Univerzitet u Beogradu

Elektrotehnički fakultet



**What's Cooking? (Kernels Only)**

Projektni zadatak iz predmeta Softversko inženjerstvo velikih baza podataka

Mihailo Tomašević, 2023/3081

Beograd, januar 2024.

Projektni zadatak

Projektni zadatak predstavlja jedan od problema sa sajta <https://www.kaggle.com>. Konkretno, u pitanju je problem <https://www.kaggle.com/competitions/whats-cooking-kernels-only>.

Problem spada u kategoriju višeklasne klasifikacije. Zadatak je odrediti kojem tipu kuhinje pripada zadato jelo na osnovu njegovih sastojaka.

Ulazne podatke predstavljaju 2 fajla: ***train.json*** i ***test.json***.

Fajl *train.json* sadrži niz JSON objekata. Svaki objekat predstavlja jedno jelo i sadrži 3 polja:

* ’’id’’ – ceo broj koji predstavlja jedinstveni identifikator datog jela
* ’’cuisine’’ – string koji predstavlja kuhinju kojoj jelo dato pripada
* ’’ingredients’’ – niz stringova koji predstavljaju sastojke koji čine dato jelo

Ovaj fajl nam služi da se istrenira model kako bi našli vezu između sastojaka i kuhinje.

Fajl *test.json* sadrži niz JSON objekata. Svaki objekat predstavlja jedno jelo i sadrži 2 polja:

* ’’id’’ – ceo broj koji predstavlja jedinstveni identifikator datog jela
* ’’ingredients’’ – niz stringova koji predstavljaju sastojke koji čine dato jelo

Podaci koji se nalaze u ovom fajlu se prosleđuju modelu, koji je prethodno istreniram podacima iz fajla *train.json* i pomoću njega se, na osnovu sastojaka koji čine jelo, predviđa kojoj od kuhinja iz fajla *train.json* to jelo pripada.

Rešenje predstavlja fajl ***submission.csv*** koji sadrži 2 kolone:

* ’’id’’ – ceo broj koji predstavlja jedinstveni identifikator jela iz fajla *test.json*
* ’’cuisine’’ – string koji predstavlja kuhinju kojoj smo predvideli da jelo pripada

Primer izlaznog fajla:

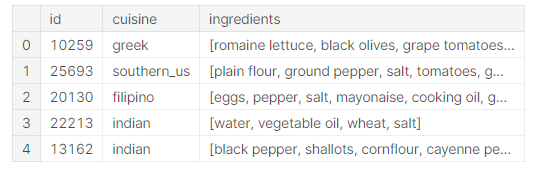
id,cuisine  
35203,italian  
17600,italian  
35200,italian  
17602,italian  
...  
etc.

Kvalitet modela se određuje na osnovu procenta tačno klasifikovanih jela to jest procenta jela za koje smo tačno predvideli kojoj kuhinji pripadaju. Evaluacija se vrši tako što se dobijeno rešenje preda preko kaggle notebooka-a, a kao odgovor se dobije procenat tačno klasifikovanih jela.

Ulazni podaci

U ovoj sekciji su prikazane različite informacija koje se mogu dobiti na osnovu ulaznih fajlova kako bi se stekao bolji utisak o podacima koji se koriste za obučavanje modela pomoću kog će biti vršena klasifikacija i o podacima koji će biti klasifikovani.

U fajlu *train.json* se nalaze podaci o 39774 jela. Na slici 1 je prikazano prvih pet linija fajla *train.json.*



***Slika 1***

U fajlu *test.json* se nalaze podaci o 9944 jela. Na slici 2 je prikazano prvih pet linija fajla *test.json.*

A screenshot of a recipe

Description automatically generated

***Slika 2***

Na slici 3 je grafički prikazan broj jela koji pripadaju svakoj od kuhinja iz fajla *train.json*.

A graph of a bar graph

Description automatically generated with medium confidence

***Slika 3***

Na slici 4 je grafički prikazano koliko ima jela koja sadrže određeni broj sastojaka, za jela iz fajla *train.json*.

A graph of a number of years

Description automatically generated with medium confidence

***Slika 4***

Na slici 5 je grafički prikazano 20 sastojaka koji se najčešće pojavljuju u jelima iz fajla *train.json* i broj njihovih pojavljivanja.

A graph of different colored bars

Description automatically generated

***Slika 5***

# Rešenje

Za rešavanje ovog problema su iskorišćena 3 modela mašinskog učenja: naivni Bajes ***(eng. Naive Bayes*),** metoda potpornih vektora ***(eng. Support Vector Machine)*** i neuralne mreže ***(eng. Neural Networks*).**

Svaki od ovih modela je dao različite rezultate po pitanju procenta ispravno klasifikovanih jela.

Pri korišćenju svakog od ovih modela, nad podacima je izvršen proces pretprocesiranja.

