Primjena modificiranog k-adskog Jaccardovog koeficijenta sličnosti za usporedbu dvaju skupova binarnih klasifikatora

Petričević, Rafaela Brigita

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

RAFAELA BRIGITA PETRIČEVIĆ

Primjena modificiranog k-adskog Jaccardovog koeficijenta sličnosti za usporedbu dvaju skupova binarnih klasifikatora

DIPLOMSKI RAD

Split, rujan 2023.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

Primjena modificiranog k-adskog Jaccardovog koeficijenta sličnosti za usporedbu dvaju skupova binarnih klasifikatora

DIPLOMSKI RAD

Neposredna voditeljica:
dr. sc. Ana Perišić

Studentica:
Rafaela Brigita Petričević

Mentorica:
prof. dr. sc. Borka Jadrijević

Split, rujan 2023.

TEMELJNA DOKUMENTACIJSKA KARTICA

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU
ODJEL ZA MATEMATIKU

DIPLOMSKI RAD

Primjena modificiranog k-adskog Jaccardovog koeficijenta sličnosti za usporedbu dvaju skupova binarnih klasifikatora

Rafaela Brigita Petričević

Sažetak:

Klasifikacija je jedna od najvažnijih tehnika analize podataka, a koristi se za razvrstavanje podataka u unaprijed definirane kategorije na temelju njihovih svojstava. U ovom radu ocjenjujemo sličnost u konsenzusu o prisutnosti unutar jednog i između dvaju skupova klasifikatora. U prvom dijelu rada uvodimo potrebnu teorijsku podlogu te pobliže upoznajemo dva odabrana klasifikacijska algoritma, logističku regresiju i slučajne šume. Nadalje, predstavljamo koeficijente sličnosti kao mjere povezanosti objekata ili skupina objekata, s naglaskom na Jaccardov koeficijent sličnosti i njegove modifikacije. U drugom dijelu rada obrađujemo dva realna skupa podataka. Gradimo modele logističke regresije i primjenjujemo dvije različite implementacije algoritma slučajnih šuma. Nakon izgradnje i ocjene modela, računamo sličnost u konsenzusu o prisutnosti unutar jednog i između dvaju skupova klasifikatora i dajemo interpretaciju dobivenih rezultata.

Ključne riječi:

binarna klasifikacija, sličnost skupova binarnih klasifikatora, Jaccardov koeficijent sličnosti

TEMELJNA DOKUMENTACIJSKA KARTICA

Podatci o radu:

44 stranice, 1 slika, 12 tablica, 12 literaturnih navoda, izvornik na hrvatskom jeziku

Mentorica: *prof. dr. sc. Borka Jadrijević*

Neposredna voditeljica: *dr. sc. Ana Perišić*

Član povjerenstva: *Marcela Mandarić, mag. math.*

Povjerenstvo za diplomski rad je prihvatilo ovaj rad *4. rujna 2023.*

TEMELJNA DOKUMENTACIJSKA KARTICA

FACULTY OF SCIENCE, UNIVERSITY OF SPLIT
DEPARTMENT OF MATHEMATICS

MASTER'S THESIS

An application of the modified k-adic Jaccard similarity coefficient for comparing two sets of binary classifiers

Rafaela Brigita Petričević

Abstract:

Classification is one of the most important data analysis techniques, used to categorize data into predefined classes based on their attributes. In this paper, we assess similarity in consensus agreement on presence within and between two sets of classifiers. In the first part of the paper, we introduce the necessary theoretical background and discuss in detail the two selected classification algorithms, logistic regression and random forests. Furthermore, we introduce similarity coefficients as measures of association among objects or sets of objects, with an emphasis on the Jaccard similarity coefficient and its modifications. In the second part of the paper, we analyze two real-world datasets. We build logistic regression models and apply two different implementations of the random forest algorithm. After constructing and evaluating the models, we calculate similarity in consensus agreement on presence within and between two sets of classifiers, providing an interpretation of the obtained results.

Key words:

binary classification, similarity of sets of binary classifiers, Jaccard similarity coefficient

TEMELJNA DOKUMENTACIJSKA KARTICA

Specifications:

44 pages, 1 image, 12 tables, 12 references, original in Croatian

Mentor: *professor Borka Jadrijević*

Immediate mentor: *Ana Perišić, PhD*

Committee: *Marcela Mandarić, mag. math.*

This thesis was approved by a Thesis committee on *September 4th, 2023.*

Uvod

Živimo u svijetu u kojem se svake sekunde generira ogromna količina podataka. Bilo da se radi o znanstvenim ili svakodnevnim podacima, njihov utjecaj na naše odluke i saznanja je neupitan. Međutim, brzo rastuća količina i različiti oblici podataka nam otežavaju izdvajanje korisnih informacija, stoga su nam potrebni moćni alati za obradu i analiziranje podataka. Klasifikacija je jedna od najvažnijih tehnika analize podataka, a nalazi primjenu u raznim područjima, poput medicine, ekonomije, obrade jezika i ekologije. Zadaća klasifikacije je razvrstavanje podataka u unaprijed definirane kategorije ili klase na temelju njihovih svojstava. Na primjer, uz primjenu klasifikacijskih algoritama moguće je prepoznati bolest na temelju simptoma pacijenta ili razvrstati književna djela na temelju žanra. U ovom radu promatramo odabrane klasifikacijske algoritme i uspoređujemo njihovu sličnost na dvama skupovima podataka iz stvarnog svijeta.

Prvo poglavlje započinjemo uvođenjem osnovnih ideja klasifikacije i detaljnijim opisom klasifikacije kao procesa u dva koraka. Dalje navodimo najčešće korištene mjere za ocjenjivanje prediktivne moći klasifikatora i dajemo smjernice za njihovo korištenje. Nakon upoznavanja s osnovnim pojmovima prelazimo na odabrane klasifikacijske algoritme, logističku regresiju i slučajne šume.

Logistička regresija je jedna od najkorištenijih tehnika za opisivanje veze

između nezavisnih i zavisne varijable i to u slučaju kad je zavisna varijabla kategorijska. Predstavljamo univarijatni logistički model i njegove dvije generalizacije, multivarijatni i multinomni logistički model. Također, opisujemo metode za procjenu parametara i izgradnju logističkih modela. Konačno, prikazujemo postupak ocjenjivanja performansi logističkih modela na problemima klasifikacije.

Drugi klasifikacijski algoritam kojeg obrađujemo je algoritam slučajnih šuma. Zbog svoje robusnosti i jednostavnosti korištenja stekao je veliku popularnost u strojnom učenju i rudarenju podataka. Dajemo preporuke za izbor hiperparametara i komentiramo metodu za procjenu pogreške predviđanja. Opisujemo dva programska paketa za rad sa slučajnim šumama koja koristimo za izgradnju modela u posljednjem poglavlju.

U drugom poglavlju se bavimo koeficijentima sličnosti. Koeficijentima sličnosti se nastoji kvantificirati sličnost ili povezanost među objektima ili skupinama objekata. Predstavljamo generalizaciju poznatog Jaccardovog koeficijenta sličnosti na slučaj k objekata, a zatim i modifikaciju navedenog koja se koristi za usporedbu sličnosti dvaju skupova objekata. Također, opisujemo korištenje takve modifikacije za usporedbu sličnosti u konsenzusu dvaju skupova klasifikatora.

U posljednjem poglavlju analiziramo dva medicinska skupa podataka. Gradimo modele logističke regresije i slučajnih šuma s različitim prediktorima odnosno hiperparametrima. Primjenjujemo dvije različite implementacije algoritma slučajnih šuma, *randomForest* i *cforest*. Nakon izgradnje, ocjenjujemo prediktivne sposobnosti modela na skupu za testiranje pomoću ranije navedenih mjera. Na kraju računamo sličnost u konsenzusu unutar jednog i između dvaju skupova klasifikatora i dajemo interpretaciju dobivenih rezultata.

Sadržaj

Poglavlje 1

Klasifikacija

Klasifikacija je oblik analize podataka u kojem se izdvajaju modeli za opisivanje važnih klasa podataka. Takvi modeli, zvani klasifikatorima, predviđaju kategorijske (diskretne, neuređene) oznake klasa. Na primjer, možemo izgraditi klasifikacijski model koji kategorizira zahtjeve za bankovni kredit kao sigurne ili rizične. Takva analiza nam može pomoći u detaljnijem razumijevanju podataka. Istraživači koji se bave strojnim učenjem, prepoznavanjem uzoraka i statistikom predložili su mnoge klasifikacijske metode. Detekcija prevare, ciljani marketing, predviđanje performansi, proizvodnja i medicinska dijagnostika samo su neka od područja primjene klasifikacije.
Počinjemo uvođenjem osnovnih ideja klasifikacije nakon čega razmatramo ocjenjivanje i usporedbu različitih klasifikatora.

1.1 Osnovni koncepti klasifikacije

U ovom potpoglavlju uvodimo koncept klasifikacije i opisujemo općeniti pristup klasifikaciji kao procesu u dva koraka. U prvom koraku gradimo klasifikacijski model na temelju podataka za treniranje, dok u drugom koraku

određujemo je li točnost modela prihvatljiva i, ako je, koristimo isti model za klasifikaciju novih podataka.

U praksi često nailazimo na primjenu klasifikacijskih metoda. Zajmodavac treba analizu podataka da bi odredio koji podnositelj zahtjeva za zajam je "siguran", a koji "rizičan" za banku u kojoj radi. Marketinški menadžer u trgovini elektroničkom opremom treba analizu podataka da bi procijenio hoće li kupac određenog profila kupiti novo računalo. Medicinski istraživač analizira podatke o raku dojke da odredi koju bi terapiju, od tri različite, trebala primiti određena pacijentica. U svakom od navedenih primjera zadatak analize podataka je klasifikacija, gdje je model (klasifikator) izgrađen da predviđa (kategorijske) oznake klasa, poput "siguran" ili "rizičan" za podatke o kandidatima za zajam; "da" ili "ne" za marketinške podatke; ili "terapija A", "terapija B" ili "terapija C" za medicinske podatke. Ove kategorije mogu biti predstavljene diskretnim vrijednostima, čiji poredak nije važan. Na primjer, vrijednosti 1, 2 i 3 se mogu koristiti za reprezentaciju terapija A, B i C, gdje poredak među ovim režimima liječenja nije impliciran.

Klasifikacija podataka je dvokoračni proces, točnije, sastoji se od koraka učenja (u kojem se izgrađuje klasifikacijski model) i koraka klasifikacije (u kojem se taj model koristi za predviđanje oznaka klasa danih podataka). Slika 1.1 prikazuje postupak klasifikacije na primjeru zahtjeva za zajam (podaci su pojednostavljeni za potrebe ilustracije. U stvarnosti je očekivano da će se razmatrati puno više atributa).

