Analiza podataka o receptima

Aleksandra Mitro, IN 8/2018, sandramitro99@gmail.com

I. Uvod

Izveštaj se bavi analizom prisustva odnosno odsustva sastojaka u receptima i njihovim poreklom. Baza podataka sadrži 10566 recepata koji su južno-američkog, francuskog, grčkog, meksičkog, italijanskog, japanskog, kineskog, tajlanskog i britanskog porekla. Posmatrano je prisustvo 150 različitih sastojaka poput soli, ulja, šećera, različitih vrsta mesa, začina, voća i povrća. Analiziranje ovih podataka i uočavanje eventualne pravilnosti i zavisnosti između atributa bi moglo biti dobra teorijska osnova za projektovanje klasifikatora koji će recepte klasifikovati prema poreklu na osnovu njegovih sastojaka.

II. ANALIZA PODATAKA

U okviru posmatrane baze podataka nema nedostajajućih podataka stoga nijedan od uzoraka nije izostavljen. Uklonjena je kolona Unnammed koja nije korisna prilikom analize podataka jer predstavlja oznaku recepta. Prosečan broj sastojaka po receptu je 12.

A. Raspodela porekla recepata u bazi podataka

SLIKA 1 : Prikaz broja recepata u zavisnosti od porekla
Sa slike 1 uočavamo da je najviše recepata u bazi podataka južno-američkog porekla, dok je najmanje recepata britanskog porekla.

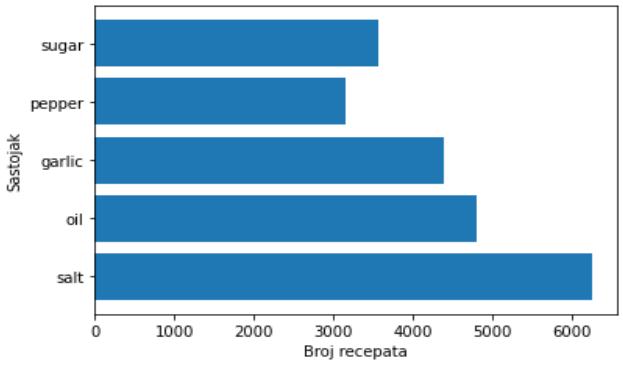
B. Pristustvo/odustvo sastojaka u zavisnosti od porekla

Poreklo	**Najređe korišćen sastojak**	**Najčešće korišćen sastojak**
Južno-američko	Susam	So
Francusko	Origano	Sos
Grčko	Riblje ulje	So
Meksičko	Susamovo ulje	Ulje
Italijansko	Riblje ulje	Ulje
Japansko	Origano	Sos
Kinesko	Mirin	So
Tajlandsko	Tortilje	So
Britansko	Prašak za pecivo	Sos

TABELA 1 : Prikaz najređe i najčešće korišćenih sastojaka u zavisnosti od porekla

Iz tabele 1 uočavamo da su so i sos najčešće korišćeni sastojci u većini recepata bez obzira na njihovo poreklo. Takođe primećujemo da je u japanskoj ishrani najređe korišćen sastojak origano što je očekivano jer je origano začin karakterističan za južnu Evropu, dok su sa druge strane tortilje najređe korišćen sastojak u tajlandskoj ishrani što je očekivano jer su tortilje karakteristične za južno-američko podneblje.

C. Najčešće korišćeni sastojci u receptima

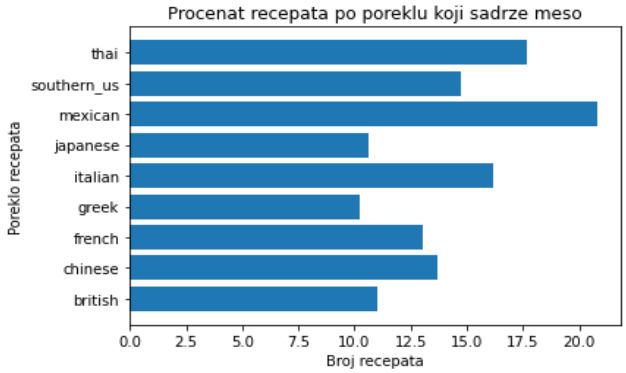


SLIKA 2 : Pet najčešće korišćeni sastojci u receptima
Sa slike 2 uočavamo da je so najčešće korišćen sastojak u receptima, a zatim slede ulje, luk, šećer i biber, što je očekivano uzimajući u obzir da su prethodno navedeni sastojci osnovni sastojci većine jela i predstavljaju njihovu osnovu, bez obzira na njihovo poreklo.