## Pretprocesiranje

Pre nego što se podaci pošalju modelu na treniranje, nad njima se vrši proces pretprocesiranja. Ovaj proces se vrši pomoću funkcije koja je prikazana na slici 6.

Funkcija kao parametar prima string koji se sastoji iz niza sastojaka odvojenih razmakom i nad njim vrši određene modifikacije. Ova funkcija se primenjuje na svako polje u koloni *’’ingredients’’*  koja pripada fajlu *train.csv.*

Funkcija radi sledeće:

* sva velika slova u stringu pretvara u odgovarajuća mala slova
* iz stringa se izbacuju svi karakteri iz sledećeg skupa: ***’’,’’ , ’’.’’, ’’\’’, ’’?’’, ’’:’’, ’’(’’, ’’)’’, ’’ ’’ ’’*** jer nisu relevantni za proces klasifikacije
* nakon toga iz stringa se izbacuju sve reči koje sadrže karaktere koji nisu slova
* na kraju se iz stringa izbacuju reči koje ne predstavljaju same sastojke već samo njihovu meru (masu ili zapreminu)
* funkcija na kraju vraća modifikovani string koji sadrži niz sastojaka odvojenih razmakom

import re

import pandas as pd

import unidecode as unidecode

def preprocess(ingredients):

ingredients = ingredients.lower()

ingredients = re.sub('[,\.!?:()"]', '', ingredients)

ingredients = re.sub('[^a-zA-Z"]', ' ', ingredients)

ingredients = ingredients.replace('-', ' ')

words = []

for word **in** ingredients.split():

word = re.sub((r'\b(oz|ounc|ounce|pound|lb|inch|inches|kg|to)\b'), ' ', word)

if len(word) <= 2: continue

word = unidecode.unidecode(word)

if len(word) > 0: words.append(word)

return ' '.join(words)

***Slika 6***

## Naivni Bajes

Ovaj model mašinskog učenja je najjednostavniji od korišćenih. On pretpostavlja da su sastojci u jelu međusobno nezavisni što možda nije slučaj kod ovog problema. To je jedan od razloga što je ovo model koji je dao najlošije rezultate. Prednosti ovog modela su to što je jednostavan i vrlo brz i efikasan prilikom izračunavanja.

Pri realizaciji je korišćen model *MultinomialNB* iz Python paketa *sklearn*.

Najbolji rezultat koji je ovaj model dao iznosi ***72,274%*** ispravno klasifikovanih jela.

Na slici 7 su prikazani rezultati koje je ovaj model dao u vidu grafa koji prikazuje koliko jela je klasifikovano da pripada svakoj od kuhinja.

A graph with different colored bars

Description automatically generated

***Slika 7***

## Metoda potpornih vektora

Ovaj model mašinskog učenja postiže najbolji balans između kompleksnosti i robusnosti. On pronalazi hiperravan koja na najbolji način deli tačke u višedimenzionom prostoru na različite klase. Tačke predstavljaju jela, dok klase predstavljaju kuhinje. Koriste se različite kernel funkcije koje efikasno prave podele koje se ne bi mogle izvršiti pomoću linearnih funkcija. Veoma je bitno izabrati odgovarajuće kernel funkcije. Iako se podele vrše na kompleksan način i obavlja se dosta zahtevnih izračunavanja, model se efikasno štitit od preobučavanja.

Iz navedenih razloga ovaj model je dao najbolje rezultate. Jedine mane ovog modela su to što je računski dosta zahtevan i što odabir odgovarajuće hiperravni može imati dosta velikih uticaj na performanse i zato izvršavanje traje dosta duže nego kada se koriste druga dva modela.

Pri realizaciji je korišćen model *SVC* iz Python paketa *sklearn*.

Najbolji rezultat koji je ovaj model dao iznosi ***80,044%*** ispravno klasifikovanih jela.

Na slici 8 su prikazani rezultati koje je ovaj model dao u vidu grafa koji prikazuje koliko jela je klasifikovano da pripada svakoj od kuhinja.

