo dobili što bolje rezultate.

Upsampling-om je veštački povećan broj uzoraka, tako što su svi uzorci koji nisu iz klase kolačića, poduplani. Ovo je urađeno da bi neuralna mreža češće uzimala u obzir uzorke manje zastupljenih klsa.

Ponovo su određeni parametri tačnost i osetljivost za svaku klasu (Tabela 2.), dok prosečna tačnost iznosi 95,98%, a prosečna osetljivost 93, 89%. Veći procenat tačnosti i osetljivosti ukazuje na to da je novi model dao bolje rezultate.

Rezultati predviđanja drugog modela, koje iznosi 93,97%, prikazani su na toplotnoj mapi, izraženi u procentima (Slika 5.).

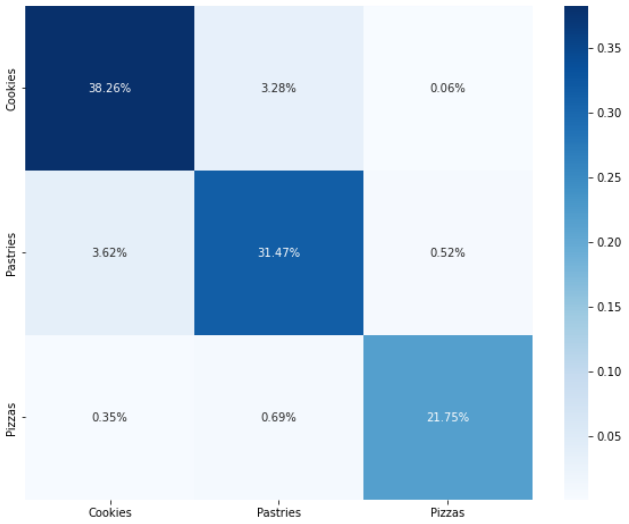
U Tabeli 3. prikazani su rezultati matrica konfuzija i tačno predviđenih uzoraka prvog i drugog modela, čime ih možemo uporediti.

	Tačnost [%]	*Osetljivost [%]*
Kolačići	94.55	90.04
Testa	94.33	93.78
Pice	99.05	97.85

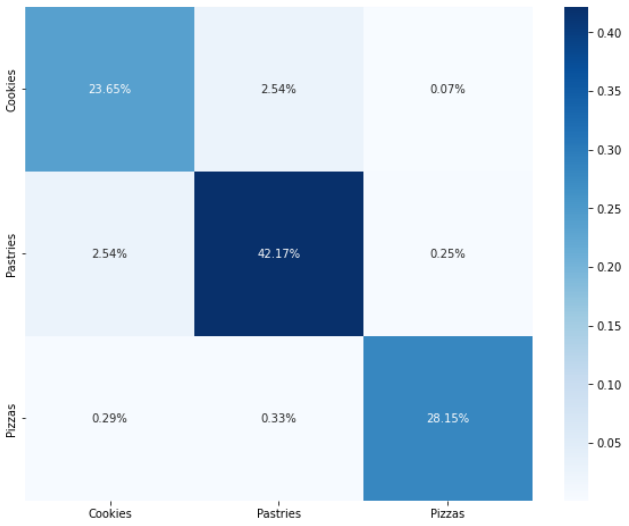
Tabela 2. Tačnost i osetljivost

	Prvi MLPC	*Drugi MLPC*
Matrica konfuzije	655 57 1 63 547 9 6 12 378	651 70 2 70 1161 7 8 9 775
Tačno predviđeni uzorci	91,48%	93,97%

Tabela 3. Rezultati prvog i drugog modela



Slika 4. Prikaz koeficijenta korelacije pomoću toplotne mape za prvi model



Slika 5. Prikaz koeficijenta korelacije pomoću toplotne mape za drugi model

Obučavanjem klasifikatora na test skupu, procenat pogođenih uzoraka iznosi 93,78%. Upoređivanjem rezultata unakrsne validacije (93,97%), sa rezultatima urađenim na test skupu, dobijen je približno isti rezultat, pa se može zaključiti da je klasifikator dobro obučen.

2) Mašina na bazi vektora nosača - SVM

Unakrsnom validacija sa 10 podskupova obučen je drugi klasifikator – SVM, na trening podacima. Iterativnom metodom dolazimo do iteracije sa najvećom tačnošću, te iz te iteracije dolazimo do najboljih parametara za ovaj klasifikator. Obukom modela na trening podacima, dolazimo do rezultata (Tabela 4.). Klasifikator je takođe obučen i na test skupu podataka.

Matrica konfuzije	*672*	*51*	*0*
	49	*559*	*11*
	10	*12*	*373*
Procenat pogođenih uzoraka	92,29%		

Tabela 4. Rezultati obuke modela

Određeni su parametri – tačnost i osetljivost (Tabela 5.), za svaku od 3 klase. Prosečna tačnost iznosi 94,32%, a prosečna osetljivost 91, 93%.

	Tačnost [%]	*Osetljivost [%]*
Kolačići	93.67	92.95
Testa	92.86	90.31
Pice	98.04	94.19

Tabela 5. Tačnost i osetljivost

IV. Rezultati

Kao konačne mere uspešnosti klasifikatora sa više klasa, računaju se prosečna tačnost, stopa greške, osetljivost, preciznost i F-mera.

U Tabeli 6. prikazani su mere uspešnosti klasifikatora, dobijene obukom modela neuronskih mreža, kao i mašine na bazi vektora nosača.

[%]	Neuronske mreže	SVM
Prosečna tačnost	95.85	97.59
Mikro preciznost	93.78	96.37
Mikro F-mera	93.78	96.37
Makro preciznost	94.73	97.02
Makro F-mera	94.36	96.47
Procenat pogođenih uzoraka	93.78	96.37

Tabela 6. Upoređene vrednosti obuke modela na trening uzorcima

[%]	Neuronske mreže	SVM
Procenat pogođenih uzoraka	93.78	96.37
Mikro preciznost	93.78	96.37
Makro preciznost	94.73	97.02
Mikro osetljivost	93.78	96.37
Makro osetljivost	93.98	95.92
Mikro F-mera	93.78	96.37
Makro F-mera	94.34	96.42

Tabela 7. Upoređene vrednosti obuke modela na test uzorcima

V. Zaključak

Upoređivanjem konačnih mera uspešnosti klasifikatora trening i test uzoraka, dolazimo do zaključka da se bolji rezultati postižu obučavanjem mašine na bazi vektora modela.