D. Recepti u kojima se koristi meso

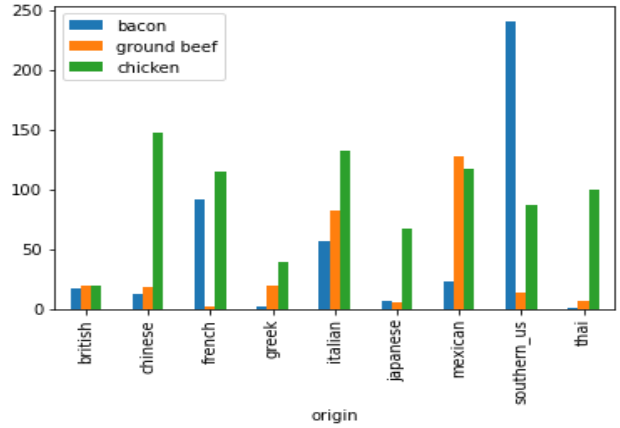
U korišćenoj bazi podataka 14.75% recepata sadrži neku od sledećih vrsta mesa : govedina, slanina i više vrsta

piletine.



SLIKA 3 : Prikaz procenata recepata po poreklu koji sadrže meso

Sa slike 3 uočavamo da se meso najčešće koristi u meksičkoj kuhinji dok se najređe koristi u grčkoj kuhinji, takođe se meso veoma retko koristi i u japanskoj kuhinji što je posledica toga što je u grčkoj i japanskoj kuhinji riba zastupljenija od mesa, a koristi se i povrće u izobilju.



SLIKA 4 : Prikaz zastupljenostii različitih vrsta mesa u receptima u zavisnosti od porekla

Sa slike 4 uočavamo da je piletina najčešće korišćeno meso u svim kuhinjama sem u južno-američkoj i meksičkoj. U južno-američkoj kuhinji je najzastupljenije meso slanina, što je posledica toga što je upravo u južno-američkoj kuhinji svinjsko meso najzastupljenije, dok je u meksičkoj kuhinji najzastupljenija govedina koja se koristi u nekim od najpopularnijih meksičkih jela poput Takosa, Čili kon karne-a i raznih varijanti jela sa tortiljama.

III. KLASIFIKATORI

A. *Neuronska mreža*

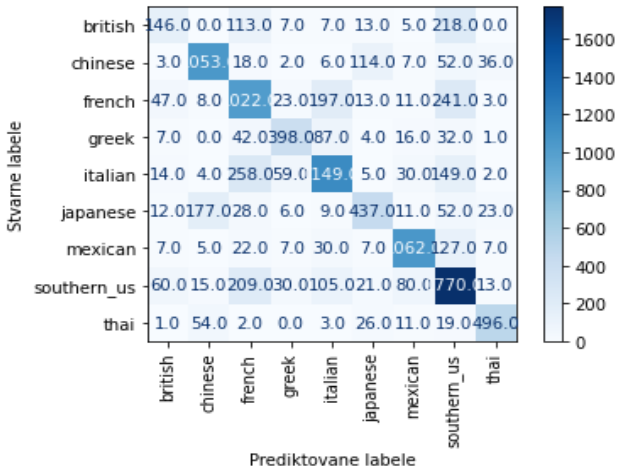
Neuronske mreže su inspirisane procesom učenja koji se odvija u ljudskom mozgu. Sastoje se od veštačke mreže funkcija nazvanih parametri, koji omogućavaju računaru da uči i da se podešavaju analizirajući nove podatke. Svaki parametar, koji se drugačije naziva i neuronom, je funkcija koja proizvodi izlaz, nakon što primi jedan ili više ulaza. Prethodno spomenuti izlazi se zatim prosleđuju sledećem sloju neurona, koji ih koriste kao ulaze i proizvode dalje izlaze. Ovaj proces se ponavlja sve dok se svaki sloj neurona ne uzme u obzir, a terminalni neuroni ne dobiju svoj ulaz. Izlaz terminalnih neurona predstavlja konačni rezultat za model.

Prilikom obuke klasifikatora korišćenjem neuronskih mreža korišćena je unakrsna validacija, a u sledećoj tabeli je prikazan procenat tačno predviđenih uzoraka u zavisnosti od različitih parametara koji su menjani a to su broj skrivenih slojeva neuronske mreže, broj neurona u skrivenim slojevima i maksimalan broj iteracija.

