Klasifikacija recepata Stablo odluke i logistička regresija

Sara Popov, IN 41/2017, popov.sara@uns.arc.rs

I. OPIS PROBLEMA

Klasifikatori koji će biti implementirani i opisani u ovom radu će na osnovu prisustvu ili odsustvu određenih sastojaka odlučiti da li uzorak predstavlja recept za kolače, peciva ili pice. Korišćeni su klasifikator logističke regresije i stabla odluke, takođe na kraju su upoređene njihove uspešnosti.

Tačnost je korišćena kao mera uspešnosti, jer su podjednako bitne klase, odnosno podjednako je loša greška klasifikovati, na primer, klasu kolača u klasu peciva, tako i za ostale.

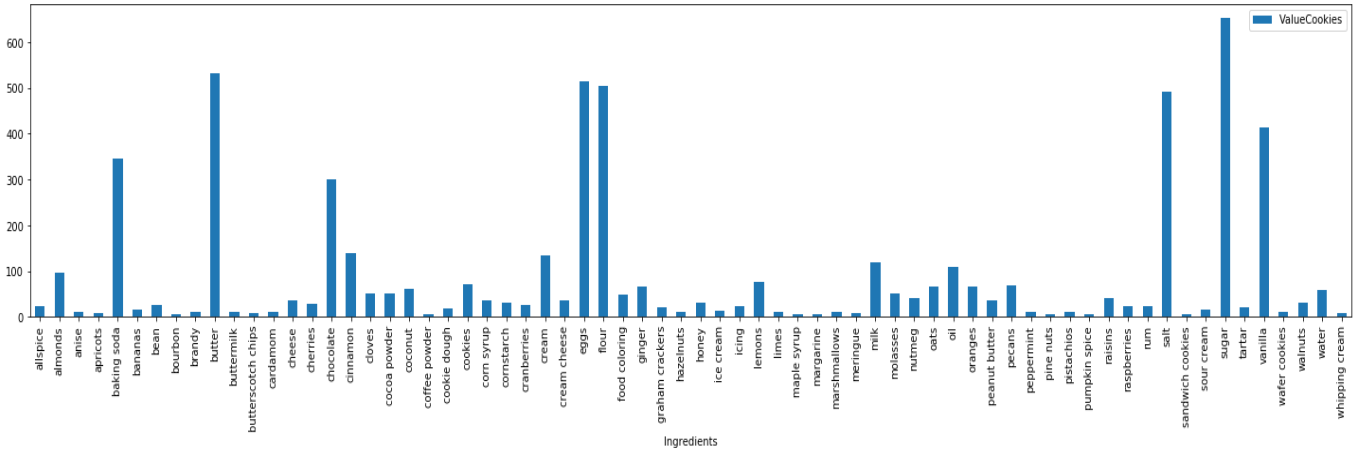
II. BAZA PODATAKA

Inicijalno je baza podataka podeljena i na osnovu toga postoje trening i test skup. Trening skup se sastoji iz 1738 uzoraka, dok test skup ima 193 uzoraka. Oba skupa imaju po 134 obeležja. Obeležja, sva sem poslednjeg, predstavljaju sastojak koji je potencijalno potreban za recept, a to da li je on stvarno potreban za recept ili ne se predstavlja kao vrednost obeležja, gde 0 vrednost predstavlja da taj sastojak nije deo recepta, dok 1 predstavlja da je sastojak potreban za recepta. Poslednje obeležje predstavlja klasnu labelu, u kojoj se nalazi informacija o tome da li je dati recept za kolače, pecivo ili picu. Kako je cilj ovog klasifikatora da na osnovu sastojaka koji imaju vrednost 1, odnosno potrebni su za recept, klasifikuje za šta je dati recept, poslednje obeležje predstavlja izlaz klasifikatora, odnosno podatak kojoj klasi uzorak zaista pripada. Upravo na osnovu grupisanja poslednjeg obeležja je moguće uočiti da u skupu za obuku nije podjednaka zastupljenost recepata za sve tri klase. Od svih uzoraka u trening skupu čak 41.59% njih predstavljaju recepte za kolače, dok 35.61% njih predstavljaju recepte za peciva, a najmanje recepata ima za pice, 22.78%.

