Primjena strojnog učenja u ocjenjivanju kakvoće mora za kupanje

Džal, Daniela

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

DANIELA DŽAL

PRIMJENA STROJNOG UČENJA U OCJENJIVANJU KAKVOĆE MORA ZA KUPANJE

DIPLOMSKI RAD

Split, siječanj 2021.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

PRIMJENA STROJNOG UČENJA U OCJENJIVANJU KAKVOĆE MORA ZA KUPANJE

DIPLOMSKI RAD

Neposredna voditeljica:
dr. sc. Ivana Nižetić Kosović, znanstveni suradnik

Studentica:
Daniela Džal

Mentor:
doc. dr. sc. Ivo Ugrina

Split, siječanj 2021.

Ovaj diplomski rad dijelom je nastao u sklopu rada na istraživačkom projektu odjela ETK Research, kompanije Ericsson Nikola Tesla d.d.

Zahvaljujem dr.sc. Ivani Nižetić Kosović na podršci i stručnom vodstvu pri izradi ovog rada.

Također, hvala Zavodu za javno zdravstvo na podacima za istraživanje i dr. sc. Slavenu Joziću sa Instituta za oceanografiju i ribarstvo u Splitu koji nam je stavio svoje stručno znanje na raspolaganje.

Uvod

S obzirom na velik broj kupača na plažama, praćenje i održavanje kakvoće mora za kupanje važan je javnozdravstveni čimbenik. Uredbom Vlade RH propisan je način, učestalost i prostorna raspoređenost praćenja kakvoće mora uzorkovanjem. Ovaj je rad nastao iz nastojanja da se metodama strojnog učenja nadopune biološke metode kako bi se dobio pouzdan i učinkovit sustav za obavještavanje kupača o kakvoći mora za kupanje koji bi radio kontinuirano i u stvarnom vremenu.

U prvom poglavlju detaljnije će se objasniti priroda problema. Opisat će se osnove strojnog učenja kao alata za rješenje problema, a onda i detaljno definirati željeni oblik rješenja.
Premda je nemoguće iz loših podataka dobiti dobar model, priprema podataka često je podcijenjen korak u procesu strojnog učenja. Drugo poglavlje posvećeno je opisivanju svih poduzetih koraka kako bi se dobro pripremili podaci za daljnje istraživanje.
Konačno, treće poglavlje opisat će korake gradnje i odabira modela te naravno predstaviti i analizirati dobivene rezultate.

Sadržaj

Poglavlje 1

Opis problema

1.1 Upravljanje kakvoćom mora

Upravljanje kakvoćom mora za kupanje uređeno je Direktivom Europskog parlamenta i vijeća [1] te Uredbom Vlade Republike Hrvatske [2]. Uzorkuje se morska voda na plažama i to svakih 15 dana u sezoni kupanja, a promatraju se mikrobiološki pokazatelji onečišćenja - broj izraslih kolonija bakterija: escherichia coli i crijevni enterokok. Uzorci se uzimaju na udaljenosti od najmanje 1 metar od obalne linije, u moru dubine najmanje 1 metar, 30 centimetara ispod površine.

Definirano je da sezona kupanja traje od 1. lipnja do 15. rujna, a praćenje kakvoće mora obavlja se od 15. svibnja do 30. rujna. Kalendar uzorkovanja definira se unaprijed prije početka svake sezone kupanja. Uzorci mora ne uzimaju se za vrijeme jake kiše, jakog vjetra, velikih valova ili pojave proliferacije makroalgi/fitoplanktona, a propušteno uzorkovanje nadoknađuje se čim prestanu nepovoljni uvjeti.

Pokazatelj	Izvrsna	Dobra	Zadovoljavajuća	Nezadovoljavajuća
crijevni enterokok	< 60	$61 - 100$	$101 - 200$	> 200
escherichia coli	< 100	$101 - 200$	$201 - 300$	> 300

Tablica 1.1: Klasifikacija morske vode

Prema količini bakterija e. coli i crijevni enterokok u uzorku, uzorci se ocjenjuju kao u tablici 1.1. Ako bakterije pokazuju različitu kakvoću mora za kupanje, za opću ocjenu tog uzorka uzima se lošija.

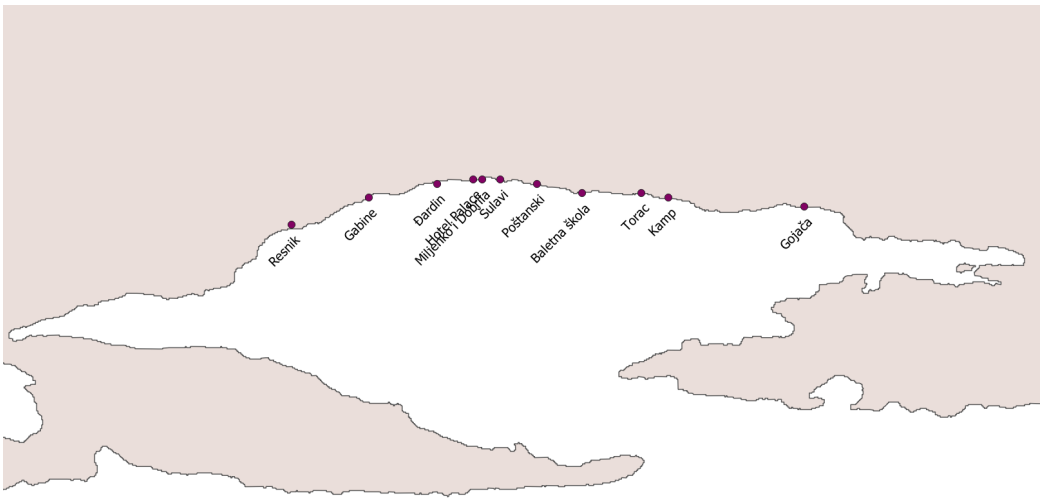
Kada bilo koji mikrobiološki pokazatelj dobiven uzorkovanjem prelazi graničnu vrijednost za ocjenu zadovoljavajuće smatra se da je došlo do kratkotrajnog onečišćenja. U tom slučaju uzorkovanje se na spornoj plaži ponavlja svakodnevno do prestanka onečišćenja te još jednom tjedan dana nakon zadnjeg zabilježenog onečišćenja.

Osim ovakve, pojedinačne ocjene, na temelju svakog uzorkovanja, stvaraju se dugoročnije ocjene plaža za kupanje. Godišnja ocjena računa se na temelju svih uzoraka u jednoj sezoni, a konačna na temelju prethodnih 5 sezona.