	Broj skrivenih slojeva	**Broj neurona**	**Maksimalan broj iteracija**	**Procenat tačno predviđenih uzoraka**
1	3	64	100	70.67%
2	5	64	100	71.07%
3	3	72	100	70.80%
4	5	72	100	70.56%
5	3	82	100	71.02%
6	5	82	100	70.68%
7	3	50	100	71.29%
8	5	50	100	71.08%

TABELA 2 : Prikaz uspešnosti klasifikacije novih uzoraka u zavisnosti od promene parametara

Iz tabele 2 uočavamo da je najbolji procenat tačno predviđenih uzoraka imala neuronska mreža sa tri skrivena sloja, sa po pedeset neurona i najviše sto iteracija i tada procenat tačnosti klasifikacije novih uzoraka iznosi 71.29%.



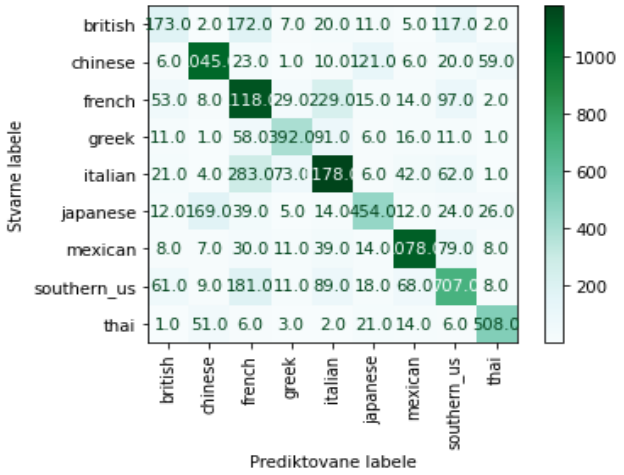
SLIKA 5 : Matrica konfuzije klasifikatora

Sa slike 5 na kojoj je prikazana matrica konfuzije uočavamo da klasifikator pravi greške prilikom klasifikacije recepata koji potiču sa sličnih podneblja što se tiče ishrane poput italijanske i francuske kuhinje, kao i kineske i japanske kuhinje. Takođe uočavamo da je velik broj recepata klasifikovan kao južno-američki pošto ih ima najviše u bazi podataka, njih čak 2303 recepta dok za njima slede italijanski recepti kojih ima 1670. Nebalansirani skupovi podataka često dovode do loših performansi klasifikatora, stoga ćemo u nastavku primeniti smanjivanje broja uzoraka najzastupljenije klase tako što će biti uklonjen svaki drugi uzorak koji pripada klasi južno-američkih recepata, gde su prilikom obuke korišćeni isti parametri kao i pre obuke nebalansiranih podataka.

	Broj skrivenih slojeva	Broj neurona	Maksimalan broj iteracija	Procenat tačno previđenih uzoraka
1	3	64	100	70.66%
2	5	64	100	70.80%
3	3	72	100	70.80%
4	5	72	100	70.16%
5	3	80	100	70.26%
6	5	80	100	70.28%
7	3	50	100	70.31%
8	5	50	100	70.44%

TABELA 3 : Prikaz uspešnosti klasifikacije novih uzoraka u zavisnosti od promene parametara

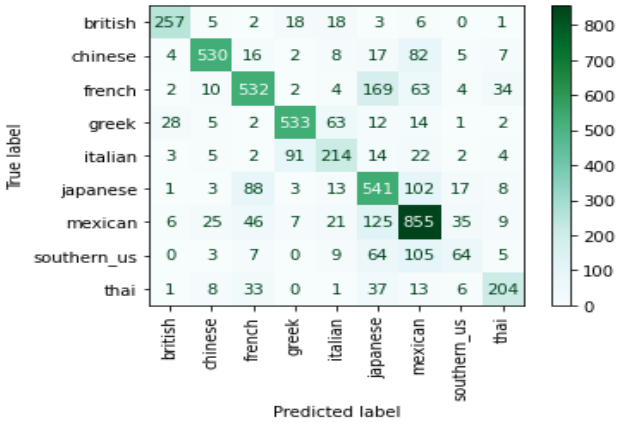
Iz tabele 3 uočavamo da najveći procenat uspešnog klasifikovanja novih uzoraka ima neuronska mreža sa 5 skrivenih slojeva i 64 neurona i neuronska mreža sa 3 skrivena sloja i 72 neurona. Takođe uočavamo da izbacivanjem svakog drugog uzorka iz klase južno-američkih recepata nemamo napredak u pogledu tačnosti klasifikovanja novih uzoraka.