Ni u trening ni u test skupu se ne javljaju nedostajuće vrednosti i svako obeležje ima pozitivnu vrednost u bar jednom uzorku, tako da obeležja koja mogu biti izbačena selekcijom nisu očigleda. Na osnovu toga obeležja nisu selektovana i samim tim nije ni smanjena dimenzionalnost.

Radi lakšeg vizuelne koncepcije učestalosti pojavljivanja određenih sastojaka za svaku klasu je predstavljen histogram koji upravo opisuje broj pojavljivanja svakog sastojka za svaku klasu. Usled velikog broja obeležja i zbog bolje vizuelizacije za svaku klasu su izbačena ona obeležja koja se ni jednom ne pojavljuju kao potreban sastojak za tu klasu.

Na osnovu histograma moguće je uočiti da obeležja, odnosno sastojci, koji se potencijalno mogu pridružiti klasi kolača ima najmanje, tačnije 68, dok za peciva ima najviše 83, a za pice ima 71. Obeležja koja se najviše javljaju za sve tri klase su, redom: šećer, puter, jaja, brašno i so. So je karakteristično obeležje, jer je približno podjednako zastupljeno kao sastojak koji se najviše javlja u sve tri klase. Intuitivnim razmišljanjem moguće je zaključiti da ovo obeležje neće u velikom meri biti korisno klasifikatoru, isto kao ni ostala obeležja koja se nalaze u podjednakoj zastupljenosti za sve tri klase. Takva obeležja mogu biti uzeta u razmatranje prilikom selekcije. Obeležja koja su bitnija za donošenje odluke su ona koja se pojavljuju pretežno za samo jednu klasu. Klasa pica imaju najviše takvih obeležja, dok klasa peciva ima najmanje. Odnosno ova klasa najviše deli obeležja sa ostalima i kao takva na početku ima mogućnost za teže prepoznavanje, pa samim tim i manju tačnost.



Slika 1. Histogram pojavljivanja sastojaka za klasu kolače

III. STABLO ODLUKE

Za prvu implementaciju klasifikacije korišćen je klasifikator stabla odluke.

Metodom unakrsne validacije i uz tačnost kao meru uspešnosti utvrđeni su optimalni parametri za stablo odluke. Tako je utvrđeno da je za kriterijum na osnovu kog će se vršiti odlučivanje pogodnije da bude Đini indeks, koji predstavlja meru ukupne varijanse kroz K klasa, i definiše se izrazom $G_m = \sum_{k=1}^{K} p_{mk}\,(1 - p_{mk})$, za čvor m. Takođe, pokazano je da je najpovoljnije da stablo završi sa grananjem kad dostigne dubinu stabla 15 ili ako bi novi čvor imao manje od 1% uzoraka. Najbolji rezultat postiže se korišćenjem balansiranog stabla, odnosno potrebno je postaviti težine obrnuto proporcionalno zastupljenosti klase. Pomoću balansiranog stabla se ublažava problem nejednake zastupljenosti klasa i sprečava se da klasifikator da veću šansu uzoraku da pripada, na primer, klasi kolača u odnosu na klasu pica, na osnovu toga što u skupu za obuku ima skoro pa duplo više uzoraka koji pripadaju klasi kolača.

MATRICA 1 : MATRICA KONFUZIJE ZA STABLO ODLUKE NA SKUPU ZA OBUKU

$$\begin{bmatrix} 716 & 7 & 0 \\ 91 & 521 & 7 \\ 6 & 14 & 376 \end{bmatrix}$$

Bitno je napomenuti da je matrica konfuzije data u takvom obliku da se na glavnoj dijagonali nalaze ispravno klasifikovani uzorci. Pomoću nje moguće je uočiti da je 1% uzoraka koji pripadaju klasi kolača klasifikator prepoznao kao peciva, a da ni jednom nije pomešao i klasifikovao recept za kolač kao recept za picu. Kod peciva se javlja najviše odstupanja, gde je čak 15% uzoraka pogrešno klasifikovano. Gde klasifikator više brka kolače i peciva, nego pice i peciva. Mešanje kolača i peciva se događa zbog veće unije obeležja, nego što je to između ostalih obeležja. Kod klasifikacije recepata za picu javlja se mešanje i sa receptima za kolače i sa receptima za pecivo, međutim klasifikator ne pravi toliku grešku, čak i najtačnije vrši klasifikaciju. Jedan od potencijalnih razloga je veća koncetracija obeležja koja se javljaju samo za klasu pice.