U prvom koraku se izgrađuje klasifikator koji opisuje unaprijed određen skup klasa podataka ili koncepata. Ovo je korak učenja (ili treniranja), u kojem klasifikacijski algoritam gradi klasifikator analiziranjem ili "učenjem od" skupa za treniranje, koji se sastoji od podataka iz baze i njima pridruženih oznaka klasa. Podatak, u oznaci X, je reprezentiran n-dimenzionalnim vekto-

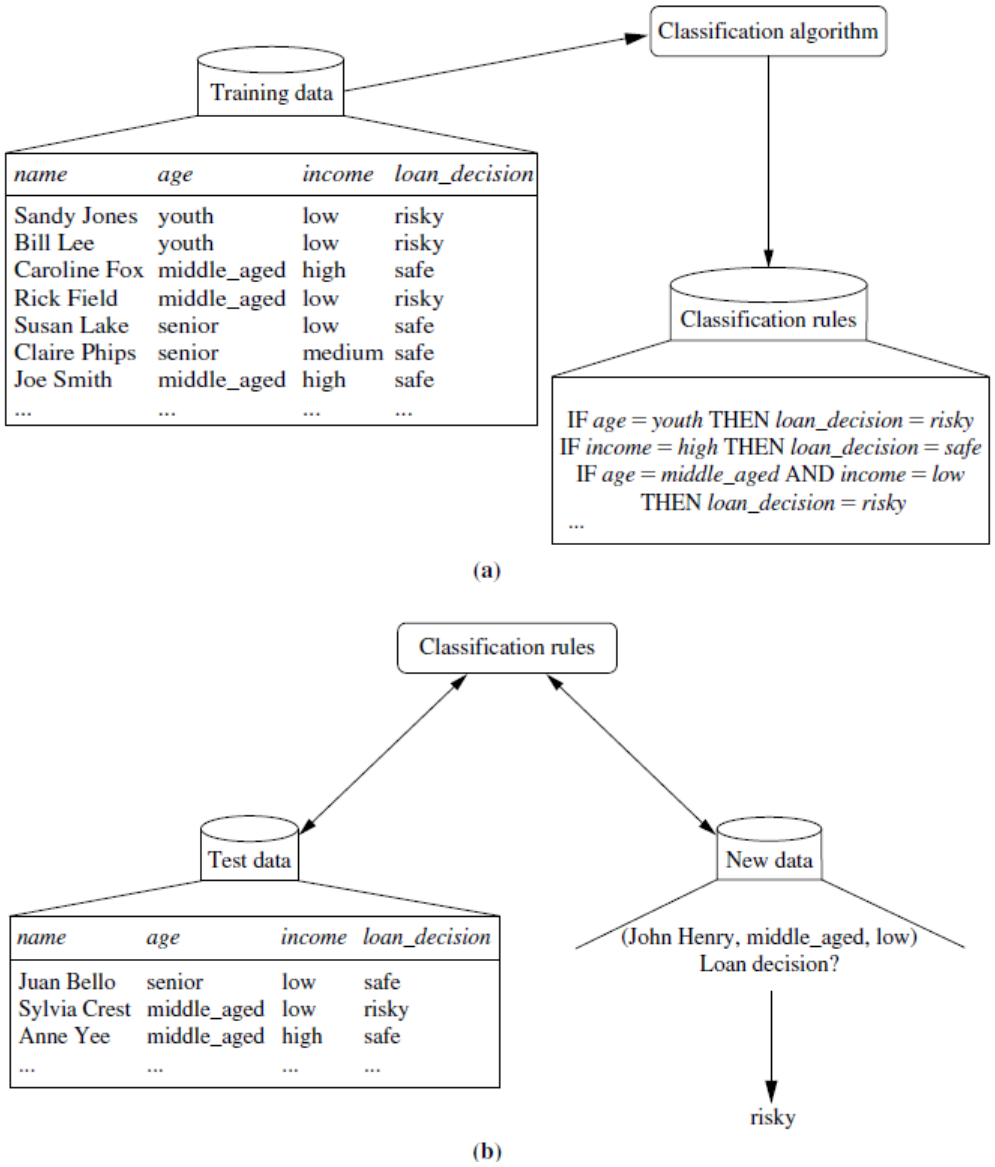
rom atributa $X = (x_1, \ldots, x_n)$ prikazujući n mjerenja načinjenih na podatku za n atributa baze podataka, redom $A_1, \ldots, A_n$.

Pretpostavlja se da svaki podatak pripada predefiniranoj klasi koja je određena posebnim atributom baze podataka, takozvanim atributom oznake klase. Atribut oznake klase ima diskretnu vrijednost i neuređen je. Kategorijski je (ili nominalan) utoliko što svaka vrijednost služi kao kategorija ili klasa. Pojedinačni podaci koji čine skup za treniranje nazivaju se podacima za treniranje i slučajno se uzorkuju iz baze podataka koja se analizira. U kontekstu klasifikacije nazivaju se još i uzorcima, primjercima, instancama, podatkovnim točkama ili objektima.

Ovaj korak je poznat i pod nazivom nadzirano učenje jer je za svaki podatak skupa za treniranje dana oznaka pripadajuće klase (odnosno, učenje klasifikatora je "nadzirano" u smislu da je poznato kojoj klasi pripada koji podatak). U suprotnosti je s nenadziranim učenjem (ili klasteriranjem) kod kojeg oznaka klase niti jednog podatka nije poznata te broj ili skup klasa ne mora biti unaprijed poznat.

„Što je s točnošću klasifikacije?“ U drugom koraku (slika 1.1(b)) model se koristi za klasifikaciju. Najprije se procjenjuje prediktivna točnost klasifikatora. Ako bismo koristili skup za treniranje za mjerenje točnosti klasifikatora, procjena bi vjerojatno bila optimistična jer su klasifikatori skloni pretreniranju ili prenaučenosti (*"overfit"*), to jest tijekom učenja mogu uzeti u obzir određene nepravilnosti skupa za treniranje koje općenito nisu prisutne u cijelom skupu podataka. Dakle, koristi se skup za testiranje sačinjen od podataka za testiranje i njima odgovarajućih oznaka klasa. Podaci za testiranje ne ovise o podacima za treniranje, što znači da nisu korišteni za izgradnju klasifikatora.

Točnost klasifikatora na danom skupu za testiranje je postotak podataka



Slika 1.1: Proces klasifikacije podataka: (a) *Učenje: Klasifikacijski algoritam analizira podatke skupa za treniranje. U ovom primjeru, atribut oznake klase je loan_decision i naučeni model je predstavljen u obliku klasifikacijskih pravila.* (b) *Klasifikacija: Radi procjene točnosti klasifikacijskih pravila koriste se podaci skupa za testiranje. Ako se točnost smatra prihvatljivom, pravila se mogu primijeniti na klasifikaciju novih podataka.*

iz skupa za testiranje koje je taj klasifikator točno klasificirao. Oznaka klase pridružena svakom podatku za testiranje uspoređuje se s oznakom klase koju je naučeni klasifikator predvidio za taj podatak. Ako se točnost klasifikatora smatra prihvatljivom, može se dalje koristiti za klasifikaciju budućih podataka čije oznake klasa nisu poznate. Na primjer, klasifikacijska pravila na slici 1.1(a) naučena analizom podataka o prethodnim zahtjevima za zajam se mogu koristiti za odobrenje ili odbijanje novih zahtjeva za zajam.

1.2 Evaluacija modela

Nakon izgradnje klasifikacijskog modela postavljaju se mnoga pitanja. Na primjer, pretpostavimo da smo za izgradnju klasifikatora koji predviđa ponašanje kupaca koristili podatke o prethodnim kupovinama. Htjeli bismo procijeniti koliko točno klasifikator može predvidjeti ponašanje budućih kupaca, to jest podatke o budućim kupcima na kojima klasifikator nije bio treniran. Možda smo čak, pomoću različitih metoda, izgradili više klasifikatora i želimo usporediti njihovu točnost. No što je zapravo točnost? Kako je možemo procijeniti? Jesu li neke mjere točnosti klasifikatora primjerenije od ostalih? Odgovore na ova pitanja dajemo u sljedećem potpoglavlju.

1.2.1 Mjere za ocjenu klasifikatora

U ovom potpoglavlju su predstavljene mjere kojima se procjenjuje koliko je dobar, odnosno koliko "točno" klasifikator predviđa oznake klasa podataka. Promatrat ćemo slučaj u kojem su klase prilično ravnomjerno raspodijeljene i slučaj u kojem su klase neuravnotežene (na primjer, gdje je klasa od interesa malobrojna). Mjere za ocjenu klasifikatora predstavljene u ovom poglavlju su sažete u tablici 1.2. Među njima se nalaze točnost (također poznata kao stopa prepoznavanja), osjetljivost (ili opoziv), specifičnost, preciznost, F_1 i F_β. Primijetimo da se, iako je točnost posebna mjera, riječ "točnost" također koristi kao općenit pojam koji se odnosi na prediktivne sposobnosti klasifikatora.

Korištenje skupa za treniranje za izgradnju klasifikatora i procjenu točnosti dobivenog, naučenog modela može dati zavaravajuće, preoptimistične procjene zbog pretjerane specijalizacije algoritma učenja na podacima. Točnost klasifikatora je bolje mjeriti na skupu za testiranje, koji se sastoji od poda-

taka koji nisu korišteni za treniranje modela.

Prije rasprave o različitim mjerama napomenimo da ćemo u ostatku poglavlja govoriti o pozitivnim podacima (podaci koji pripadaju klasi od interesa) i negativnim podacima (svi ostali podaci). Ako su dane dvije klase, na primjer, pozitivni podaci mogu biti oni za koje je *buys_computer* = *yes*, a negativni podaci oni za koje je *buys_computer* = *no*.

Pretpostavimo da koristimo klasifikator na skupu za testiranje s označenim podacima i neka je P broj pozitivnih, a N broj negativnih podataka. Za svaki podatak uspoređujemo predviđenu s već poznatom oznakom klase.

Navedimo još četiri pojma koja trebamo poznavati i koji su temelj za računanje mnogih mjera za ocjenjivanje klasifikatora:

- Stvarno pozitivni podaci (u oznaci TP, *"true positives"*): Podaci koje je klasifikator ispravno prepoznao kao pozitivne. Neka je TP broj stvarno pozitivnih podataka.

- Stvarno negativni podaci (u oznaci TN, *"true negatives"*): Podaci koje je klasifikator ispravno prepoznao kao negativne. Neka je TN broj stvarno negativnih podataka.

- Lažno pozitivni podaci (u oznaci FP, *"false positives"*): Podaci koje je klasifikator pogrešno prepoznao kao pozitivne. Neka je FP broj lažno pozitivnih podataka.

- Lažno negativni podaci (u oznaci FN, *"false negatives"*): Podaci koje je klasifikator pogrešno prepoznao kao negativne. Neka je FN broj lažno negativnih podataka.

Navedeni pojmovi su sažeti u donjoj tablici.

	Predviđena klasa		
Stvarna klasa	Da	Ne	Ukupno
Da	TP	FN	P
Ne	FP	TN	N
Ukupno	P'	N'	P + N

Tablica 1.1: Matrica zabune za slučaj dviju klasa.

Matrica zabune (matrica konfuzije, *"confusion matrix"*) daje dobar uvid u sposobnost klasifikatora da ispravno prepoznaje podatke različitih klasa. TP i TN ukazuju na ispravan rad klasifikatora, a FP i FN nam govore kad klasifikator griješi, odnosno krivo označava podatke. Ako je zadano m klasa (gdje je $m \geq 2$), matrica zabune je veličine barem $m \times m$. Element $C_{i,j}$ i-tog retka i j-tog stupca matrice zabune predstavlja broj podataka klase i koje je klasifikator označio kao podatke klase j, za $i, j = 1, \ldots, m$. Da bi klasifikator imao dobru točnost, većina podataka bi se trebala nalaziti na dijagonali matrice zabune, od elementa $C_{1,1}$ do $C_{m,m}$, a brojevi van dijagonale bi trebali biti nula ili blizu nuli. Drugim riječima, FP i FN bi trebali biti blizu nuli.