1.2 Prikupljeni podaci

Za potrebe ovog rada promatrat će se povijesni podaci o uzorcima uzetima na 11 plaža u Kaštelanskom zaljevu. Promatrane plaže su:

- Gojača - Kaštel Sućurac
- Kamp - Kaštel Gomilica
- Torac - Kaštel Gomilica
- Baletna škola - Kaštel Kambelovac



Slika 1.1: Mapa promatranih plaža

- Poštanski - Kaštel Lukšić
- Šulavi - Kaštel Lukšić
- Miljenko i Dobrila - Kaštel Lukšić
- Hotel Palace - Kaštel Stari
- Đardin - Kaštel Stari
- Gabine - Kaštel Štafilić
- Resnik - Kaštel Štafilić

Korišteni podaci prikupljani su u sezonama kupanja između 2015. i 2019. godine, a promatrane plaže prikazane su na slici 1.1.

Podaci su prikupljani iz različitih izvora čiji je kratki pregled dan u tablici 1.2.

Izvor	**Parametri**	**Izvor**	**Parametri**
IZOR	lokacija	DHMZ	temperatura
	vrijeme		vlažnost zraka
	e.coli		tlak zraka
	crijevni enterokok		tendencija tlaka
	slanost		relativni tlak na površini mora
	temperatura mora		brzina vjetra
	vjetar		smjer vjetra
	kiša		pokrivenost oblacima
	opis vremena		količina kiše u posljednja 24 sata
	najviša razina mora	NASA	termalno infracrveno zračenje
	najniža razina mora		osunčanost vrha atmosfere
TZ	broj turističkih noćenja		osunčanost horizontalne površine
			bistrina osunčanosti

Tablica 1.2: Pregled parametara i izvora iz kojih su prikupljeni

1.2.1 IZOR

Institut za oceanografiju i ribarstvo u Splitu upravlja bazom podataka o kakvoći mora za kupanje. Podaci su prikupljeni od strane Zavoda za javno zdravstvo. Osim mikrobioloških pokazatelja kakvoće mora, pri uzorkovanju se bilježi još slanost i temperatura mora te kategorički parametri koji opisuju meteorološke uvjete.
Parametar `vjetardane` poprima vrijednosti odsutan/prisutan, a u slučaju prisustva vjetra, bilježeni su još parametri `vjetar_jačina` i `vjetar_smjer` - također kategorički. Jačina poprima vrijednosti umjeren/jak, a smjer strane svijeta.
Parametar `vrijeme_opis` poprima vrijednosti sunčano/poluoblačno/oblačno, a parametri `kiša_dan_prije` i `kiša_dan_uzorkovanja` imaju vrijednosti prisutna/odsutna.
Ukupno nam je ustupljeno 612 uzoraka s već navedenih plaža u spomenutom periodu.

Na web lokaciji IZOR[1] su javnodostupni podaci bilježenih razina mora u Splitu. Preuzete su vrijednosti parsirane i izdvojene su dnevne najviše i najniže razine mora za svaki dan uzorkovanja.
Korištenjem navedenih parametara, svakom uzorku dodani su neki novi, agregirani parametri. Za svaki je uzorak izračunata aritmetička sredina količine bakterija na istoj plaži u prethodnoj sezoni kupanja, a osim tog parametra, dodane su još i količine bakterija u prošlom uzorku na istoj plaži.

Kaštelanski zaljev, nažalost još nema kvalitetno i potpuno provedenu kanalizaciju i pročišćivač te diljem zaljeva postoje "divlji" ispusti iz kućanstava.

[1] https://acta.izor.hr/wp/mjerni-sustavi-u-stvarnom-vremenu/visoke-i-niske-vode/

Radi se o takozvanim potocima koji skupljaju slivne vode, ali i otpadne vode iz kućanstava i ulijevaju se u more. Ukazano nam je na lokacije nekih takvih izvora onečišćenja u Kaštelanskom zaljevu. Izračunata je matrica svih udaljenosti između plaža i izvora onečišćenja, a zatim je svakom uzorku pridružena udaljenost do najbližeg izvora onečišćenja kao novi parametar.

Aritmetička sredina količine bakterija na nekoj plaži i udaljenost do najbližeg izvora onečišćenja na neki način stvaraju sliku o plaži, odnosno govore ima li neka plaža sklonost prema onečišćenju.

1.2.2 DHMZ

Drugi je važan izvor informacija Državni hidrometeorološki zavod iz čije su arhive prikupljeni parametri: temperatura zraka, vlažnost, tlak zraka, tendencija tlaka, relativni tlak na površini mora, količina kiše, smjer i brzina vjetra te prekrivenost oblacima. Ovi su podaci prikupljeni kao proširenje i poboljšanje opisnih parametara koje prate biolozi s Instituta. Osim što su mjereni, a ne uočeni, ovi su podaci numerički. Zbog toga su precizniji i upotrebljiviji u smislu strojnog učenja.

Mjerenja se, ovisno o meteorološkoj postaji, bilježe u različitim vremenskim intervalima pa je potrebno na temelju dostupnih podataka procijeniti meteorološke uvjete u trenutku uzorkovanja na lokacijama gdje su uzorci uzeti. Tražene vrijednosti potrebno je interpolirati prostorno vremenski. Interpolacija je obavljena metodom koja izražava traženu vrijednost u točki kao ponderiranu sumu dostupnih vrijednosti u najbližim susjednim točkama. Svi navedeni parametri interpolirani su za trenutak u kojem je uzorak uzet i točno 24 sata ranije.

1.2.3 NASA

Podaci vezani uz sunčevu radijaciju preuzeti su s web lokacije Power data access viewer[2]. Ovaj sustav namijenjen je istraživanju obnovljivih izvora energije diljem svijeta u svrhu njihovog što boljeg iskorištavanja. Na temelju satelitskih snimki, dostupne su dnevne srednje vrijednosti meteoroloških parametara, kao i parametara vezanih uz sunčevo zračenje.
Za sve datume uzorkovanja, preuzeti su podaci na istoj lokaciji - na sredini Kaštelanskog zaljeva. Iskorištena su četiri parametra: osunčanost vrha atmosfere, osunčanost horizontalne površine, termalno infracrveno zračenje i bistrina osunčanosti. Motivacija za dodavanje ovih parametara, također je proizašla iz pojava koje promatraju biolozi s Instituta za oceanografiju. Već je navedeno da oni pri uzorkovanju bilježe je li vrijeme oblačno, polublačno ili sunčano, a to rade jer postoje naznake veze između količine bakterija i sunčevog zračenja. Ovdje je kategoričko opažanje nadopunjeno numeričkim podacima.