Potrebno je razmotriti rezultate klasifikatora na novom, neviđenom skupu i za to je korišćen test skup. Primetno je da je tačnost kako pojedinačnih, tako i prosečnog klasifikatora veća na skupu za obuku nego na testnom skupu, što je i očekivano jer skup za obuku daje malo optimističnije rezultate u odnosu na realne, jer su na njemu pronađeni optimalni parametri klasifikatora i na osnovu toga su ti uzorci već viđeni, a i stablo odluke je generalno podložno preobučavanju, što čini dodatni razlog za lošijim rezultatima na neviđenom skupu.

MATRICA 2 : MATRICA KONFUZIJE ZA STABLO ODLUKE NA TEST SKUPU

$$\begin{bmatrix} 74 & 5 & 1 \\ 9 & 60 & 0 \\ 3 & 3 & 38 \end{bmatrix}$$

Test skup u određenoj meri prati odstupanja skupa za obuku, odnosno procenat pogrešno klasifikovanih uzoraka je približan procentu pogrešno klasifikovanih uzoraka skupa za obuku. Jedna od razlika je da na testnom skupu klasifikator ni jednom nije pomešao i uzorku umesto pecivo dodelio da pripada klasi pice, što nije bio slučaj u skupu za obuku. Na testnom skupu, ukoliko se radi o uzorku koji pripada klasi pasta, manja je verovatnoća javljanja lažnih pozitivnih uzoraka nego što je to kod skupa za obuku.

TABELA 1 : TAČNOST KLASIFIKATORA STABLA ODLUKE NA TRENING I TEST SKUPU

Naziv klase	Skup za obuku	Skup za test
Kolači	94.02%	90.67%
Peciva	93.15%	91.19%
Pice	98.45%	96.37%
Prosečna tačnost klasifikatora	95.20%	92.75%

Stablo odluke pogodno je za jednostavnu vizualizaciju (dato u kodu) i interpretaciju. Relativno velika dimenzionalnost, što se može smatrati da se javlja u ovom skupu ne predstavlja problem za stablo odluke. Još jedna prednost ovog klasifikatora je pogodnost utvrđivanja najvažnijih obeležja, što može biti korisno i pri selekciji dimenzionalnosti. Najvažnija obeležja su ona prema kojima se vrše prve podele, jer ona najviše informacija klasifikatoru otkrivaju i tako redom, kako su obeležja bliža listovima stabla tako manje utiču na tačnost klasifikatora. Na osnovu vizualnog prikaza stabla odluke uočljivo je da se prva podela vrši po tome da li recept sarži sir ili ne i na taj način se grana stablo koje u jednom čvoru ima značajno manji procenat uzoraka kojima je pridružena klasna labela pica, u odnosu na drugi, desni, čvor sa 70% uzoraka sa ovom klasnom labelom. Pri vrhu stabla, sa dubinom 2 na osnovu dva pitanja moguće je potencijalno odgovoriti da li se radi o receptu za pecivo.

Problem preobučavanja je jedna od negativnih tendencija stabla odluke, to je i razlog zašto je veće odstupanje uspešnosti klasifikatora na skupu za obuku u odnosno na skup za testiranje. Takođe, korišćenje jednog stabla odluke ne daje dovoljno dobre rezultate klasifikacije, kao što bi to dali neki drugi klasifikatori.