Matrica zabune može sadržavati dodatne retke ili stupce koji sadrže ukupne iznose. Na primjer, u gornjoj matrici zabune (tablica 1.1) su pokazani P (broj pozitivnih podataka) i N (broj negativnih podataka). Dodatno, P' je broj podataka koji su označeni kao pozitivni ($TP + FP$) i N' je broj podataka koji su označeni kao negativni ($TN + FN$). Ukupan broj podataka je $TP + TN + FP + TN$, ili $P + N$, ili $P' + N'$. Uočimo da se, iako je na slici dana matrica za problem binarne klasifikacije, matrice zabune lako i na sličan način mogu konstruirati i za probleme s više klasa.

Promotrimo sada mjere ocjene, počevši s točnošću. Točnost klasifikatora

(*"accuracy"*) je omjer broja ispravno klasificiranih podataka i broja svih podataka, to jest definiramo

$$točnost = \frac{TP + TN}{P + N}.$$

Točnost je najbolje upotrijebiti kad je raspodjela klasa poprilično ravnomjerna (balansirana).

Promotrimo problem nebalansiranosti klasa gdje je klasa od interesa malobrojna. Drugim riječima, distribucija skupa podataka ukazuje na značajnu većinu negativne klase i manjinu pozitivne klase. Na primjer, kod problema detekcije prevara klasa koja nas zanima (pozitivna klasa) je *"fraud"* i ona se pojavljuje puno rjeđe nego negativna, *"nonfraudulant"* klasa. U medicinskim podacima također može postojati malobrojna klasa, na primjer klasa koja označava prisutnost bolesti. Pretpostavimo da smo istrenirali klasifikator da klasificira medicinske podatke, gdje je atribut oznake klase *"cancer"* s mogućim vrijednostima "da" ili "ne". Točnost od npr. 97 % može zvučati jako dobro, ali što ako 3 % podataka za treniranje zapravo predstavlja rak? Očito, stopa točnosti od 97 % ne mora biti prihvatljiva – klasifikator može točno klasificirati samo podatke koji predstavljaju odsutnost raka, a krivo razvrstati sve podatke koji predstavljaju prisutnost raka. Zaključujemo da, umjesto točnosti, trebamo druge mjere koje nam daju uvid u to koliko dobro klasifikator raspoznaje pozitivne (*cancer = yes*) i negativne podatke (*cancer = no*).

U ovu svrhu možemo koristiti mjere osjetljivosti i specifičnosti. Osjetljivost se često naziva i stopom (prepoznavanja) stvarne pozitivnosti (*"true positive (recognition) rate"*) (to jest udio pozitivnih podataka koji su ispravno prepoznati), dok je specifičnost stopa stvarne negativnosti (*"true negative (recognition) rate"*) (to jest udio negativnih podataka koji su ispravno pre-

poznati). Navedene mjere su definirane kao

$$osjetljivost = \frac{TP}{P}$$

$$specifičnost = \frac{TN}{N}.$$

Lako se pokaže da je točnost funkcija osjetljivosti i specifičnosti:

$$točnost = osjetljivost\frac{P}{P+N} + specifičnost\frac{N}{P+N}.$$

Mjere preciznosti (*"precision"*) i opoziva (*"recall"*) također imaju široku primjenu u klasifikaciji. Na preciznost se može gledati kao na mjeru egzaktnosti (to jest koliki postotak podataka označenih kao pozitivno zapravo jest pozitivno), dok je opoziv mjera potpunosti (koliki je postotak pozitivnih podataka označen kao pozitivno). Primijetimo da je opoziv jednak osjetljivosti. Ove mjere se mogu izračunati kao

$$preciznost = \frac{TP}{TP+FP}$$

$$opoziv = \frac{TP}{TP+FN} = \frac{TP}{P}.$$

Savršena preciznost 1.0 za klasu C znači da svaki podatak koji je klasifikator dodijelio klasi C uistinu pripada klasi C. Međutim, taj rezultat ne govori ništa o broju podataka klase C koje je klasifikator pogrešno razvrstao. Savršen opoziv 1.0 za C znači da je svaki podatak iz klase C označen kao takav, ali ne možemo zaključiti koliko je preostalih podataka pogrešno pridijeljeno klasi C. Nadalje, postoji inverzna veza između preciznosti i opoziva, stoga je moguće povećati jednu vrijednost nauštrb druge. Na primjer, medicinski klasifikator može ostvariti visoku preciznost točno označavajući sve podatke s određenim svojstvom kao slučajeve raka, ali imati nizak opoziv ako pogrešno razvrstava mnoge druge slučajeve raka. Preciznost i opoziv se obično koriste zajedno, i to tako da se za fiksnu vrijednost opoziva uspoređuju

Mjera	**Formula**
Točnost	$\frac{TP+TN}{P+N}$
Osjetljivost (opoziv)	$\frac{TP}{P}$
Specifičnost	$\frac{TN}{N}$
Preciznost	$\frac{TP}{TP+FP}$
F (F_1)	$\frac{2\times preciznost\times opoziv}{preciznost+opoziv}$
$F_\beta, \beta \in \mathbb{R}_0^+$	$\frac{(1+\beta^2)\times preciznost\times opoziv}{\beta^2\times preciznost+opoziv}$

Tablica 1.2: Mjere za ocjenu klasifikatora.

različiti iznosi preciznosti ili obrnuto. Na primjer, možemo uspoređivati vrijednosti preciznosti za fiksnu vrijednost opoziva 0.75.

Drugi način korištenja preciznosti i opoziva je njihovo kombiniranje u jednu mjeru – F mjeru (*"F_1-score", "F-score"*) ili F_β mjeru. Ove mjere su definirane kao

$$F = \frac{2 \times preciznost \times opoziv}{preciznost + opoziv}$$

$$F_\beta = \frac{(1+\beta^2) \times preciznost \times opoziv}{\beta^2 \times preciznost + opoziv},$$

gdje je β nenegativan realan broj. F mjera je harmonijska sredina preciznosti i opoziva koja objema daje jednaku težinu. F_β je ponderirana mjera preciznosti i opoziva, točnije, daje β puta veću ($\beta > 1$) ili manju ($\beta < 1$) težinu opozivu u odnosu na preciznost. Često korištene F_β mjere su F_2 (koja opozivu daje dvaput veću težinu) i $F_{0.5}$ (koja daje dvaput veću težinu preciznosti).

Uz prethodno navedene mjere, klasifikatori se mogu uspoređivati i na temelju sljedećih karakteristika:

- Brzina - Odnosi se na troškove izračuna potrebnih za generiranje i korištenje danog klasifikatora

- Robusnost - Sposobnost klasifikatora da napravi točna predviđanja ako su dani podaci sa šumom ili nedostajućim vrijednostima. Robusnost se obično procjenjuje nizom sintetičkih skupova podataka rastućeg stupnja šuma i nedostajućih vrijednosti.

- Skalabilnost - Mogućnost efikasne konstrukcije klasifikatora ako je dana velika količina podataka. Skalabilnost se obično procjenjuje nizom skupova podataka rastuće veličine.

- Interpretabilnost - Razina razumijevanja i uvida u klasifikator. Ovo svojstvo je subjektivno, stoga ga je teže ocijeniti. Stabla odluke i klasifikacijska pravila mogu biti jednostavna za tumačenje, ali njihova razumljivost može opadati povećanjem kompleksnosti.

Ukratko, predstavili smo nekoliko mjera za ocjenjivanje klasifikatora. Mjeru točnosti je najbolje koristiti kad su klase podataka prilično ravnomjerno raspodijeljene. Druge mjere, poput osjetljivosti (ili opoziva), specifičnosti, preciznosti, F i F_β, su primjerenije za probleme s nebalansiranim klasama gdje je klasa od interesa malobrojna.

1.3 Logistička regresija

Regresijske metode su sastavni dio svake analize podataka koja se bavi opisivanjem veze između zavisne varijable (varijable odziva) i jedne ili više nezavisnih (eksplanatornih) varijabli. Čest slučaj je da je zavisna varijabla kategorijska i poprima dvije ili više mogućih vrijednosti. Upravo u takvim situacijama se kao standardna metoda analize nameće logistička regresija.

1.3.1 Univarijatni logistički model

Zadaća svakog regresijskog problema je predviđanje srednje vrijednosti varijable odziva uz danu vrijednost nezavisne varijable. Tu vrijednost označavamo s $\mathbb{E}(Y|x)$, pri čemu Y označava varijablu odziva, a x predstavlja vrijednost nezavisne varijable. U slučaju binarne (dihotomne) varijable odziva treba vrijediti

$$0 \leq \mathbb{E}(Y|x) \leq 1.$$

Za analizu takvih podataka predložene su razne funkcije, no zbog jednostavnosti primjene i interpretacije najprihvaćenija je logistička.

Radi jednostavnosti zapisa uvedimo oznaku $\pi(x) = \mathbb{E}(Y|x)$. U ovom potpoglavlju pretpostavljamo da je zavisna varijabla binarna (poprima vrijednost 0 ili 1) i da je dana samo jedna nezavisna varijabla. Promatramo model dan s

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Gornji model nazivamo univarijatnim ili jednostavnim logističkim modelom. β_0 i β_1 su nepoznati parametri modela, čijom se procjenom bavimo u potpoglavlju 1.3.4. Također, neophodno je navesti i transformaciju $\pi(x)$ koju

nazivamo *logit transformacijom*, a definira se kao

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x.$$

Važnost ove transformacije leži u tome što $g(x)$ ima mnoga poželjna svojstva linearnog regresijskog modela. Logit, $g(x)$, je linearan, može biti neprekidan i može poprimati vrijednosti od $-\infty$ do $+\infty$, ovisno o rasponu varijable x.

Navedeni model ima široku primjenu u različitim područjima. Primjerice, u biostatistici, gdje su binarne zavisne varijable česta pojava (npr. pacijent je živ ili mrtav, ima/nema srčanu bolest i sl.). U idućim potpoglavljima promatramo dvije generalizacije univarijatnog modela i procjenjujemo njihove parametre.

1.3.2 Multivarijatni logistički model

U prethodnom potpoglavlju je predstavljen logistički model sa samo jednim prediktorom. Ipak, u praksi se često javljaju problemi koji zahtijevaju modele s više prediktora. U ovom potpoglavlju dajemo generalizaciju univarijatnog logističkog modela na slučaj više od jedne nezavisne varijable, to jest promatramo multivarijatni logistički model.

Neka je dano $p \in \mathbb{N}\backslash\{1\}$ međusobno nezavisnih slučajnih varijabli $X_1, X_2, \ldots, X_p$ i neka je $X = (X_1, \ldots, X_p)$ slučajan vektor. Zasad pretpostavimo da su sve nezavisne varijable numeričke. Radi jednostavnosti s $\pi(x) = P(Y = 1|x)$ označimo uvjetnu vjerojatnost pozitivnog ishoda, pri čemu je $x = (x_1, \ldots, x_p)$ realizacija slučajnog vektora X. Logit multivarijatnog logističkog modela je dan s

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

a multivarijatni logistički model je oblika

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}.$$

Ako je neka od nezavisnih varijabli kategorijska, uvode se indikatorske (" *dummy*") varijable. Pretpostavimo da je nezavisna varijabla X_j kategorijska za neki $j = 1, 2, \ldots, p$ i neka poprima $k \in \mathbb{N}$ različitih vrijednosti $L_1, L_2, \ldots, L_k$. Tada se uvodi $k-1$ indikatorskih varijabli i logit multivarijatnog logističkog modela je dan izrazom

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \sum_{l=1}^{k-1} \beta_{jl} D_{jl} + \ldots + \beta_p x_p,$$

gdje su D_{jl} indikatorske varijable zadane s:

$$D_{jl} = \begin{cases} 1, & x_j = L_l \\ 0, & inače \end{cases},$$

$l = 1, 2, \ldots, k-1$.