1.2.4 Turistička zajednica

Iz Turističke zajednice grada Kaštela dobiveni su podaci o broju turističkih noćenja u Kaštelima za datume u 2016. godini za koje imamo uzorke. Od tih podataka napravljen je splajn koji je korišten za aproksimaciju broja noćenja za ostale datume uzorkovanja.
Primijenjen je Forsythe, Malcolm i Molerov kubični splajn iz paketa `stats`.

Ovaj je podatak gruba aproksimacija stvarnog broja turističkih noćenja koja leži na pretpostavci da je vremenska distribucija broja noćenja svake godine ista. Slično kao što udaljenost od poznatih izvora onečišćenja ukazuje

[2]https://power.larc.nasa.gov/data-access-viewer/

na potencijalno rizičnu plažu, ovaj bi podatak mogao ukazati na potencijalno rizičan datum u godini.

1.3 Strojno učenje

Neka je X uzorak u nekom skupu podataka opisan sa p nezavisnih varijabli $X = [X_1, \ldots, X_p]$ i neka je Y odgovarajući zavisni parametar. Pretpostavimo da postoji veza između Y i X koja može biti zapisana matričnom jednadžbom

$$Y = f(X) + \epsilon.$$

Funkcija f je neka fiksna, ali nepoznata funkcija, a ϵ je greška.
Strojno učenje je grana umjetne inteligencije koja uključuje metode i algoritme za automatsko stvaranje ovakvih funkcija uz minimiziranje greške. Ove funkcije još nazivamo modelima, a pod automatskim stvaranjem podrazumijevamo mogućnost algoritma da na temelju podataka (iskustva) generira traženu funkciju uz minimalnu grešku. Za razliku od sustava baziranih na pravilima koji izvršavaju zadatak praćenjem naredbi i koji će svaki put izvršiti zadatak na isti način, performanse sustava baziranog na strojnom učenju mogu se poboljšavati kroz treniranje, odnosno izlaganjem sustava novim podacima. Tada kažemo da sustav uči.

Ustaljena je podjela strojnog učenja na nadzirano i nenadzirano. Kada govorimo o nenadziranom učenju, podrazumijevamo da algoritam prima vektore vrijednosti nezavisnih varijabli X za svaki uzorak no nema uvid u vrijednost izlazne varijable Y. U ovakvom tipu strojnog učenja traže se veze između nezavisnih varijabli ili između uzoraka. Na traženju veza između nezavisnih varijabli zasnivaju se tehnike smanjenja dimenzionalnosti. Uočavanjem veze među nekim varijablama, njih više može se zamijeniti jednom koja sadrži

približno istu informaciju što može biti jako korisno za daljnje učenje iz tih podataka. Traženjem veza među uzorcima, skup podataka može se grupirati - klasterirati. Više uzoraka povezujemo u klastere gdje na temelju sličnosti nezavisnih varijabli očekujemo i sličnost zavisne varijable.

S druge strane kod nadziranog učenja algoritam za svaki uzorak prima vektor vrijednosti nezavisnih varijabli X i odgovarajuću vrijednost izlazne varijable Y te mapira vezu među njima s ciljem predviđanja nezavisne varijable na budućim uzorcima ili boljeg razumijevanja veze između nezavisnih i zavisne varijable. Ovisno je li izlazna varijabla numeričkog ili kategoričkog tipa, govorimo o regresiji ili klasifikaciji (binarnoj ili višestrukoj ovisno o broju kategorija zavisne varijable). Većina algoritama nadziranog strojnog učenja mogu se prilagoditi da rade i klasificiranje i regresiju.

Bitna razlika između klasifikacije i regresije leži upravo u načinu mjerenja greške modela. Kako ocijeniti koliko je dobro neki model "naučio" iz podataka? Kod regresije je bitno koliko su predviđene vrijednosti udaljene od stvarnih. Obično se gleda srednja vrijednost kvadratnog odstupanja predviđanja od greške ili srednja vrijednost apsolutnih vrijednosti tog odstupanja. Bilo kvadriranjem, apsolutnom vrijednosti ili nekom trećom funkcijom, treba na neki način onemogućiti da se greške pozitivnog i negativnog predznaka međusobno ponište. S druge strane, kod klasifikacije je važno samo koliko je točno, a koliko pogrešno klasificiranih uzoraka. Ovisno o podacima, pogodno je gledati broj dobro klasificiranih u odnosu na sve ili pak u odnosu na uzorke iz samo jedne klase.

Uobičajeni tijek izrade modela strojnog učenja sastoji se od sljedećih koraka:

1. **Prikupljanje podataka** prvi je i osnovni korak u strojnom učenju. Važno je prikupiti što točnije i potpunije podatke. U ovom je koraku važno i ekspertno, odnosno domensko znanje kako bi odabrane nezavisne varijable imale što čvršću i smisleniju vezu sa zavisnom.

2. **Odvajanje skupa za evaluaciju** sljedeći je važan korak kako bi mogli izraziti koliko je dobar izgrađeni model. Ovdje izdvojeni podaci ostat će netaknuti do kraja procesa, a zovu se još i holdout podaci. Izgrađeni je model potpuno neovisan o ovim podacima i oni služe isključivo za evaluaciju modela.

3. **Preprocesiranje podataka** uključuje obradu outliera, nepoznatih vrijednosti i svih nepravilnosti u podacima. Također, često je nužno normalizirati podatke radi boljih performansi treniranog modela.

4. **Treniranje i validacija modela** - skup podataka koji je preostao nakon odvajanja holdout skupa u ovom se koraku na neki način dijeli na skup za treniranje i skup za validaciju. Na skupu za treniranje model se trenira - uči, a na skupu za validaciju procjenjuje točnost koju bi izgrađeni model trebao imati na novim, nepoznatim podacima. Ovisno o odabranoj metodi validacije, ovaj se postupak može, ali ne mora ponavljati više puta, a cilj je odrediti koji je model u stanju najbolje generalizirati naučeno, odnosno primijeniti stečeno "znanje" na nove podatke.

5. **Evaluacija odabranog modela** finalni je korak u kojem doznajemo kakve su zaista performanse istreniranog odabranog modela na nepoznatim podacima. Izražavamo točnost, odnosno grešku modela brojevima, tablicama i grafovima te analiziramo u kojim je uvjetima model

povoljan i iskoristiv, a u kojim uvjetima nije pouzdan i traži daljnje poboljšanje.