Metoda slučajne šume u određenoj meri rešava pomenute probleme obučavajući mnoštvo stabala odluke i donošenjem krajnje odluke glasanjem. Za razliku od klasičnog klasifikatora pomoću jednog stabla odluke kod kojeg je interpretacija relativno jednostavna, kod metode slučajne šume je dosta teža, međutim za dovoljno velik skup podatak metoda slučajne šume se pokazala kao bolja opcija u odnosu na jedno stablo odluke i zbog toga je i u ovom radu implementiran. Mera uspešnosti klasifikatora ukazuje na to da je bolji izbor metoda slučajne šume. Bitno je uočiti da je kod metode slučajne šume mnogo manja razlika procenta uspešnosti između skupa za obuku i za testiranje, nego što je to kod klasifikacije pomoću jednog stabla odluke, što je posledica toga da ova metoda nije podložna preobučavanju u toj meri u kojoj je klasifikacija jednim stablom odluke podložna. Na osnovu upoređene uspešnosti na testnom skupu bolje je koristiti klasifikaciju metodom slučajne šume.

MATRICA 3 : MATRICA KONFUZIJE ZA METODU SLUČAJNE ŠUME NA SKUPU ZA OBUKU

$$\begin{bmatrix} 717 & 6 & 0 \\ 103 & 508 & 8 \\ 5 & 11 & 380 \end{bmatrix}$$

MATRICA 4 : MATRICA KONFUZIJE ZA METODU SLUČAJNE ŠUME NA TEST SKUPU

$$\begin{bmatrix} 79 & 1 & 0 \\ 7 & 62 & 0 \\ 3 & 0 & 41 \end{bmatrix}$$

TABELA 2 : TAČNOST KLASIFIKATORA METODE SLUČAJNE ŠUME NA TRENING I TEST SKUPU

Naziv klase	Skup za obuku	Skup za test
Kolači	93.44%	94.30%
Peciva	92.63%	95.85%
Pice	98.61%	98.44%
Prosečna tačnost klasifikatora	94.89%	96.20%

IV. LOGISTIČKA REGRESIJA

Pored stabla odluke klasifikator koji je korišćen je multinomijalna logistička, takođe je tačnost korišćena kao mera uspešnosti radi ispravnog upoređivanja uspešnosti ova dva klasifikatora.

Utvrđeno je da je potrebno da optimalni klasifikator dodeljuje težine klasa obrnuto proporcijalno njihovom učestalosti.

MATRICA 5: MATRICA KONFUZIJE ZA LOGISTIČKU REGRESIJU NA SKUPU ZA OBUKU

$$\begin{bmatrix} 684 & 39 & 0 \\ 60 & 550 & 9 \\ 5 & 4 & 387 \end{bmatrix}$$

MATRICA 6: MATRICA KONFUZIJE ZA LOGISTIČKU REGRESIJU NA TEST SKUPU

$$\begin{bmatrix} 73 & 7 & 0 \\ 3 & 64 & 2 \\ 2 & 0 & 42 \end{bmatrix}$$

Kao i kod klasifikatora stablom odluke klasifikator logističke regresije ne svrstava uzorke koji pripadaju klasi kolača u klasu pize, ni u skupu za obučavanje ni u testnom skupu. Malo lošije, 3%, uzorak koji pripada klasi peciva razvrstava u tu klasu, veći je procenat greške da će biti razvrstan u klasu pica nego što je to bilo na skupu za obuku. Sa druge strane, uzorci koji pripadaju klasi pica su dodeljeni klasi pica u većem procentu na test nego na obučavajućem skupu. Potencijalni razlog za to je veća učestalost uzoraka u test skupu koji imaju izrazito karakteristična obeležja za klasu pica.

Razlika između prosečne uspešnosti klasifikatora na skupu za obuku i test skupu je relativno mala, oko 1%.

Iako je logističa regresija inicijalno za binarne probleme, na lak način se može proširiti i na multinomijalnu regresiju, kao što je predstavljeno kroz ovaj rad. Pored ove prednosti i u poređenju sa klasifikatorom stabla odluke logistička regresija nije toliko podložna preobučavanju, pa razlika između rezultata na skupu za obuku i testnom skupu nije toliko značajna. Jedna od potencijalnih metoda za poboljšanje modela logističke regresije je korišćenje regularizacionih metoda i na taj način se postigao kompromis između pristrasnosti i varijanse.