1.3.3 Multinomni logistički model

U ovom potpoglavlju promatramo generalizaciju multivarijatnog logističkog modela koja se koristi u slučajevima kad zavisna varijabla poprima jednu od $K \geq 3$ vrijednosti, odnosno pripada jednoj od K klasa. Takav model nazivamo multinomnim logističkim modelom, a dan je sljedećim izrazima (uz napomenu da je $\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})$, $i = 1, \ldots, K-1$):

$$\begin{aligned} \ln\left[\frac{P(Y=1|X=x)}{P(Y=K|X=x)}\right] &= \beta_{10} + x\beta_1^T \\ \ln\left[\frac{P(Y=2|X=x)}{P(Y=K|X=x)}\right] &= \beta_{20} + x\beta_2^T \\ &\vdots \\ \ln\left[\frac{P(Y=K-1|X=x)}{P(Y=K|X=x)}\right] &= \beta_{(K-1)0} + x\beta_{K-1}^T. \end{aligned}$$

Model je određen $(K-1)$-om logit transformacijom, odražavajući ograničenje

da je zbroj vjerojatnosti jednak jedan. Iako je u nazivnicima gornjih izraza korištena posljednja (K-ta) klasa, odabirom bilo koje druge klase se dobije ekvivarijantna procjena. Jednostavan račun daje

$$P(Y = k|X = x) = \frac{\exp(\beta_{k0} + x\beta_k^T)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + x\beta_l^T)}, \quad k = 1, \ldots, K-1,$$
$$P(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + x\beta_l^T)}$$

i lako se vidi da ovi izrazi u sumi daju jedan. Kako bismo naglasili ovisnost o cijelom skupu parametara $\theta = \{\beta_{10}, \beta_1, \ldots, \beta_{(K-1)0}, \beta_{K-1}\}$, uvedimo oznaku $P(Y = k|X = x) = p_k(x; \theta)$.

1.3.4 Procjena parametara modela

Parametri modela logističke regresije se obično procjenjuju metodom maksimalne vjerodostojnosti (MLE, "*maximum likelihood estimation*"), koristeći pri tome uvjetnu vjerojatnost $P(Y|X)$. Funkcija vjerodostojnosti za N opservacija je dana s

$$l(\theta) = \prod_{i=1}^{N} p_{y_i}(x_i; \theta),$$

gdje je $p_k(x_i; \theta) = P(Y = k|X = x_i; \theta)$. Radi jednostavnosti se najčešće maksimizira logaritmirana funkcija vjerodostojnosti koju nazivamo log-vjerodostojnost. Log-vjerodostojnost za N opservacija je

$$L(\theta) = \ln(l(\theta)) = \sum_{i=1}^{N} \ln p_{y_i}(x_i; \theta).$$

Uočimo da je, jer je ln strogo rastuća funkcija, maksimizacija log-vjerodostojnosti ekvivalentna maksimizaciji funkcije vjerodostojnosti.

Dalje ćemo, radi jednostavnosti, promatrati slučaj u kojem je zavisna varijabla binarna. Neka je $p_1(x; \theta) = p(x; \theta)$ i $p_0(x; \theta) = 1 - p(x; \theta)$. Log-

vjerodostojnost se može zapisati kao

$$L(\beta) = \sum_{i=1}^{N} \Big\{ y_i \ln p(x_i;\beta) + (1-y_i)\ln(1-p(x_i;\beta)) \Big\}$$
$$= \sum_{i=1}^{N} \Big\{ y_i x_i \beta^T - \ln(1+e^{x_i\beta^T}) \Big\},$$

pri čemu je $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, a (x_i, y_i), $i = 1, \ldots, N$, je dani uzorak od N, $N \in \mathbb{N}$, nezavisnih opservacija. Pri tome je y_i opažena vrijednost zavisne varijable te $x_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$, gdje je $x_{i0} = 1$, sadrži vrijednosti nezavisnih varijabli i-tog člana uzorka.

Da bismo maksimizirali log-vjerodostojnost, izjednačavamo njene derivacije s nulom:

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} (y_i x_{ij} - p(x_i;\beta)x_{ij}) = 0, \quad j = 0, \ldots, p,$$

pri čemu dobivamo $p+1$ nelinearnih jednadžbi. Često korištena metoda za rješavanje je Newton-Raphsonova iterativna metoda. Za primjenu ove metode potrebno je poznavati i derivacije drugog reda:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{N} x_{ij} x_{il}\, p(x_i;\beta)(1-p(x_i;\beta)), \quad j, l = 0, \ldots, p.$$

Počevši s β^{stari}, novu vrijednost β dobivamo na način:

$$\beta^{novi} = \beta^{stari} - (\mathbf{H})^{-1} \frac{\partial L(\beta)}{\partial \beta},$$

gdje je $\mathbf{H}$ Hesseova matrica funkcije L u točki β^{stari}, a $\frac{\partial L(\beta)}{\partial \beta} = \left(\frac{\partial L(\beta)}{\partial \beta_0}, \ldots, \frac{\partial L(\beta)}{\partial \beta_p} \right)^T$.

Zgodno je gornje izraze prikazati matričnim zapisom. Neka $\mathbf{y}$ označava vektor s vrijednostima y_i, $\mathbf{X}$ matricu tipa $N \times (p+1)$ čiji retci sadrže podatke za svaki element uzorka i oblika su $(1, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, N$, $\mathbf{p}$ vektor procijenjenih vjerojatnosti s i-tim elementom $p(x_i; \beta^{stari})$ i $\mathbf{W}$ dijagonalnu

matricu reda N s i-tim elementom dijagonale $p(x_i;\beta^{stari})(1-p(x_i;\beta^{stari}))$. Slijedi

$$\frac{\partial L(\beta)}{\partial \beta} = \mathbf{X}^T(\mathbf{y}-\mathbf{p})$$
$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = \left[-\mathbf{X}^T\mathbf{W}\mathbf{X}\right]_{(j+1)(l+1)}, \quad j,l = 0,\ldots,p.$$

Newtonov korak je tada

$$\beta^{novi} = \beta^{stari} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}).$$

1.3.5 *Stepwise* metode za izgradnju modela

Nakon upoznavanja s logističkim modelima i metodom za procjenu njihovih parametara prelazimo na temu izgradnje modela. Čest problem u praksi je koje nezavisne varijable uključiti u model da bi on bio što "bolji", stoga su razvijene razne metode za njihov odabir. Međutim, za uspješnu izgradnju modela se ne možemo osloniti samo na takve metode, već i na teorijsko znanje, iskustvo i zdrav razum. Na primjer, čest kriterij za uključivanje prediktora u multivarijatni model je njegova značajnost u univarijatnom modelu. No ponekad varijable koje se nisu pokazale značajnim u univarijatnim modelima mogu biti teorijski povezane sa zavisnom varijablom i pridonijeti značajnosti multivarijatnog modela. U ovom potpoglavlju opisujemo dvije *stepwise* metode za izgradnju modela: metodu unaprijed i metodu unatrag.

Metoda unaprijed ("*forward stepwise selection*") započinje rad s praznim modelom i u njega, s obzirom na unaprijed određen kriterij, uključuje jedan po jedan prediktor. Točnije, u svakom koraku uključuje prediktor koji u najvećoj mjeri poboljšava prilagodbu modela. Metoda staje s radom nakon što se uključe svi prediktori ili ako uključivanje novog prediktora značajno ne poboljšava model.

S druge strane, metoda unatrag ("*backward stepwise selection*") započinje

rad s punim modelom (modelom koji sadržava sve prediktore) i izbacuje jedan po jedan prediktor dok nije zadovoljen određeni kriterij zaustavljanja. U svakom koraku izbacuje se prediktor koji ima najmanji utjecaj na prilagodbu modela.

Osim dviju navedenih, postoje i kombinirane stepwise metode. Takve metode u svakom koraku izbacuju neku od uključenih ili uključuju novu varijablu u model. Često se kao kriterij za uključivanje odnosno izbacivanje prediktora koristi Akaikeov informacijski kriterij (AIC, *"Akaike information criterion"*). AIC vrijednost se računa formulom

$$\mathrm{AIC} = 2k - 2\ln(\hat{l}),$$

gdje je k broj procijenjenih parametara, a $\hat{l}$ maksimizirana vrijednost funkcije vjerodostojnosti modela. Prilikom usporedbe dvaju modela, boljim smatramo model s manjom AIC vrijednosti.

1.3.6 Logistička regresija kao klasifikator

Pretpostavimo da smo izgradili model logističke regresije s dihotomnom zavisnom varijablom. Prirodno se postavlja pitanje koliko učinkovito taj model predviđa vrijednosti zavisne varijable. Jednostavan način evaluacije ovakvog modela je korištenjem matrice zabune (tablica 1.1). Najprije treba usporediti stvarne vrijednosti varijable odziva i nove, dihotomne varijable čije se vrijednosti izvode iz procijenjenih vjerojatnosti logističkog modela.

Da bismo dobili izvedenu dihotomnu varijablu, prvo trebamo odrediti graničnu vrijednost c i zatim je usporediti s procijenjenim vjerojatnostima. U slučaju da je procijenjena vjerojatnost veća od c, vrijednost izvedene varijable postavljamo na 1, dok u suprotnom stavljamo da je jednaka 0. Često se za graničnu vrijednost uzima $c = 0.5$.

1.3. Logistička regresija

U potpoglavlju 1.2 smo naveli razne mjere za evaluaciju rada klasifikatora. Osim njih, za ocjenjivanje logističkih modela se često koriste ROC krivulje ("*receiver operating characteristic curve*") odnosno vrijednosti površina ispod tih krivulja (AUC, "*area under the curve*"). ROC krivulja prikazuje vjerojatnosti ispravne klasifikacije pozitivnih podataka (osjetljivost) uz prisutnost šuma, odnosno lažno pozitivnih podataka (1 - specifičnost) za sve moguće granične vrijednosti c. Očito, ako tražimo optimalan c za naš model, biramo c koji maksimizira osjetljivost i specifičnost.

Površina ispod krivulje, AUC, je mjera sposobnosti klasifikatora da razlikuje pozitivne od negativnih podataka i poprima vrijednosti između 0 i 1. Konačno, dajemo smjernice za evaluaciju klasifikatora na temelju AUC-a:

- Ako je AUC $= 0.5$, klasifikator je bezvrijedan
- Ako je AUC $\in [0.7, 0.8\rangle$, klasifikator je prihvatljiv
- Ako je AUC $\in [0.8, 0.9\rangle$, klasifikator je vrlo dobar
- Ako je AUC ≥ 0.9, klasifikator je izvrstan.