A graph with different colored bars

Description automatically generated

***Slika 8***

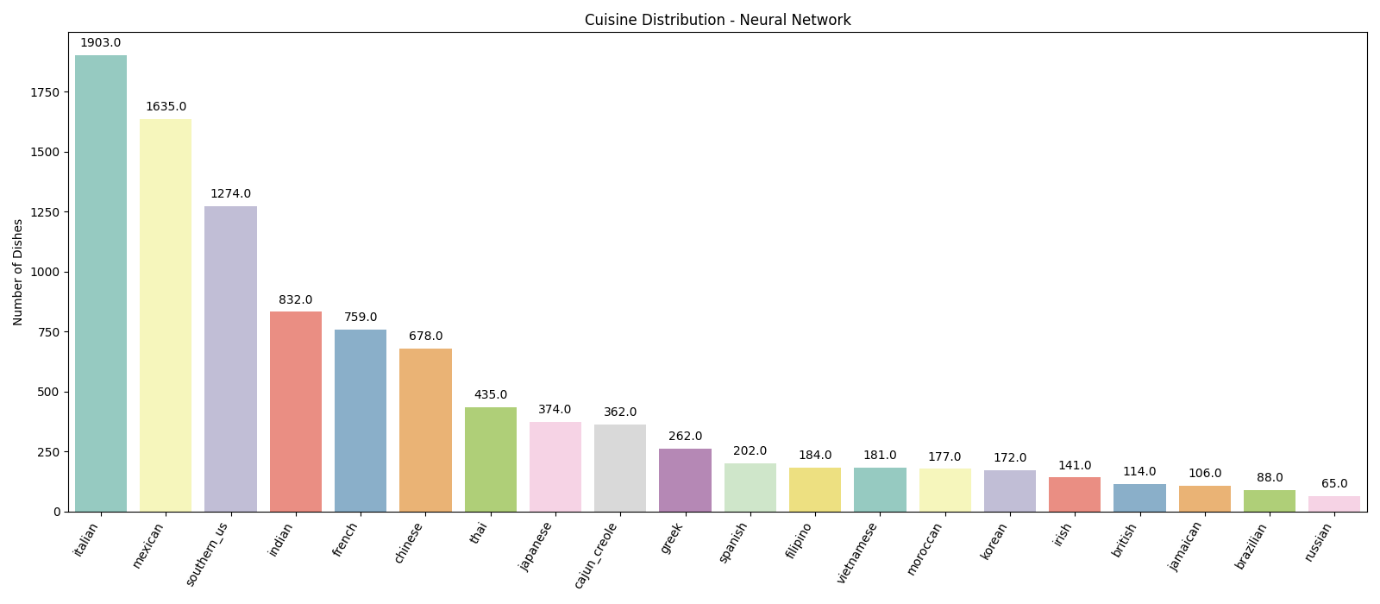
## Neuralna mreža

Ovaj model mašinskog učenja je najkompleksniji od korišćenih. Sastoji se iz više slojeva međusobno povezanih neurona i u stanju je da pronađe kompikovane veze između podataka ukoliko su parametri dobro podešeni. Pri korišćenju ovog modela, podaci iz fajla *train.csv* su podeljeni na skup za treniranje i skup za validaciju i mreža je obučavana u više epoha sve dok se u narednoj epohi dobijaju bolji rezultati nego u prethodnoj. Ovaj model zahteva dosta veliku količinu podataka ukoliko se želi da klasifikacija bude kvalitetna. Kao u prethodnom modelu, i u ovom slučaju se vrši dosta računski zahtevnih operacija i od ključnog je značaja odabrati dobre vrednosti parametara i dobru arhitekturu neuralne mreže.

Pri realizaciji su korišćeni slojevi neuralnih mreža iz Python paketa *keras*.

Najbolji rezultat koji je ovaj model dao iznosi ***75,241%*** ispravno klasifikovanih jela.

Na slici 9 su prikazani rezultati koje je ovaj model dao u vidu grafa koji prikazuje koliko jela je klasifikovano da pripada svakoj od kuhinja



***Slika 9***

## Poređenje rezultata

U tabeli 1 su upoređeni rezultati koje su dala opisana 3 modela mašinskog učenja. Za svaku kuhinju je navedeno koliko jela je svaki od modela klasifikovao da toj kuhinji pripada.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Naive Bayes | Support Vector Machine | Neural Network |
| Italian | 2568 | 2061 | 1903 |
| Mexican | 1788 | 1624 | 1635 |
| Southern\_US | 1663 | 1166 | 1274 |
| Indian | 904 | 802 | 832 |
| French | 587 | 756 | 759 |
| Chinese | 972 | 731 | 678 |
| Thai | 457 | 399 | 435 |
| Japanese | 203 | 295 | 374 |
| Cajun\_Creole | 294 | 374 | 362 |
| Greek | 126 | 289 | 262 |
| Spanish | 30 | 174 | 202 |
| Filipino | 35 | 150 | 184 |
| Vietnamese | 59 | 191 | 181 |
| Moroccan | 76 | 177 | 177 |
| Korean | 87 | 177 | 172 |
| Irish | 14 | 111 | 141 |
| British | 22 | 176 | 114 |
| Jamaican | 20 | 97 | 106 |
| Brazillian | 30 | 100 | 88 |
| Russian | 9 | 94 | 65 |
| Total | 9944 | 9944 | 9944 |

***Tabela 1***

Naivni Bajes je najjednostavniji model sa najmanje kompleksnom podelom. Iz toga razloga je dosta više jela klasifikovano da pripada kuhinjama koje imaju najviše jela, u odnosu na metodu potpornih vektora i neuralnu mrežu. Kod metode potpornih vektora i neuralne mreže je klasifikacija uniformnija, s tim što metoda potpornih vektora ima oko 5% uspešniju klasifikaciju kao što je prikazano u prethodnom poglavlju.