Jedna od mana logističke regresije je što joj veći problem pravi visoka dimenzionalnost nego što to predstavlja problem za klasifikator stabla odluke.

TABELA 3 : TAČNOST KLASIFIKATORA LOGISTIČKE REGRESIJE NA TRENING I TEST SKUPU

Naziv klase	Skup za obuku	Skup za test
Kolači	94.01%	93.78%
Peciva	93.55%	93.78%
Pice	98.96%	97.93%
Prosečna tačnost klasifikatora	95.51%	95.16%

V. POREĐENJE KLASIFIKATORA

Kako metoda slučajne šume daje bolje rezultate od klasifikacije stablom odluke u nastavku će biti upoređivani rezultati klasifikatora metodom slučajne šume i logističke regresije.

Klasifikacija logističkom regresijom daje lošije rezultate na oba skupa od klasifikacije pomoću metode slučajne šume, kada je tačnost mera uspešnosti, a napomenuto je u prethodnom poglavlju da je tačnost mera koja će najviše uticati na odluku koji klasifikator daje bolje rezultate.

Preciznost, osetljivost i specifičnost za svaku klasu takođe imaju bolje rezultate kada se koristi klasifikator metode slučajne šume, upravo iz razloga što je logistička regresija sklonija preobučavanju, pa na neviđenom skupu daje lošije rezultate.

Potrebno je zapaziti da klasifikator metode slučajne šume za klasu pice dostiže preciznost i specifičnost od 100%. Odnosno, ni jedan uzorak nije svrstan u klasu pica, a da njoj ne pripada.

Uzorci koji pripadaju klasi kolača su u najvećoj meri klasifikovani da pripadaju toj klasi, za metodu slučajne šume, dok su za logističku regresiju to uzorci koji pripadaju klasi pica.

Kako su preciznost i osetljivost veće za metodu slučajne šume, tako je i njihova harmonijska sredina, F-mera, veća i na makro i na mikro nivou.

TABELA 4 : PRECIZNOST

Naziv klase	Logistička regresija	Metoda slučajne šume
Kolači	93.58%	98.76%
Peciva	90.14%	98.41%
Pice	95.45%	100.00%
Mikro	92.75%	94.30%
Makro	93.06%	95.73%

TABELA 5 : OSETLJIVOST

Naziv klase	Logistička regresija	Metoda slučajne šume
Kolači	91.25%	98.75%
Peciva	92.75%	89.85%
Pice	95.45%	93.18%

TABELA 6: SPECIFIČNOST

Naziv klase	Logistička regresija	Metoda slučajne šume
Kolači	93.91%	99.04%
Peciva	94.35%	99.19%
Pice	98.66%	100.00%

TABELA 7: TAČNOST

Naziv klase	Logistička regresija	Metoda slučajne šume
Kolači	93.78%	99.04%
Peciva	93.78%	99.19%
Pice	97.93%	98.44%
Prosečna tačnost klasifikatora	95.16%	96.2%

TABELA 8: F-MERA

Naziv klase	Logistička regresija	Metoda slučajne šume
Mikro	77.15%	78.62%
Makro	93.11%	94.82%

VI.ZAKLJUČAK

Kada je potrebno porediti klasifikator metode slučajne šume sa klasifikatorom logističke regresije, a da je mera uspešnosti tačnost, vidno bolje rezultate daje klasifikator metodom slučajne šume, čak i ne samo za tačnost kao meru uspešnosti, nego i za osetljivost, preciznost i specifičnost. Međutim kada je potrebno uporediti klasifikator pomoću jednog stabla odluke sa logističkom regresijom, veću tačnost bi imala metoda logističke regresije, pogotovo na neviđenom skupu. Stablo odluke sa svojom dobrom vizuelizacijom i lakom interpretacijom poželjno je koristiti eventualno na početku upoznavanja sa problemom, radi lakšeg sagledavanja uticaja određenih obeležja na donošenje krajnjih odluka, a nakom toga bi bolje rezultate dala neka druga metoda klasifikacije.