1.4 Slučajne šume

Osim logističke regresije, u ovom radu ćemo koristiti i *slučajne šume* (”*random forests*”), metodu strojnog učenja koja tijekom rada konstruira više stabala odluke. Stablo odluke je jedan od moćnijih algoritama za nadzirano učenje, a koristi se za zadatke klasifikacije i regresije. Prilikom izvršavanja gradi strukturu stabla sličnu dijagramu toka, gdje svaki unutarnji čvor označava atribut, svaka grana predstavlja klasifikacijsko pravilo, a svaki list (završni čvor) sadrži oznaku klase, odnosno rezultat algoritma. Stablo se konstruira rekurzivnim dijeljenjem podataka za treniranje u podskupove na temelju vrijednosti atributa sve dok se ne ispuni kriterij zaustavljanja, poput najveće dubine stabla ili minimalnog broja uzoraka potrebnih za dijeljenje čvora.

1.4.1 *Bagging* metoda

Bagging ili *bootstrap aggregation* je metoda smanjivanja varijance procijenjene funkcije predviđanja. *Bagging* je, između ostalog, izrazito prikladan za rad sa stablima. U slučaju regresije, isto stablo se prilagođava (*”fit”*) *bootstrap*[1]-uzorkovanim verzijama podataka za treniranje i zatim se uzima prosječna vrijednost rezultata. Kod klasifikacije, svaki član *odbora (”committee”)* stabala daje glas predviđenoj klasi i algoritam za rezultat uzima najmnogobrojniji glas.

Slučajne šume su važna modifikacija *bagging* metode koja izgrađuje velik skup nekoreliranih stabala i zatim ih usrednjuje. Algoritam slučajnih šuma je stekao veliku popularnost zbog jednostavnosti treniranja i prilagodbe (*”tuning”*), stoga je implementiran u mnogim programskim paketima. Navedimo

[1] *Bootstrapping* je statistički postupak za višestruko uzorkovanje skupa podataka, pri čemu se stvara mnogo simuliranih uzoraka.

i definiciju slučajnih šuma:

Definicija 1.1 ***(slučajne šume)*** *Slučajne šume su klasifikator koji se sastoji od kolekcije stabala* $\{T(x;\Theta_k), k = 1, \dots\}$*, pri čemu su* $\{\Theta_k\}$ *nezavisni, jednako distribuirani slučajni vektori i svako stablo daje glas najpopularnijoj klasi pri ulaznom vektoru* x*.*

1.4.2 Algoritam slučajnih šuma

Varijanca prosjeka B nezavisnih, jednako distribuiranih slučajnih varijabli s varijancama σ^2 iznosi $\frac{1}{B}\sigma^2$. Ako su varijable jednako distribuirane, ne nužno i nezavisne, s međusobnom pozitivnom korelacijom ρ, varijanca prosjeka iznosi

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{1.1}$$

Ideja algoritma slučajnih šuma je smanjenje varijance smanjenjem međusobne korelacije stabala, što se postiže izgradnjom stabala pomoću slučajno odabranih varijabli. Osnovni koraci algoritma su:

1. Za $b = 1, \dots, B$, gdje je B broj stabala:
   a) Uzmi *bootstrap* uzorak veličine N iz skupa za treniranje
   b) Izgradi stablo T_b na temelju *bootstrap* uzorka rekurzivno ponavljajući sljedeće korake za svaki list stabla sve dok se ne dosegne minimalna veličina čvora n_{min}
      i. Nasumično odaberi m varijabli od p mogućih
      ii. Među m odabranih, odaberi najbolju varijablu/točku razdvajanja
      iii. Podijeli čvor na dva nova čvora (potomka)

2. Spremi skup (*"ensemble"*) stabala $\{T_b\}_1^B$.

Za predviđanje u novoj točki x:

- *Regresija:* $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.
- *Klasifikacija:* Neka je $\hat{C}_b(x)$ klasa koju je predvidjelo b-to stablo slučajne šume. Tada je $\hat{C}_{rf}^B(x) =$ većinski glas $\{\hat{C}_b(x)\}_1^B$.

Nakon izgradnje B stabala, $\{T(x;\Theta_b)\}_1^B$, (regresijski) prediktor slučajne šume je

$$\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T(x;\Theta_b).$$

Smanjenjem broja m se smanjuje korelacija među parovima stabala u ansamblu, a zatim i, kao posljedica (1.1), varijanca prosjeka.

Dodatno, navodimo i preporuke za izbor broja m:

- Za klasifikaciju, standardna vrijednost m iznosi $\lfloor\sqrt{p}\rfloor$, a minimalna 1.
- Za regresiju, standardna vrijednost m iznosi $\lfloor p/3\rfloor$, a minimalna 5.

U praksi, izbor najpovoljnijeg m ovisi o konkretnom problemu, stoga ovaj hiperparametar treba promatrati kao hiperparametar prilagodbe (*tuning*).

1.4.3 Pretreniranje, odabir varijabli i OOB podaci

Kad je ukupan broj varijabli velik, a udio značajnih varijabli mali, veliki su izgledi da će slučajne šume raditi loše u slučaju izbora male vrijednosti hiperparametra m. Pri svakom razdvajanju čvorova, vjerojatnost odabira značajnih varijabli je mala. S druge strane, povećanjem broja značajnih varijabli slučajne šume postaju izrazito otporne na povećanje šuma u podacima.

Poznato je kako su stabla odluke sklona pretreniranju (*overfitting*). Do pretreniranja, prenaučenosti ili prekomjerne specijalizacije dolazi kada model savršeno opisuje podatke na kojima je izgrađen, ali mu je prediktivna moć na novim skupovima podataka mala. Slijedi da će i u slučaju slučajnih šuma (kao skupova stabala) postojati mogućnost prenaučenosti. Ipak, uvođenjem "šume" se osigurava bolji performans na novim skupovima podataka. Jedan od načina je upravo slučajnim odabirom varijabli pri izgradnji stabala. Nadalje, povećanje vrijednosti hiperparametra B (broja stabala) ne uzrokuje pretreniranje slučajnih šuma. Napomenimo da se pretreniranje rijetko viđa kod primjene slučajnih šuma na klasifikacijske probleme, dok su regresijske šume osjetljivije na taj fenomen.

Još jedna bitna karakteristika slučajnih šuma je korištenje *out of bag* (OOB) podataka. To su podaci koji su izostavljeni u procesu uzorkovanja, to jest podaci koji ne pripadaju *bootstrap* uzorku. Za svaku opservaciju se izgrađuje prediktor slučajne šume, i to usrednjujući samo ona stabla koja odgovaraju uzorcima u kojima se ta opservacija ne pojavljuje.

OOB procjena greške je približno jednaka procjeni dobivenoj N-strukom unakrsnom validacijom[2]. Dakle, može se koristiti i za procjenu greške slučajnih šuma. Prednost korištenja OOB procjene greške je mogućnost istovremenog treniranja i testiranja modela. U trenutku stabilizacije OOB greške proces treniranja se može prekinuti.

1.4.4 Programski paketi u R-u

Kao što smo ranije naveli, slučajne šume su jedan od najpopularnijih i najkorištenijih algoritama strojnog učenja. Jednostavne su za korištenje i robusne, što skraćuje vrijeme potrebno za pretprocesiranje ulaznih podataka.

[2]Jedna od najkorištenijih metoda za procjenu pogreške predviđanja.

U R-u, statističkom programu koji ćemo koristiti za izgradnju modela, su dostupni razni paketi za rad sa slučajnim šumama. Budući da ćemo dva takva paketa primijeniti u Poglavlju 3, dajemo kratak opis svakog od njih.

randomForest je standardni paket u kojem je implementiran istoimeni algoritam slučajnih šuma. Navedeni algoritam se temelji na najosnovnijoj logici slučajnih šuma. Jako je robustan i lagan za korištenje. Moguće je ugađati hiperparametre poput broja i dubine stabala i broja slučajno odabranih varijabli m. *randomForest* se može koristiti za regresijske i klasifikacijske probleme.

cforest je implementacija algoritma slučajnih šuma dostupna u paketu *party*. *cforest* je računski skuplji od *randomForest* paketa, ali često daje bolje rezultate. Sporiji je i može obraditi manje podataka za jednaku količinu memorije. Da bi dobio konačan ansambl, koristi ponderirani prosjek izgrađenih stabala. Važna prednost *cforesta* nad *randomForest* paketom je izgradnja nepristranih stabala. Naime, jednostavni algoritmi poput onih implementiranih u *randomForest* paketu često favoriziraju varijable s mnogo potencijalnih točaka razdvajanja (varijable s mnogo kategorija ili kontinuirane varijable). Ukoliko su računalni resursi dovoljni, preporučuje se korištenje *cforesta*.

Poglavlje 2

k-adski Jaccardov koeficijent

Razni podaci se mogu predstaviti binarnim nizovima. Često se binarnim brojevima označava prisutnost ili odsutnost određenog svojstva nekog objekta. Na primjer, u psihologiji, objekti mogu biti osobe koje mogu i ne moraju imati određenu osobinu; u ekologiji, objekti mogu biti regije ili okruzi u kojima se određene vrste pojavljuju ili ne pojavljuju (ili obrnuto, objekti su vrste koje koegzistiraju na više lokacija); u arheologiji, objekti mogu biti grobovi u kojima mogu biti pronađene određene vrste artefakata itd. U ovom odjeljku se bavimo koeficijentima sličnosti razvijenim za binarne skupove podataka.

2.1 Definicija k-adskog koeficijenta sličnosti

Koeficijentom sličnosti (*"similarity coefficient"*) nazivamo bilo koju statističku mjeru sličnosti (povezanosti) dvaju objekata. Takve mjere se mogu koristiti i za usporedbu dvaju klastera unutar danog skupa podataka.

Neka je O konačan skup objekata (označenih s $j_1, j_2, j_3, \dots$) i označimo s n broj atributa ($n > 0$). Dijadski (*"2-adic"*) koeficijent sličnosti se definira

kao preslikavanje $S : O \times O \to \mathbb{R}$ takvo da vrijedi

$$S(j_1, j_1) \geq S(j_1, j_2) \quad \text{i} \quad S(j_1, j_2) = S(j_2, j_1), \quad \forall j_1, j_2 \in O.$$

Mnogi koeficijenti sličnosti imaju svojstvo $S(j_1, j_1) = 1$.

Osim na parovima objekata, koeficijenti sličnosti se mogu definirati i na trojkama, četvorkama, ili, općenito, k-torkama objekata. Trijadski (*"3-adic"*) koeficijent sličnosti se definira kao preslikavanje $S : O \times O \times O \to \mathbb{R}$ takvo da je

$$S(j_1, j_1, j_1) \geq S(j_1, j_1, j_2) \geq S(j_1, j_2, j_3)$$

i vrijedi trosmjerna simetrija,

$$S(j_1, j_2, j_3) = S(j_1, j_3, j_2) = S(j_2, j_1, j_3)$$
$$= S(j_2, j_3, j_1) = S(j_3, j_1, j_2) = S(j_3, j_2, j_1),$$

$\forall j_1, j_2, j_3 \in O$. Nadalje, zahtijeva se da trijadski koeficijent sličnosti zadovoljava $S(j_1, j_1, j_2) = S(j_1, j_2, j_2), \forall j_1, j_2 \in O$, odnosno, ako su dva objekta identična, sličnost među neidentičnim objektima treba biti jednaka, bez obzira na to koja dva objekta su jednaka.

Definicija k-adskog (*"k-adic"*) koeficijenta sličnosti $S(j_1, j_2, \ldots, j_k),\ k \geq 2$, uključujući k-smjernu simetriju, je analogna definiciji trijadskog koeficijenta sličnosti. Očito, k-adski koeficijenti sličnosti se mogu koristiti za istovremenu usporedbu k binarnih nizova.