SLIKA 6 : Matrica konfuzije klasifikatora

Sa slike 6 na kojoj je prikazana matrica konfuzije klasifikatora uočavamo da klasifikator pravi greške prilikom klasifikacije recepata koji potiču sa sličnih podneblja međutim ima poboljšanja u odnosu na prethodni klasifikator u broju recepata koji se pogrešno klasifikuju kao južno-američki što je posledica uklanjanja polovine južno-američkih recepata.

Prilikom obučavanja modela korišćen je ceo trening skup, a parametri koji su korišćeni su da je broj skrivenih slojeva 3, broj neurona 50 i broj maksimalnih iteracija 100, pošto su ovi parametri dali najbolje rezultate prilikom unakrsnevalidacije.



SLIKA 7 : Matrica konfuzije obučene neuronske mreže

Sa slike 7 uočavamo da neuronska mreža najviše grešaka pravi prilikom klasifikacije južno-američkih recepata, gde čak 105 recepata klasifikuje kao meksičke. Prosečna tačnost po klasi iznosi 93.41%, dok je prosečna osetljivost po klasi 67.09%.

Poreklo	Tačnost	Osetljivost
Južno-američko	87.11%	75.73%
Francusko	87.20%	67.72%
Grčko	96.80%	67.33%
Meksičko	96.12%	78.99%
Italijansko	90.84%	64.88%
Japansko	94.69%	59.94%
Kinesko	95.27%	80.75%
Tajlandsko	98.14%	82.90%
Britansko	95.02%	24.90%

TABELA 4 : Tačnost i osetljivost po klasama

Iz tabele 4 uočavamoda neuronska mreža ostvaruje najveću tačnost prilikom klasifikacije tajlandskih recepata, a najmanju za francuske recepte što možemo uočiti na matrici konfuzije gde se čak 169 francuskih recepata pogrešno klasifikuje kao japanski. Najveću osetljivost imaju tajlandski recepti dok najmanju imaju britanski recepti, i njihova osetljivost je značajno manja od ostalih.

B. Mašine na bazi vektora nosača (SVM)

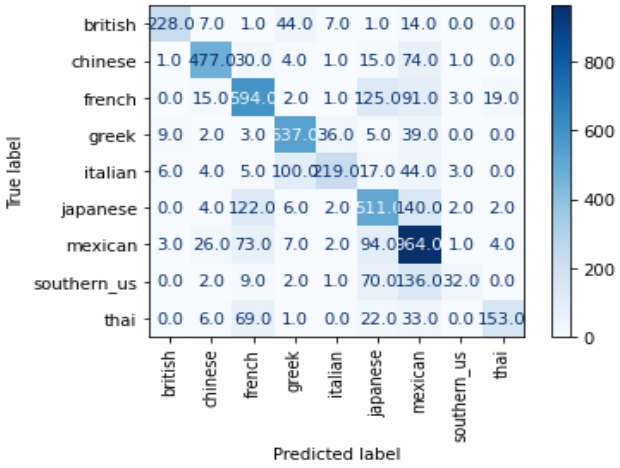
Mašina na bazi vektora nosača je relativno jednostavan algoritam nadgledanog učenja koji se može koristiti za probleme klasifikacije i regresije. Cilj SVM algoritma je da stvori najbolju liniju ili granicu odlučivanja,koja se takođe naziva i hiperravan, koja može da razdvoji n-dimenzionalni prostor u klase tako da možemo lako da stavimo novu tačku podataka u ispravnu kategoriju u budućnosti. SVM bira ekstremne tačke/vektore koji pomažu pri kreiranju hiperravni. Pomenuti vektori se nazivaju vektori nosači po kojima je pomenuti algoritam i dobio naziv. U poređenju sa novijim algoritmima poput prethodno obrađenih neuronskih mreža, SVM ima dve glavne prednosti a to su : veća brzina i bolje performanse sa ograničenim brojem uzoraka (u hiljadama). Prilikom obuke klasifikatora korišćenjem SVM algoritma menjani su sledeći parametri : one-versus-rest (jedan protiv svih), one-versus-one (svako protiv svakog), parametar break_ties(koji je postavljen na vrednost True(1) što znači

da će se odluka doneti na osnovu toga koliko je klasifikator siguran u svoju odluku), takođe menjan je parametar C, koji određuje toleranciju na pogrešnu klasifikaciju, a kreće se u vrednostima [1,5,10,20] i vrsta kernela(rbf i linearan). U nastavku su prikazani procenti tačno predviđenih uzoraka za najuspešnije kombinacije parametara za različite pristupe.