1.3.1 Korišteni alati

Za obradu podataka i provedbu statističke analize i strojnog učenja korišten je programski jezik R (verzija 3.6.3. - "Holding the windsock") uz pripadajuće okruženje RStudio. To je besplatni software široko korišten u ove svrhe kompatibilan s raznim operacijskim sustavima (Windows, MacOS i Unix). Nastao je kao nasljednik programskog jezika S. Nudi širok spektar statističkih i grafičkih alata, a postoji i mnoštvo nadogradnji u vidu paketa koje svakodnevno oblikuju i održavaju korisnici prilagođavajući ih svojim potrebama.
Za potrebe ovog rada, korišteni su sljedeći paketi:

- `dyplr` - prikupljanje i parsiranje podataka, općenito baratanje podacima, filtriranje, stvaranje agregiranih parametara, izdvajanje željenih parametara...
- `ggplot2` - crtanje i uređivanje grafova
- `caret` - istraživanje raznih modela i metoda strojnog učenja
- `lubridate` - parsiranje datuma i vremena
- `geosphere` - računanje udaljenosti među geografskim lokacijama
- `ROSE` - balansiranje podataka
- `nnet` - izgradnju neuronskih mreža
- `MLmetrics` - računanje raznih ocjena performansa modela

1.4 Definicija problema

Mikrobiološka obrada uzorka traje u prosjeku 2.2 dana [3] što je predug period da bi sustav obavještavanja kupača imao smisla i uopće bio učinkovit. Nadalje, čak i u slučaju trenutačne obrade uzoraka, 15 dana je predug period bez uzorkovanja jer neka istraživanja upućuju na to da se situacija mijenja na dnevnoj, čak i satnoj bazi. Ovaj je rad rezultat pokušaja gradnje modela kojim će se omogućiti pravovremeno i redovitije obavještavanje kupača o zagađenjima.

Iako sustav izgrađen metodama strojnog učenja ima ograničenu točnost i ne može biti precizan i pouzdan kao biološko uzorkovanje i mjerenje, ipak može biolozima sugerirati kojim plažama treba posvetiti više pažnje. Time se biolozima daje mogućnost da bolje usmjere ljudske i materijalne resurse prema plažama s većim faktorima rizika.
Osim obavještavanja kupača, redovitije praćenje kakvoće i bolje uočavanje zagađenja trebalo bi biti od koristi nadležnim službama za sprječavanje zagađenja u budućnosti, odnosno otkrivanje i saniranje izvora onečišćenja.

S obzirom da raspolažemo količinom bakterija u svakom uzorku (ne samo kategorijom u koju uzorak spada) logično bi bilo koristiti regresijske metode. U tom slučaju model predviđa količinu bakterija u novom uzorku s nekom greškom ϵ.
Jedan od pristupa u ovom istraživanju bio je na osnovu toga napraviti klasifikaciju na tri kategorije - pozitivno, negativno i neodlučeno, uzimajući u obzir graničnu vrijedost za ocjenu nezadovoljavajuće t. Klasifikacija bi bila realizirana na način da uzorci čija je predviđena vrijednost manja od $t - \epsilon$ budu klasificirani kao zadovoljavajući za kupanje, uzorci čija je predviđena

vrijednost iznad $t + \epsilon$ budu klasificirani kao nezadovoljavajući, a za uzorke čija je predviđena vrijednost "blizu" graničnoj vrijednosti (predviđanje u intervalu $[t - \epsilon, t + \epsilon]$) nema odluke.
Predikcije se donose za svaku bakteriju posebno, a konačna odluka na temelju zaključaka o obje bakterije. Najbolji bi bio onaj model koji ima najmanji interval $[t - \epsilon, t + \epsilon]$.

Međutim, u slučaju kada ne postoje jake veze između nezavisnih varijabli i zavisne varijable, regresijski modeli pokušavaju umanjiti ukupnu grešku na način da se predviđanja približavaju srednjoj vrijednosti zavisne varijable. Zbog toga se gore opisani pristup pokazao beskorisnim jer većina uzoraka upadne u klasu "neodlučeno".

Drugi bi pristup bio uobičajena klasifikacija, s obzirom da je željeni zaključak upravo kategorija kakvoće vode za kupanje. U tom slučaju uzorci se klasificiraju u odnosu na svaku bakteriju, a konačna je odluka opet temeljena na zaključku o obje bakterije.

U ovom radu neće se promatrati klasifikacija na četiri kategorije već će se izvrsno, dobro i zadovoljavajuće promatrati kao jedna, a nezadovoljavajuće kao druga kategorija. Dakle, radi se o binarnoj klasifikaciji koja ima smisla u primjeni kada je bitno razlučiti je li se sigurno kupati.

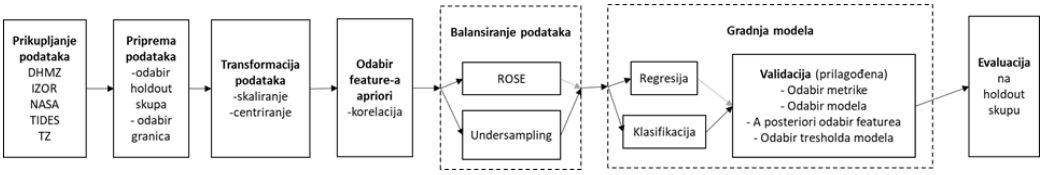
Jasno je da možemo programirati strojno učenje da predviđa postojanje onečišćenja koje je uvjetovano barem jednom od spomenutih bakterija - jedan model detektira onečišćenje u uzorku, nebitno kojom bakterijom. Ipak, s obzirom da su bakterije različite i različito reagiraju na okolišne čimbenike, u ovom se radu predlaže da se grade odvojeni modeli za svaku bakteriju. U tom se slučaju odabiru parametri za gradnju modela strojnog učenja koji su

relevantni za svaku bakteriju posebno.

Osim toga, predviđanje onečišćenja uzrokovanog bilo kojom bakterijom jednim modelom predstavlja samo jedno dodatno kodiranje i otegotnu okolnost algoritmu strojnog učenja.