2.2 Jaccardov koeficijent sličnosti

Paul Jaccard (1912) je proučavao rasprostranjenost određenog bilja u Alpama. U njegovom području istraživanja objekti su bili tri različita alpska okruga, a atributi vrste bilja. Da bi izmjerio sličnost dvaju okruga u smislu

prisutnosti vrsta bilja, Jaccard je koristio omjer

$$S_{Jac}^{(2)} = \frac{\textit{Broj vrsta zajedničkih dvama okruzima}}{\textit{Ukupan broj vrsta u dvama okruzima}}.$$

Uvedimo sljedeće varijable za dijadski koeficijent sličnosti $S(j_1, j_2)$:

- a = broj atributa prisutnih i u j_1 i u j_2
- b = broj atributa prisutnih u j_1, ali ne i u j_2
- c = broj atributa prisutnih u j_2, ali ne i u j_1
- d = broj atributa odsutnih i u j_1 i u j_2.

Uočimo da je $a + b + c + d = n$ (prisjetimo se, n je broj atributa).

Nadalje, umjesto varijabli a, b, c i d, za k binarnih n-vektora $j_1, j_2, \ldots, j_k$ definirajmo tri varijable:

- $x^{(k)}$ = broj atributa prisutnih u $j_1, j_2, \ldots, j_k$
- $z^{(k)}$ = broj atributa odsutnih u $j_1, j_2, \ldots, j_k$
- $y^{(k)} = n - x^{(k)} - z^{(k)}$, to jest broj nepodudaranja.

Iz navedenog slijedi $x^{(2)} = a$, $z^{(2)} = d$ i $y^{(2)} = b + c$, a zatim i

$$S_{Jac}^{(2)} = \frac{a}{a+b+c} = \frac{x^{(2)}}{x^{(2)} + y^{(2)}} = \frac{x^{(2)}}{n - z^{(2)}}. \tag{2.1}$$

Očita trijadska formulacija Jaccardovog koeficijenta sličnosti bi glasila

$$S_{Jac}^{(3)} = \frac{\textit{Broj vrsta zajedničkih trima okruzima}}{\textit{Ukupan broj vrsta u trima okruzima}} = \frac{x^{(3)}}{x^{(3)} + y^{(3)}},$$

a k-adska formulacija (2.1) je

$$S_{Jac}^{(k)} = \frac{x^{(k)}}{x^{(k)} + y^{(k)}} = \frac{x^{(k)}}{n - z^{(k)}}. \tag{2.2}$$

2.3 Konsenzus (”*consensus agreement*”)

Skupovi binarnih podataka mogu, između ostalog, biti generirani dvama skupovima binarnih klasifikatora. Ocjenjivanje njihove sličnosti je problem koji se javlja u mnogim istraživačkim područjima, na primjer u biomedicini, psihologiji ili ekonomiji. U ovom radu naglasak je na procjeni sličnosti u konsenzusu dvaju skupova binarnih klasifikatora. Kažemo da skup klasifikatora ima konsenzus o klasifikaciji danog objekta ako ga svi klasifikatori iz tog skupa jednako klasificiraju. Glavna ideja je da su klasifikatori primijenjeni na istom skupu objekata sličniji ako rezultiraju sličnijim klasifikacijama odnosno, u kontekstu konsenzusa, ako induciraju klasifikacije sa sličnijim konsenzusom. Dalje, uvedimo osnovne definicije.

Označimo s $C = \{c_1, \ldots, c_l\}$ konačan skup klasifikatora i s $O = \{o_1, \ldots, o_n\}$ konačan skup objekata koji mogu biti klasificirani kao 1 (označava prisutnost svojstva) ili 0 (označava odsutnost svojstva) na temelju klasifikatora $c_1, \ldots, c_l$. Navedimo definicije spomenutih pojmova:

Definicija 2.1 *(**binarni klasifikator**) Neka je $O = \{o_1, \ldots, o_n\}$ konačan skup objekata. Funkciju $c : O \to \{0, 1\}$ nazivamo binarnim klasifikatorom na O.*

Definicija 2.2 *(**konsenzus**) Neka je $C = \{c_1, \ldots, c_l\}$ konačan skup klasifikatora i $O = \{o_1, \ldots, o_n\}$ konačan skup objekata.*

(1) Skup klasifikatora C ima konsenzus o prisutnosti svojstva za objekt $o \in O$ ako je $c_j(o) = 1,\ j = 1, 2, \ldots, l$.

(2) Skup klasifikatora C ima konsenzus o odsutnosti svojstva za objekt $o \in O$ ako je $c_j(o) = 0,\ j = 1, 2, \ldots, l$.

Veličina konsenzusa o prisutnosti unutar skupa se može izraziti kao broj objekata koje su svi klasifikatori iz tog skupa klasifikatora klasificirali kao 1. Slično vrijedi i za konsenzus o odsutnosti, to jest promatra se broj objekata klasificiranih kao 0 od strane svih klasifikatora. Da bismo ocijenili sličnost u konsenzusu o prisutnosti unutar jednog skupa, koristimo prethodno spomenutu k-adsku formulaciju Jaccardovog koeficijenta (formula (2.2)). Očito, veće vrijednosti navedenog koeficijenta sugeriraju veću sličnost u konsenzusu. U sljedećem potpoglavlju navodimo varijantu k-adskog Jaccardovog koeficijenta koja se koristi za ocjenu sličnosti dvaju skupova.

2.4 k-adski Jaccardov koeficijent za dva skupa

Nakon upoznavanja s konsenzusom prelazimo na konstrukciju k-adskog Jaccardovog koeficijenta sličnosti dvaju skupova (*"2-group k-adic Jaccard similarity coefficient"*). Prilikom ocjenjivanja sličnosti dvaju skupova treba uzeti u obzir i sličnost unutar svakog od njih. Koeficijent čiju konstrukciju navodimo je primjeren za ocjenu sličnosti u konsenzusu o prisutnosti svojstva dvaju skupova binarnih klasifikatora.

Neka su $C_1 = \{c_{11}, \ldots, c_{1l_1}\}$ i $C_2 = \{c_{21}, \ldots, c_{2l_2}\}$ dva skupa klasifikatora na skupu objekata $O = \{o_1, \ldots, o_n\}$, pri čemu se oba skupa sastoje od barem dvaju klasifikatora ($l_1, l_2 \geq 2$). Neka je a_l broj objekata za koje vrijedi da su klasificirani kao 1 za svaki klasifikator iz $C_l,\ l = 1, 2$. Neka je d_l broj objekata za koje vrijedi da su klasificirani kao 0 za svaki klasifikator iz $C_l,\ l = 1, 2$. k-adski Jaccardov koeficijent sličnosti dvaju skupova se računa na sljedeći način:

1. Izračunaju se k-adski Jaccardovi koeficijenti posebno za svaki skup:

$$J_{C_l} = \frac{a_l}{n - d_l}, \quad l = 1, 2. \tag{2.3}$$

2. Klasifikatori iz odabranih skupova se spoje u jedan skup $C = C_1 \cup C_2 = \{c_{11}, \ldots, c_{1l_1}, c_{21}, \ldots, c_{2l_2}\}$, pri čemu C sadržava $l_1 + l_2$ klasifikatora iz C_1 i C_2. Zatim se izračuna k-adski Jaccardov koeficijent J_C za skup klasifikatora C primjenjujući $J_C = \frac{a_C}{n - d_C}$, pri čemu je a_C broj objekata za koje vrijedi da su klasificirani kao 1 za svaki klasifikator iz C, a d_C je broj objekata koji su klasificirani kao 0 za svaki klasifikator iz C.

3. Izračuna se k-adski Jaccardov koeficijent sličnosti dvaju skupova primjenom formule

$$J_{group}(C_1, C_2) = \frac{J_C}{\frac{1}{2}(J_{C_1} + J_{C_2})}.$$

Problem neodređenosti gornjih koeficijenata se javlja u slučaju $\frac{1}{2}(J_{C_1} + J_{C_2}) = 0$ (odnosno $J_{C_1} = J_{C_2} = 0$) ili kao posljedica neodređenosti k-adske formulacije Jaccardovog koeficijenta. Prvi slučaj, $J_{C_1} = J_{C_2} = 0$, nastaje kad ni u jednom skupu nema konsenzusa o prisutnosti, to jest kad je $a_1 = a_2 = 0$, što znači da $\nexists o \in O$ takav da je $c_{1j}(o) = 1,\ j = 1, 2, \ldots, l_1$ i $\nexists o \in O$ takav da je $c_{2j}(o) = 1,\ j = 1, 2, \ldots, l_2$. U tom slučaju stavljamo $J_{group}(C_1, C_2) = 1$. Neodređenost k-adske formulacije Jaccardovog koeficijenta je vezana za slučaj kad su svi objekti klasificirani kao 0, to jest $n - d_l = 0,\ l = 1, 2$. Tada stavljamo da je $J_{group}(C_1, C_2) = 1$. Primijetimo da smo sličnost među skupovima procijenili u odnosu na srednju vrijednost sličnosti unutar skupova. Osim ovog, predloženi su i drugi načini izračuna, no njima se nećemo baviti u ovom radu.

Poglavlje 3

Primjena na stvarne skupove podataka

3.1 Opis problema

Procjena sličnosti binarnih varijabli je čest problem koji se javlja u raznim područjima poput ekologije, biomedicine, kemije, psihologije, računalnog jezikoslovlja i ekonomije. Nastojanja da se kvantificira njihova povezanost, odnosno sličnost rezultirala su različitim mjerama. S obzirom na to da izbor odgovarajuće mjere ovisi o području istraživanja i obilježjima podataka koji se analiziraju, ne postoji jedinstvena strategija koju bi trebalo primijeniti. Velik dio istraživanja u polju binarne sličnosti je usmjeren na evaluaciju sličnosti dvaju ili više binarnih vektora. Za probleme usporedbe dvaju skupova binarnih vektora teorijska pozadina i dalje nije dovoljno razvijena. Skupovi binarnih vektora mogu odgovarati skupovima stručnjaka, testnih instrumenata ili algoritama. Na primjer, u medicini nailazimo na različite skupove stručnjaka/algoritama koji klasificiraju slučajeve bolesti kao dobroćudne ili zloćudne, dok u ekonomiji susrećemo skupove klasifikatora koji identificiraju

projekte kao prihvatljive ili neprihvatljive.

U ovom radu mjerimo sličnost u konsenzusu o prisutnosti unutar jednog i između dvaju skupova binarnih klasifikatora. Primijenit ćemo ranije navedene klasifikacijske algoritme na dva stvarna skupa podataka i dati interpretaciju dobivenih rezultata. Obrada podataka i izgradnja klasifikacijskih modela je u cijelosti provedena u programu R.