Različiti pristupi	Parametri	Procenat tačno predviđenih uzoraka
OVR	C = 1 Kernel = rbf	70.32%
OVO	C = 1 Kernel = rbf	70.32%
OVR sa break_ties	C = 5 Kernel = rbf	70.26%

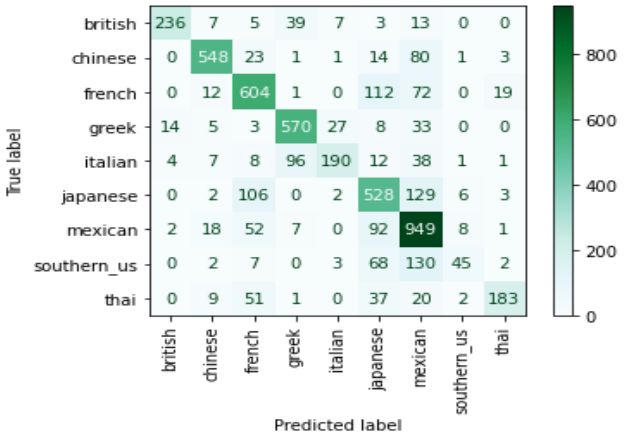
TABELA 5 : Procenat tačno predviđenih uzoraka SVM

Iz tabele 5 uočavamo da najbolje rezultate SVM klasifikator postiže korišćenjem OVR pristupa, gde parametar C iznosi 1, a kernel rbf. U nastavku je prikazana matrica konfuzije nakon unakrsne validacije sa pomenutim parametrima. Sa slike 8 možemo uočiti da se puno japanskih recepata klasifikuje kao francuski i obrnuto, kao i pogrešne klasifikacije italijanskih recepata kao grčkih što je posledica korišćenja sastojaka koji su specifični za mediteran.



SLIKA 8 : Matrica konfuzije nakon unakrsne validacije

Prilikom obučavanja modela korišćen je ceo trening skup, sa prethodno istaknutim parametrima koji daju najbolji procenat tačno predviđenih uzoraka.



SLIKA 9 : Matrica konfuzije nakon obuke

Sa slike 9 uočavamo da klasifikator nakon obuke najviše grešaka pravi prilikom klasifikacije južno-američkih recepata gde čak 130 klasifikuje kao meksičke, što je posledica istih sastojaka koji se koriste na tom podneblju. Takođe, u manjem obimu, imamo i pogrešnu klasifikaciju japanskih i francuskih recepata kao i klasifikaciju italijanskih recepata kao grčkih. Prosečna tačnost po klasi iznosi 93.98%, dok je prosečna osetljivost po klasi 66.78%

Poreklo	Tačnost	Osetljivost
Tajlandsko	98.22%	76.13%
Meksičko	96.49%	81.67%
Italijansko	91.08%	73.66%
Kinesko	95.51%	86.36%
Japansko	96.08%	53.22%
Francusko	88.75%	68.04%
Južno-američko	86.44%	84.06%
Britansko	95.64%	17.51%
Grčko	97.18%	60.39%

TABELA 6 : Tačnost i osetljivost za svaku od klasa

Iz tabele 6 uočavamo da najbolju tačnost klasifikacije imaju tajlandski recepti, dok najmanju tačnosti imaju južno-američki recepti što smo mogli uočiti iz matrice konfuzije. Takođe, najveću osetljivost imaju kineski recepti, dok najmanju osetjivost imaju britanski recepti čija je osetljivost značajno manja od osetjivosti ostalih klasa.

IV. ZAKLJUČAK

U radu su korišćeni klasifikatori na bazi neuronskih mreža i mašina na bazi vektora nosača. U nastavku je prikazana tabela gde su upoređene mere uspešnosti za obučene modele.

Mera uspešnosti	Neuronska mreža	SVM
Procenat pogođenih uzoraka	70.60%	72.93%
Preciznost mikro	70.60%	72.93%
Preciznost makro	70.52%	77.52%
Osetljivost mikro	70.60%	72.93%
Osetljivost makro	67.24%	66.78%
F-mera mikro	70.60%	72.93%
F-mera makro	68.27%	69.44%

TABELA 7 : Upoređivanje mera uspešnosti obučenih modela

Iz tabele 8 uočavamo da SVM klasifikator daje bolje rezultate od neuronske mreže po svim parametrima, iako se ti parametri ne razlikuju mnogo razlike možemo najbolje uočiti na matricama konfuzije obučenih modela.