Dakle, u ovom radu predstavit će se binarna klasifikacija kakvoće mora za kupanje. Bit će izgrađena dva neovisna modela strojnog učenja koji će klasificirati svaki na osnovu svoje bakterije i odgovarajuće granice za ocjenu zadovoljavajuće koje su ranije navedene, a njihovi rezultati u konačnici spojeni booleovom operacijom *ili*. Dakle, uzorak za koji barem jedan model predvidi da je onečišćen, bit će ocijenjen ocjenom nezadovoljavajuće. Odvojit će se uzorci koji su nezadovoljavajući za kupanje od svih ostalih. U daljnjem tekstu reći ćemo da je uzorak **pozitivan** ako ima ocjenu nezadovoljavajuće, a **negativan** ako ima ocjenu zadovoljavajuće, dobro ili izvrsno.
Cilj je pogoditi što više onečišćenja, naravno uz što manji broj lažnih uzbuna.

1.5 Sažetak procesa



Slika 1.2: Dijagram procesa izrade modela strojnog učenja

Za dobivanje učinkovitog i pouzdanog modela strojnog učenja ključno je pažljivo proći kroz cijeli proces od prikupljanja podataka do evaluacije modela, a pritom ranije opisane korake prilagoditi konkretnom problemu koji se rješava.

U ovom radu, prvo su podaci prikupljeni iz različitih izvora i spojeni po vremenu i mjestu u jedinstvenu tablicu. Sljedeći je korak odvojiti podatke za holdout skup. To su podaci koji neće biti korišteni do samog kraja odnosno evaluacije izgrađenog modela. Ovi podaci ne smiju utjecati na faktore centriranja ni skaliranja, kao ni balansiranje podataka.

Sljedeći je korak centriranje i skaliranje podataka. S obzirom da je prikupljeno puno parametara u odnosu na broj uzoraka, potrebno je nekako selektirati relevantne parametre. Ovdje je to učinjeno *a priori* selekcijom odnosno selekcijom na temelju koeficijenta korelacije sa zavisnom varijablom.

Podaci su nebalansirani, pripadnika jedne klase daleko je više od pripadnika druge, stoga ih je potrebno balansirati za optimalne rezultate. Proučene su dvije metode. Metoda ROSE sintetički stvara nove podatke, a metoda undersample nasumično odbacuje neke uzorke iz većinske kategorije da se postigne bolja balansiranost.

Nakon što su podaci ovako pripremljeni, preostaje testirati razne modele strojnog učenja i testiranjem odabrati onog s najboljim rezultatima. Da bi to bilo moguće, potrebno je odabrati metriku ocjenjivanja. To znači unaprijed odrediti kakvi su nam rezultati povoljniji za ovaj konkretan problem.

Konačno, odabrani model potrebno je fino podesiti, odnosno prilagoditi hiperparametre, a zatim evaluirati na nepoznatim podacima.

Poglavlje 2

Priprema podataka

2.1 Odvajanje holdout podataka

Kako bismo mogli evaluirati konačni model i donijeti procjenu njegove točnosti, nužno je prije svega odvojiti jedan dio podataka. Sve transformacije podataka i gradnja modela radit će se neovisno o ovom skupu podataka.

Uobičajeno je nasumično odabrati manji dio podataka koji se ostavljaju za ovu svrhu, osim kada se radi o vremenskim nizovima. Tada se ostavljaju kronološki posljednji podaci da bi podaci za evaluaciju bili neovisni o podacima za treniranje. U slučaju predviđanja kakvoće mora podaci ne sačinjavaju vremenske nizove u klasičnom smislu jer cilj ovog rada nije, na temelju nekog niza količina bakterija, pogoditi koliko će ih biti u sljedećem uzorku. Ipak, zbog agregiranih parametara: srednje vrijednosti količine bakterija u prošloj sezoni i količine bakterija u prethodnom uzorku, podaci sadrže neku informaciju o vremenu. Zbog toga ne bi bilo potpuno korektno nasumično odabrati podatke za holdout pa su izdvojeni svi uzorci iz 2019. godine, posljednje dostupne sezone kupanja.

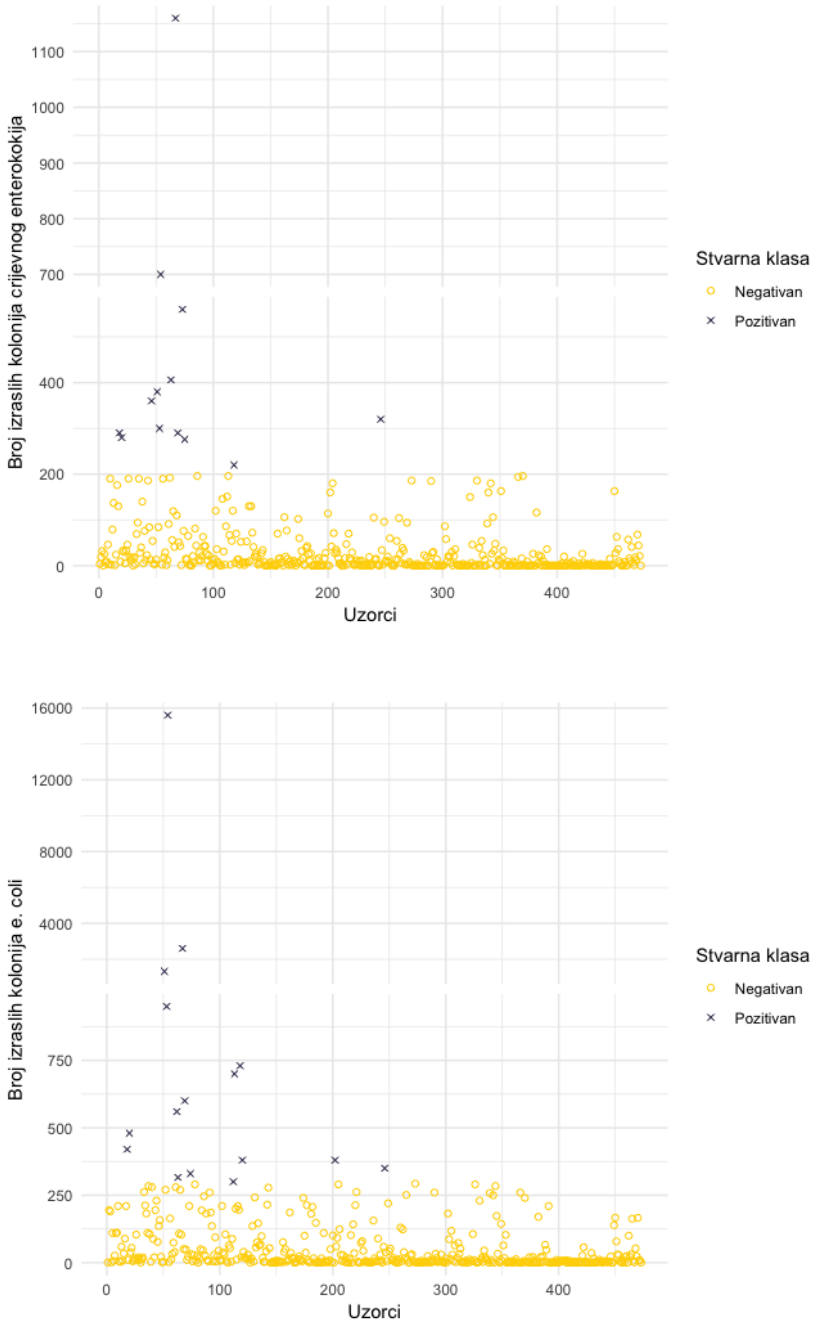
U skupu za treniranje modela preostala su 473 uzorka. Od navedenih uzoraka, 13 ih ima količinu crijevnog enterokoka koja se smatra nezadovoljavajućom za kupanje, a 15 ima prekomjernu količinu bakterije e. coli. Od navedenog, 10 se uzoraka poklapa, to jest istovremeno imaju nezadovoljavajuću količinu obiju bakterija.
U holdout skupu pojavljuje se 13 uzoraka koji imaju nezadovoljavajuću količinu e. coli, od čega dva imaju i prekomjernu količinu crijevnog enterokoka.