3.2 Opis podataka

Analiziramo dva skupa podataka iz područja medicine. Prvi od njih, *heart disease*, se sastoji od podataka prikupljenih na uzorku od 303 pacijenta bolnice u Clevelandu. Cilj je utvrditi povezanost 13 odabranih varijabli s prisutnosti bolesti srca. Nakon učitavanja podataka u programsko okruženje potrebno je promotriti njihovu strukturu. Jednostavnim naredbama se dobije da za 6 pacijenata nisu dostupne vrijednosti svih varijabli pa su takvi ispitanici isključeni iz razmatranja. Nadalje, neke varijable su pogrešno tipizirane. Na primjer, tip varijable *sex* koja označava spol pacijenta je postavljen na numerički. Navedena varijabla je pretvorena u kategorijsku varijablu s vrijednostima 0 i 1 redom za ženski i muški spol. Varijabla od interesa je *target* i poprima cjelobrojne vrijednosti od 0 do 4, pri čemu 0 označava odsutnost, a ostale vrijednosti predstavljaju različite dijagnoze srčane bolesti. S obzirom na to da nas zanima odsutnost ili prisutnost bolesti bez obzira na točnu dijagnozu, konstruirana je binarna varijabla naziva *targetbin*. Prirodno, njena vrijednost je postavljena na 1 u slučaju dijagnosticirane bolesti, a inače je jednaka 0. Od ukupno 297 pacijenata (nakon isključivanja onih s nedostajućim informacijama), njih 137 ima dijagnosticiranu srčanu bolest pa zaključujemo da su klase 0 i 1 uravnotežene.

3.2. Opis podataka

Oznaka	**Opis**	**Tip**	**Vrijednosti**
age	dob u godinama	numerička	29 - 77
sex	spol	kategorijska	1 = muški; 0 = ženski
cp	vrsta bola u prsima	kategorijska	1 = tipična angina; 2 = atipična angina; 3 = neangiozni bol; 4 = nema bola
trestbps	krvni tlak u mirovanju (u mmHg na prijemu u bolnicu)	numerička	94 - 200
chol	kolesterol u serumu (mg/dl)	numerička	126 - 564
fbs	razina šećera u krvi natašte > 120 mg/dl	kategorijska	1 = da; 0 = ne
restecg	rezultat elektrokardiografije u mirovanju	kategorijska	0 = normalan; 1 = abnormalnost ST-T vala; 2 = vjerojatna ili sigurna hipertrofija
thalach	najveći postignuti broj otkucaja srca	numerička	71 - 202
exang	angina izazvana opterećenjem	kategorijska	1 = da; 0 = ne
oldpeak	depresija ST-segmenta izazvana vježbanjem u odnosu na stanje mirovanja	numerička	0 - 6.2
slope	oblik nagiba ST-segmenta pri vrhuncu vježbanja	kategorijska	1 = rast; 2 = mirovanje; 3 = pad
ca	broj glavnih krvnih žila označenih fluoroskopijom	kategorijska	0 - 3
thal	talasemija	kategorijska	3 = normalno; 6 = nepopravljivo oštećenje; 7 = reverzibilno oštećenje
targetbin	dijagnoza srčane bolesti	kategorijska	0 = nema dijagnozu; 1 = ima dijagnozu

Tablica 3.1: Opis varijabli za skup podataka *heart disease*.

3.2. Opis podataka

Svrha drugog skupa podataka naziva *stroke prediction* je predviđanje moždanog udara na temelju raznih varijabli vezanih za pacijentovo zdravlje i način života. Originalni skup podataka sadržava informacije o 5110 pacijenata.

Oznaka	**Opis**	**Tip**	**Vrijednosti**
gender	spol	kategorijska	"Male" = muški; "Female" = ženski
age	dob u godinama	numerička	0.08 - 82
hypertension	povišen krvni tlak	kategorijska	1 = da; 0 = ne
heart_disease	srčana bolest	kategorijska	1 = da; 0 = ne
ever_married	je li pacijent ikad bio u braku	kategorijska	"Yes" = da; "No" = ne
work_type	vrsta zaposlenja	kategorijska	"children" = djeca; "Govt_job" = državni posao; "Never_worked" = nikad zaposlen; "Private" = privatni posao; "Self-employed" = samozaposlen
Residence_type	područje stanovanja	kategorijska	"Urban" = gradsko; "Rural" = seosko
avg_glucose_level	prosječna razina šećera u krvi (u mg/dl)	numerička	55.12 - 271.74
bmi	indeks tjelesne mase	numerička	10.3 - 97.6
smoking_status	pušački status	kategorijska	"formerly smoked" = bivši pušač; "never smoked" = nikad nije pušio; "smokes" = pušač
stroke	pacijent doživio moždani udar	kategorijska	1 = da; 0 = ne

Tablica 3.2: Opis varijabli za skup podataka *stroke prediction*.

Ukupan broj varijabli je 12, no jedna od njih, *id*, je izbačena odmah nakon

učitavanja podataka jer ne sadrži informacije relevantne za analizu. Slično kao i kod prvog skupa podataka, bilo je potrebno isključiti retke (ispitanike) s nedostajućim vrijednostima i promijeniti tipove nekim varijablama. Nakon uređivanja podataka preostalo je 4908 pacijenata od kojih je 209 doživjelo moždani udar. Zaključujemo da se radi o nebalansiranom skupu podataka, i to o slučaju kad je klasa od interesa (*stroke* = 1) malobrojna.

Opis varijabli i njihovih vrijednosti nakon uređivanja dan je u tablicama 3.1 i 3.2 redom za skupove podataka *heart disease* i *stroke prediction*.

3.3 Izgradnja modela i rezultati

Nakon upoznavanja s podacima prelazimo na izgradnju modela. Prvo je potrebno podijeliti skupove podataka na skupove za treniranje i testiranje. Odabir podataka koji ulaze u ta dva skupa je proveden nasumično i to u omjeru 70:30 u korist skupa za treniranje. Kako je skup podataka *stroke prediction* nebalansiran, na pripadni skup za treniranje je primijenjena kombinacija nasumičnog poduzorkovanja ("*undersampling*") i preuzorkovanja ("*oversampling*"). Takav postupak izbacuje neke podatke iz mnogobrojnije klase i stvara nove, umjetne podatke koje pridružuje manjinskoj klasi. Nakon balansiranja klasa skup za treniranje sadržava 1714 podataka za koje je *stroke* = 1 i 1731 podatak za koji je *stroke* = 0. Modele ocjenjujemo na skupu za testiranje pomoću mjera danih u tablici 1.2 i AUC vrijednosti. Dobivene vrijednosti navedenih mjera za sve modele su dane u tablicama 3.4 i 3.6.

Promatramo skup podataka *heart disease*. Prva grupa modela koje gradimo su multivarijatni logistički modeli. Krećemo od punog modela, odnosno modela s uključenim svim prediktorima (oznaka *lr_heart1*). Nakon

toga pomoću *forward*, *backward* i kombinirane *stepwise* metode tražimo najbolji model. Kao kriterij za odabir optimalnog modela uzet je Akaikeov informacijski kriterij (AIC). Sve tri *stepwise* metode su rezultirale jednakim modelima pa su, osim dobivenog (*lr_heart2*), promatrani i drugi najbolji model po *backward stepwise* metodi i drugi najbolji model po *forward stepwise* metodi (*lr_heart3* i *lr_heart4*, redom). Tablica 3.3 prikazuje popis nezavisnih varijabli uključenih i isključenih iz svakog od izgrađenih modela.

	Model			
Prediktor	*lr_heart1*	*lr_heart2*	*lr_heart3*	*lr_heart4*
age	+	-	-	-
sex	+	+	+	+
cp	+	+	+	+
trestbps	+	+	+	+
chol	+	+	+	+
fbs	+	-	-	-
restecg	+	-	-	-
thalach	+	-	+	-
exang	+	+	+	-
oldpeak	+	-	-	-
slope	+	+	+	+
ca	+	+	+	+
thal	+	+	+	+

Tablica 3.3: Prediktori uključeni ili isključeni iz logističkih modela (*heart disease*).

Za graničnu vrijednost klasifikacije c je uzeta standardna vrijednost 0.5. Dakle, pacijenti čija je procijenjena vjerojatnost bolesti srca veća od 0.5 su pridruženi klasi 1, a ostali klasi 0. Promatrajući vrijednosti AUC-a u tablici 3.4 vidimo da su svi odabrani logistički modeli prihvatljivi odnosno vrlo dobri klasifikatori.

3.3. Izgradnja modela i rezultati

Model			Točnost	Osjetljivost	Specifičnost	Preciznost	F1	AUC
logistička	*lr_heart1*		0.8193	0.8500	0.7907	0.7907	0.8193	0.8203
	lr_heart2		0.7952	0.8500	0.7442	0.7556	0.8000	0.7971
	lr_heart3		0.8072	0.8500	0.7674	0.7727	0.8095	0.8087
	lr_heart4		0.7952	0.8250	0.7674	0.7674	0.7952	0.7962
cforest	$m = 2$	$n = 50$	0.8193	0.8750	0.7674	0.7778	0.8235	0.8212
		$n = 200$	0.7952	0.8500	0.7442	0.7556	0.8000	0.7971
		$n = 500$	0.7952	0.8500	0.7442	0.7556	0.8000	0.7971
	$m = 4$	$n = 50$	0.8072	0.8500	0.7674	0.7727	0.8095	0.8087
		$n = 200$	0.8193	0.8500	0.7907	0.7907	0.8193	0.8203
		$n = 500$	0.7711	0.8500	0.6977	0.7234	0.7816	0.7738
	$m = 10$	$n = 50$	0.7831	0.8500	0.7209	0.7391	0.7907	0.7855
		$n = 200$	0.7831	0.8500	0.7209	0.7391	0.7907	0.7855
		$n = 500$	0.7831	0.8500	0.7209	0.7391	0.7907	0.7855
randomForest	$m = 2$	$n = 50$	0.8072	0.9000	0.7209	0.7500	0.8181	0.8105
		$n = 200$	0.8434	0.9000	0.7907	0.8000	0.8471	0.8453
		$n = 500$	0.8313	0.8750	0.7907	0.7955	0.8333	0.8328
	$m = 4$	$n = 50$	0.8434	0.9000	0.7907	0.8000	0.8471	0.8453
		$n = 200$	0.8193	0.9000	0.7442	0.7660	0.8276	0.8221
		$n = 500$	0.8193	0.8750	0.7674	0.7778	0.8235	0.8212
	$m = 10$	$n = 50$	0.7590	0.8000	0.7209	0.7273	0.7619	0.7605
		$n = 200$	0.7711	0.8250	0.7209	0.7333	0.7765	0.7730
		$n = 500$	0.7711	0.8250	0.7209	0.7333	0.7765	0.7730

Tablica 3.4: Vrijednosti mjera za ocjenu modela (skup podataka *heart disease*).

Različiti klasifikatori slučajnih šuma čine preostala dva skupa izgrađenih modela. U prvom od njih se nalaze modeli dobiveni pomoću algoritma *cforest* sadržanog u paketu *party*, a u drugom modeli dobiveni pomoću algoritma *randomForest*. Hiperparametri po kojima se modeli unutar svakog skupa razlikuju su broj slučajno odabranih varijabli koje se uzimaju u obzir prilikom razdvajanja čvora m i broj izgrađenih stabala slučajne šume n. Odabrane vrijednosti hiperparametra m su 2, 4 i 10, dok broj stabala poprima vrijednosti 50, 200 i 500. Iz tablice 3.4 se vidi da izgrađeni klasifikatori imaju slične performanse na skupu podataka *heart disease*. Poput logističkih, *cfo-*

rest i *randomForest* modeli se mogu smatrati prihvatljivim ili vrlo dobrim klasifikatorima.