2.2 Deskriptivna statistika

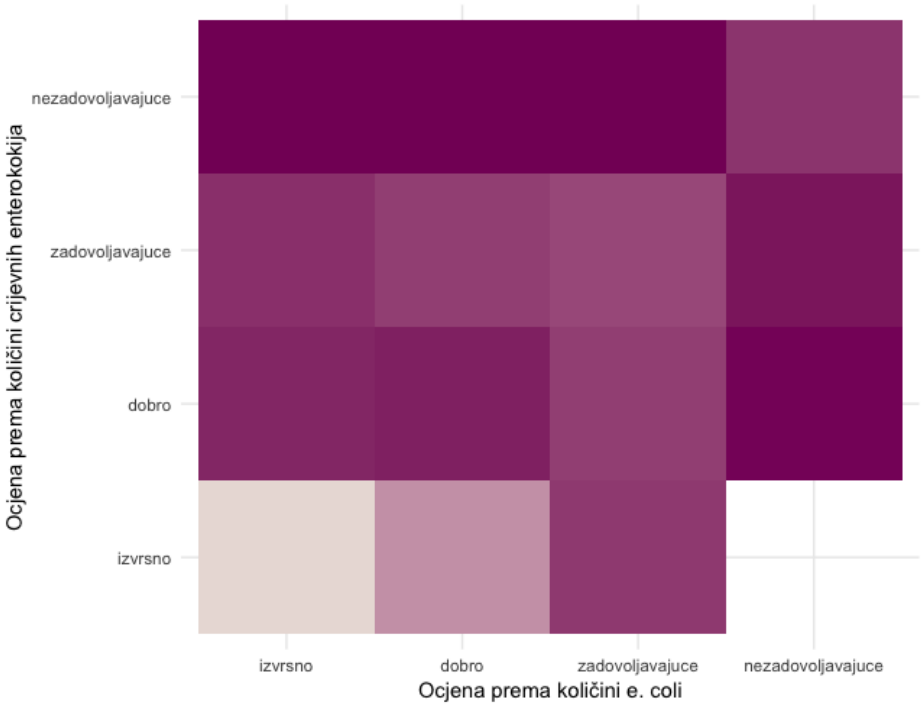
Prije samog strojnog učenja, korisno je "zaviriti" u podatke. Pogledati koliko ima pozitivnih, a koliko negativnih uzoraka, kakva je distribucija zavisnih, a kakva nezavisnih varijabli. Sve je to dio deskriptivne statistike iz koje se mogu naslutiti neki problemi i specifičnosti podataka kojima je potrebno prilagoditi proces strojnog učenja.
Iz grafova 2.1 jasno se vidi da prevladavaju negativni uzorci - oni s malom količinom mikrobioloških pokazatelja onečišćenja. Iako je veza među količinima bakterija jaka, nije rijetkost da bakterije pokažu različitu ocjenu kakvoće mora za kupanje što je i prikazano slikom 2.2.

Iz tablica 2.1 vidi se raspodjela onečišćenja i broj uzorkovanja po godinama i plažama.
Od 473 uzorka, za 441 je zabilježeno odsustvo kiše na dan uzorkovanja. Prosječna zabilježena slanost je 36.08 promila, zabilježeni minimum je 25.4, a maksimum 38.9. Temperatura mora kreće se između 18.5 °C i 28.3 °C s aritmetičkom sredinom 23.56°C.

Na slici 2.3 provjerene su veze između podataka dobivenih iz IZOR-a i DHMZ-a. Ovo bi podaci trebali sadržavati istu informaciju i veza bi trebala

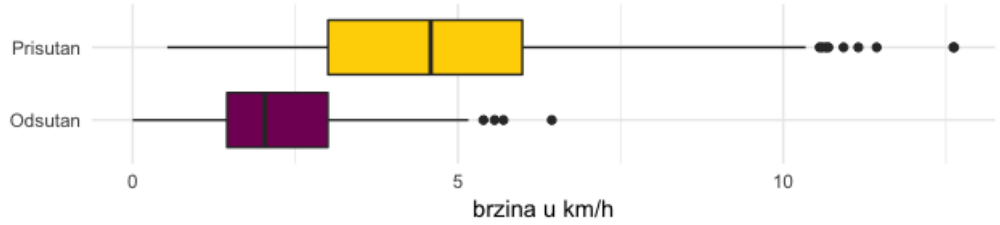


Slika 2.1: Količine bakterija u uzorcima skupa za treniranje

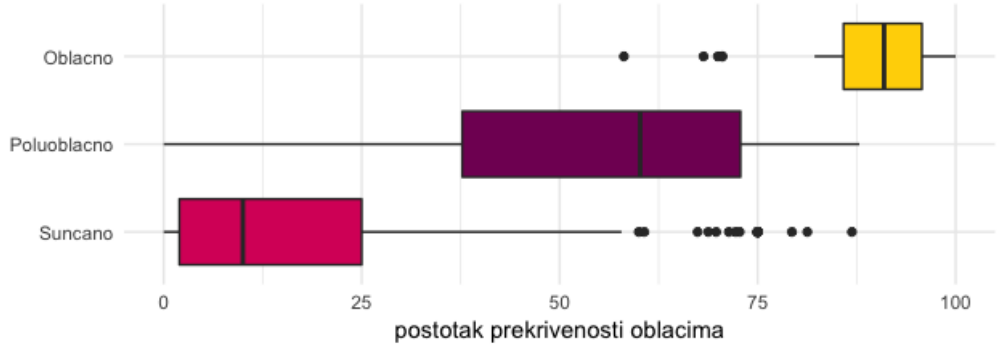


Slika 2.2: Povezanost količina bakterija: što je polje svjetlije, to je veći broj uzoraka na presjeku dvije klase.

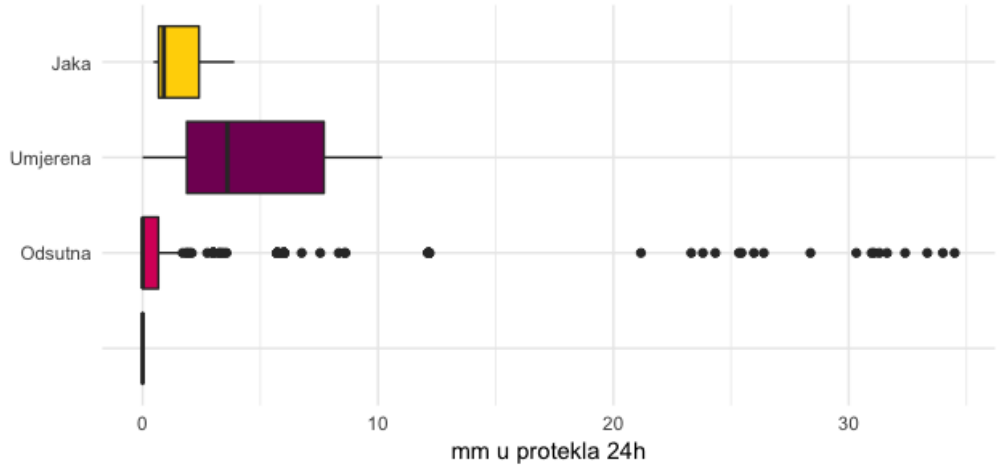
biti jaka. Kako vidimo, vjetar i prekrivenost oblacima zaista imaju očekivanu vezu, međutim za kišu numerički i opisni podaci nisu čvrsto vezani kako bismo očekivali. Uzrok tome je drukčiji pristup bilježenju količine kiše. Npr. promatranjem možemo uočiti "jaku" kišu, no ako se radi o kratkotrajnom ljetnom pljusku, možda i neće biti izmjerena velika količina vode. Također, numerički podatak izražava kumulativnu količinu kiše koja je pala u posljednja 24 sata, nešto što je u kategoričkim podacima možda zabilježeno kao jučerašnja kiša.



(a) Vjetar



(b) Prekrivenost oblacima



(c) Kiša

Slika 2.3: Veza između opisnih varijabli dobivenih iz IZOR-a i numeričkih varijabli dobivenih iz DHMZ-a

Plaža	Broj uzoraka			
	ECOLI		CE	
	P	N	P	N
Gojača	2	43	2	43
Kamp	8	43	9	42
Torac	4	45	1	48
Baletna škola	0	40	0	40
Poštanski	0	40	0	40
Šulavi	1	43	0	44
Miljenko i Dobrila	1	43	1	43
Hotel Palace	0	40	0	40
Đardin	0	40	0	40
Resnik	0	40	0	40
Gabine	0	40	0	40

Godina	Broj uzoraka			
	ECOLI		CE	
	P	N	P	N
2015	3	107	4	106
2016	13	130	9	134
2017	0	110	0	110
2018	0	110	0	110

Tablica 2.1: Tablica lijevo: broj pozitivnih i negativnih uzoraka po plažama, tablica desno: broj pozitivnih i negativnih uzoraka po godinama

2.3 Normalizacija parametara

Normalizacija parametara uobičajena je i poželjna praksa u strojnom učenju. Postoji više tipova normalizacije parametara. Da bismo ih lakše predstavili, označimo s $X = (x_1, ..., x_n)$ vektor duljine n svih vrijednosti nekog parametra.

1. Z normalizacija - naziva se još i standardizacija. Ovaj postupak uključuje centriranje, odnosno oduzimanje aritmetičke sredine parametra od svih vrijednosti, a zatim skaliranje, odnosno dijeljenje svih vrijednosti standardnom devijacijom parametra. Ovakva transformacija postavlja sred-

nju vrijednost na 0, a standardnu devijaciju svakog parametra na 1. Nove vrijednosti $\hat{x}_i$ računaju se pomoću formule:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma}$$

pri čemu je μ aritmetička sredina, a σ standardna devijacija svih vrijednosti $x_i, i = 1, ..., n$.

2. Min-max normalizacija od svake vrijednosti oduzima minimalnu, a zatim dobiveno broj dijeli s rasponom parametra. Ovaj postupak sužava raspon parametra na segment $[0, 1]$, a nove vrijednosti $\hat{x}_i$ računaju se pomoću formule:

$$\hat{x}_i = \frac{x_i - max(\{x_i, i = 1, ..., n\})}{max(\{x_i, i = 1, ..., n\}) - min(\{x_i, i = 1, ..., n\})}$$

Prednosti normalizacije podataka prije učenja su stabilnije, a samim time i brže učenje te izjednačavanje utjecaja svih parametara na izgrađeni model. Posebno je važno provesti ovaj proces kada su različiti parametri različitog reda veličine ili kad se mjere u različitim mjernim jedinicama. Na primjer, procijenjeni broj noćenja kreće se između 381 i 5302, dok se najniža zabilježena razina mora kreće između 0.01 i 0.12, zbog čega će većina algoritama dati veći značaj procijenjenom broju noćenja.

U ovom radu, na podacima je napravljena standardizacija, a centriranje i skaliranje primjenjuje se na sve nezavisne i zavisne parametre.
Potrebno je naglastiti, svi faktori izračunati na skupu za treniranje, na holdout skupu se samo primjenjuju. Holdout skup ostaje nepoznat u cijelom postupku i cijeli postupak mora bit neovisan o holdoutu kako bi mogli govoriti o korektnom strojnom učenju.