Izgradnja modela za drugi skup podataka, *stroke prediction*, je provedena na sličan način. Sve tri *stepwise* metode su dale jednake logističke modele, stoga su opet promatrani puni model (*lr_stroke1*), najbolji model po *stepwise* metodama (*lr_stroke2*), drugi najbolji model po *backward stepwise* metodi (*lr_stroke3*) i drugi najbolji model po *forward stepwise* metodi (*lr_stroke4*). Prediktori uključeni u navedene logističke modele su prikazani u tablici 3.5.

	Model			
Prediktor	*lr_stroke1*	*lr_stroke2*	*lr_stroke3*	*lr_stroke4*
gender	+	-	+	-
age	+	+	+	+
hypertension	+	+	+	+
heart_disease	+	-	-	-
ever_married	+	-	-	-
work_type	+	+	+	+
Residence_type	+	-	-	-
avg_glucose_level	+	+	+	+
bmi	+	+	+	-
smoking_status	+	+	+	+

Tablica 3.5: Prediktori uključeni ili isključeni iz logističkih modela (*stroke prediction*).

Odabrane vrijednosti hiperparametra m za klasifikatore slučajnih šuma su 2, 4 i 8. Kao kod skupa *heart disease*, broj stabala n poprima vrijednosti 50, 200 i 500. Promotrimo tablicu 3.6 u kojoj su prikazane vrijednosti mjera za ocjenu klasifikatora. Uočimo da su svi logistički i *cforest* modeli prihvatljivi, dok *randomForest* klasifikatori imaju neprihvatljivo niske vrijednosti pokazatelja specifičnosti. Nadalje, svi klasifikatori slučajne šume imaju veću osjetljivost, ali manju specifičnost od logističkih klasifikatora. Drugim riječima, bolje prepoznaju pozitivne podatke (*stroke = 1*), a slabije

negativne podatke (*stroke = 0*).

Model			Točnost	Osjetljivost	Specifičnost	Preciznost	F1	AUC
logistički	*lr_stroke1*		0.7423	0.7418	0.7544	0.9868	0.8469	0.7481
	lr_stroke2		0.7478	0.7468	0.7719	0.9878	0.8505	0.7594
	lr_stroke3		0.7416	0.7418	0.7368	0.9858	0.8466	0.7393
	lr_stroke4		0.7471	0.7468	0.7544	0.9868	0.8502	0.7506
cforest	$m = 2$	$n = 50$	0.7943	0.8001	0.6491	0.9825	0.8820	0.7246
		$n = 200$	0.7949	0.7994	0.6842	0.9842	0.8823	0.7418
		$n = 500$	0.7908	0.7952	0.6842	0.9842	0.8796	0.7397
	$m = 4$	$n = 50$	0.8195	0.8272	0.6316	0.9823	0.8981	0.7294
		$n = 200$	0.8250	0.8314	0.6667	0.9840	0.9013	0.7491
		$n = 500$	0.8230	0.8293	0.6667	0.9840	0.9000	0.7480
	$m = 8$	$n = 50$	0.8319	0.8421	0.5789	0.9801	0.9059	0.7105
		$n = 200$	0.8278	0.8371	0.5965	0.9808	0.9033	0.7168
		$n = 500$	0.8278	0.8378	0.5789	0.9800	0.9034	0.7084
randomForest	$m = 2$	$n = 50$	0.8599	0.8784	0.4035	0.9732	0.9233	0.6409
		$n = 200$	0.8619	0.8791	0.4386	0.9748	0.9245	0.6588
		$n = 500$	0.8606	0.8770	0.4561	0.9755	0.9236	0.6665
	$m = 4$	$n = 50$	0.9228	0.9509	0.2281	0.9681	0.9595	0.5895
		$n = 200$	0.9234	0.9502	0.2632	0.9695	0.9598	0.6067
		$n = 500$	0.9228	0.9509	0.2456	0.9688	0.9594	0.5979
	$m = 8$	$n = 50$	0.9132	0.9424	0.1930	0.9664	0.9543	0.5677
		$n = 200$	0.9139	0.9417	0.2281	0.9678	0.9546	0.5849
		$n = 500$	0.9139	0.9417	0.2281	0.9678	0.9546	0.5849

Tablica 3.6: Vrijednosti mjera za ocjenu modela (skup podataka *stroke prediction*).

Rezultati klasifikacije su spremljeni u tablice, posebno za svaki skup podataka i svaki skup klasifikatora. Pritom svaki stupac tablice predstavlja rezultate predviđanja jednog klasifikatora. U tablici 3.7 su prikazani rezultati predviđanja logističkih modela na skupu za testiranje skupa podataka *stroke prediction*. Napomenimo da svaki redak tablice predstavlja klasifikaciju jednog podatka skupa za testiranje različitim klasifikatorima.

	lr_stroke1	*lr_stroke2*	*lr_stroke3*	*lr_stroke4*
3	1	1	1	1
5	1	1	1	1
6	1	1	1	1
10	1	1	1	1
13	0	0	0	0
19	1	1	0	0
⋮	⋮	⋮	⋮	⋮
5110	0	0	0	0

Tablica 3.7: Rezultati klasifikacije logističkih modela (*stroke prediction*).

Nakon spremanja rezultata u tablice lako je izračunati sličnost u konsenzusu o prisutnosti unutar svakog skupa klasifikatora. Primjenom k-adskog Jaccardovog koeficijenta sličnosti (formula (2.3)) na svaki skup klasifikatora dobivene su vrijednosti prikazane u tablici 3.8.

	Skup podataka	
Skup klasifikatora	*heart disease*	*stroke prediction*
logistička regresija	0.9024	0.9211
cforest	0.8333	0.6313
randomForest	0.7674	0.2385

Tablica 3.8: Ocjena sličnosti unutar skupa klasifikatora primjenom k-adskog Jaccardovog koeficijenta sličnosti.

Uočavamo da skupovi logističkih klasifikatora imaju najveću sličnost u konsenzusu o prisutnosti unutar skupa klasifikatora, bez obzira na skup podataka. *cforest* i *randomForest* klasifikatori imaju niži stupanj sličnosti unutar skupa klasifikatora, pogotovo na skupu podataka *stroke prediction*.

Izračunajmo i sličnost u konsenzusu o prisutnosti između dvaju skupova klasifikatora. Najprije kreiramo novu tablicu "spajanjem" dviju odgovarajućih tablica, a zatim primijenimo postupak za računanje koeficijenta

sličnosti prikazan u odlomku 2.4. Dobivene vrijednosti su navedene u tablicama 3.9 i 3.10 redom za skupove podataka *heart disease* i *stroke prediction*.

Skup klasifikatora	*cforest*	*randomForest*
logistička regresija	0.8706	0.8983
cforest	1	0.8963

Tablica 3.9: k-adski Jaccardov koeficijent sličnosti dvaju skupova klasifikatora na skupu podataka *heart disease*.

Prvo promotrimo rezultate na skupu podataka *heart disease*. Interpretacija je sljedeća: prilikom udruživanja klasifikatora iz skupova *cforest* i *randomForest* u jedan skup očuvano je 89.63 % prosječne sličnosti unutar skupa klasifikatora. Ovo znači da se spajanjem tih dvaju skupova gubi 10.37 % prosječne sličnosti unutar skupa klasifikatora. Slično vrijedi i za ostale parove.

Skup klasifikatora	*cforest*	*randomForest*
logistička regresija	0.6073	0.2210
cforest	1	0.3562

Tablica 3.10: k-adski Jaccardov koeficijent sličnosti dvaju skupova klasifikatora na skupu podataka *stroke prediction*.

Uočavamo da je sličnost svakih dvaju (različitih) skupova klasifikatora nad podacima *stroke prediction* manja nego u prethodnom slučaju. Najmanje su slični skupovi *logistička regresija* i *randomForest*, a najviše *logistička regresija* i *cforest*.

Zaključak

Problem kojim smo se bavili u ovom radu je ocjena sličnosti u konsenzusu o prisutnosti pojave unutar jednog i između dvaju skupova binarnih klasifikatora. U radu je dana osnovna teorijska podloga vezana za odabrane metode klasifikacije i koeficijente sličnosti. Nakon toga je prikazana primjena koeficijenta sličnosti s ciljem usporedbe dvaju skupova klasifikatora. U programu R je provedena analiza dvaju skupova podataka iz područja medicine, *heart disease* i *stroke prediction*. Problem nebalansiranosti skupa podataka *stroke prediction* je riješen primjenom metoda nasumičnog poduzorkovanja i preuzorkovanja. Izgrađeni su logistički modeli s različitim prediktorima i primijenjene dvije implementacije algoritma slučajnih šuma, *randomForest* i *cforest*, pri čemu su izgrađeni klasifikatori uz odabir različitih vrijednosti hiperparametara. Prediktivne sposobnosti izgrađenih modela su ocijenjene na skupovima za testiranje. Na temelju vrijednosti mjera za evaluaciju klasifikatora zaključili smo da na skupu podataka *heart disease* svi klasifikatori imaju zadovoljavajuće performanse, dok na skupu *stroke prediction* isto ne vrijedi za *randomForest* modele. Nadalje, računanjem k-adskog Jaccardovog koeficijenta sličnosti smo ustanovili da skupovi logističkih klasifikatora imaju najveću sličnost unutar skupa klasifikatora, bez obzira na skup podataka. *cforest* i *randomForest* klasifikatori imaju nižu sličnost unutar skupa klasifikatora, pogotovo na skupu *stroke prediction*. Konačno, primjenom mo-

difikacije k-adskog Jaccardovog koeficijenta sličnosti smo ocijenili međusobnu sličnost između skupova klasifikatora. Zaključili smo da su na skupu podataka *heart disease* različiti skupovi klasifikatora približno jednako slični, ali i da je na skupu podataka *stroke prediction* njihova sličnost znatno manja. Najveća razina sličnosti je zabilježena u slučaju logističkih i *cforest* skupova klasifikatora, dok je najmanja razina sličnosti zabilježena između skupa logističkih i skupa *randomForest* klasifikatora.

Literatura

[1] David W. Hosmer, Stanley Lemeshow, *Applied Logistic Regression*, John Wiley Sons, 2000.

[2] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012.

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, 2009.

[4] Matthijs J. Warrens, *k-Adic Similarity Coefficients for Binary (Presence/Absence) Data*, Journal of Classification 26:227-245, 2009.

[5] Leo Breiman, *Random Forests*, University of California, 2001.

[6] Ana Perišić, Sophie Vanbelle, *Asymmetric 2-group k-adic similarity coefficient*, Book of Abstracts BIOSTAT 2023 - 26th Int. Scientific Symposium on Biometrics / Jazbec, A., Tafro, A., Šimić, D., Pecina, M., Vedriš, M., Sović, S., Brajković, V., Sonicki D., (ur.)., Zagreb, Croatian Biometric Society, 2023.

[7] Ana Perišić, Sophie Vanbelle, *2-group k-adic similarity indices*, Članak u nastajanju, 2023.

Literatura

[8] Ana Perišić, Dubravka Šišak Jung, Marko Pahor, *Churn in the mobile gaming field: Establishing churn definitions and measuring classification similarities*, Expert Systems with Applications 191 (2022): 116277.

[9] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, *Heart Disease*, UCI Machine Learning Repository, 1988., `https://doi.org/10.24432/C52P4X`

[10] Fedesoriano, *Stroke Prediction Dataset*, 2021., `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`

[11] Madhur Modi, *Different random forest packages in R*, 2016., `https://www.linkedin.com/pulse/different-random-forest-packages-r-madhur-modi`

[12] `https://www.geeksforgeeks.org/decision-tree/`