2.4 Odabir parametara

Kod strojnog učenja rizik od pretreniranja povećava se s brojem korištenih parametara. Pretreniranje je paradoks koji se događa kad model strojnog učenja "predobro" upozna podatke za treniranje i predobro im se prilagodi. Zbog toga je jako osjetljiv na varijancu zavisne varijable, odnosno, na nepoznatim podacima ima grešku koja je daleko veća od one predviđene na podacima za treniranje. Stoga je potrebno odbaciti jedan dio nezavisnih varijabli kako bi izgrađeni model bolje generalizirao.

S obzirom da se ovdje odabir parametara vrši neovisno o algoritmu strojnog učenja, radi se o a priori odabiru. Ovakav odabir vrši se na temelju koeficijenta korelacije između zavisnih i nezavisnih varijabli.

Koeficijent korelacije je statistički stupanj povezanosti dviju varijabli. Više je načina za izračun koeficijenta korelacije, a primjeri su Pearsonov i Spearmanov koeficijent. Ovdje je korišten Spearmanov koeficijent jer, za razliku od Pearsonovog, ne podrazumijeva normalnu distribuciju promatranih varijabli. Bazira se na tome da se izmjeri dosljednost povezanosti između **poredanih** varijabli. Prilikom korištenja Spearmanovog koeficijenta, vrijednosti varijabli potrebno je rangirati i na takav način svesti na zajedničku mjeru. Najjednostavniji način rangiranja je da se najmanjoj vrijednosti svake varijable pridijeli rang 1, sljedećoj po veličini rang 2 i tako sve do posljednje kojoj se pridjeljuje maksimalan rang. Izračunavanje koeficijenta radi se korištenjem vrijednosti pridijeljenih rangova. Formula za izračunavanje Spearmanovog koeficijenta korelacije je

$$r_s = 1 - \sum_{i=1}^{n} \frac{d_i^2}{n(n^2-1)}$$

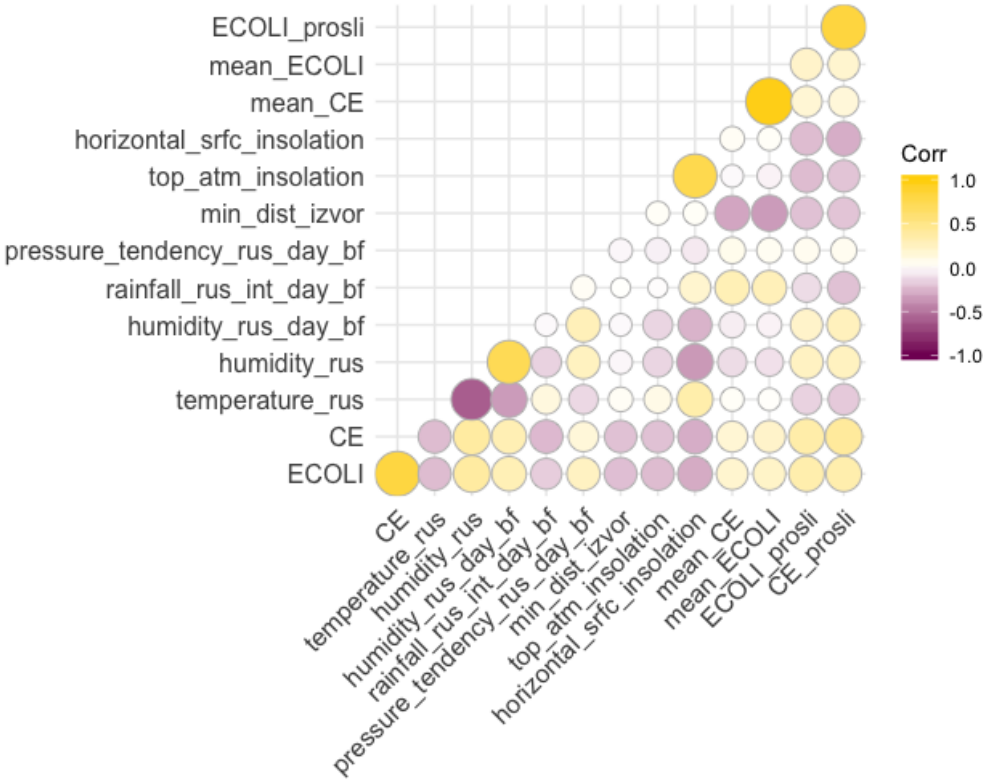
pri čemu d_i označava razliku između vrijednosti s rangom i.

E. coli		**Enterokok**	
Varijabla	Koeficijent korelacije	Varijabla	Koeficijent korelacije
temperatura zraka	0.217	temperatura zraka	0.219
vlažnost	0.368	vlažnost	0.373
vlažnost dan prije	0.258	vlažnost dan prije	0.292
kiša dan prije	0.202	kiša dan prije	0.227
tendencija tlaka dan prije	0.227	osunčanost vrha atmosfere	0.222
udaljenost do najbližeg izvora	0.214	osunčanost horiz. površine	0.253
osunčanost vrha atmosfere	0.236	srednja vrijednost e. coli	0.203
osunčanost horiz. površine	0.293	e. coli prošlog uzorka	0.303
srednja vrijednost c. e.	0.203	c.e. prošlog uzorka	0.330
srednja vrijednost e. coli	0.225		
e. coli prošlog uzorka	0.312		
c.e. prošlog uzorka	0.329		

Tablica 2.2: Parametri - a priori selekcija

Kako bi se izbjeglo pretreniranje, preporuka je da se broj parametara svede na 1 na svakih 10-20 pozitivnih slučajeva u skupu za treniranje. S obzirom na premali broj pozitivnih slučajeva, odnosno prevelik broj parametara koje bi trebalo odbaciti, za početak su izdvojene samo varijable koje s količinom bakterija imaju koeficijent korelacije barem 0.2. U tablici 2.2 su popisane te varijable.

Korelogram 2.4 prikazuje međusobnu koreliranost tih varijabli, kao i njihovu koreliranost s brojem bakterija.



Slika 2.4: Spearmanove korelacije između odabranih nezavisnih varijabli i zavisnih varijabli

2.5 Balansiranje podataka

Ranije smo već uočili da je skup podataka za treniranje nebalansiran, da je jedna klasa bitno zastupljenija od druge. U ovom slučaju, uzoraka koji imaju nezadovoljavajuću količinu escherichie coli ima tek 3.2%, a uzoraka koji imaju nezadovoljavajuću količinu crijevnog enterokokija ima 2.7%.

Većina algoritama za binarnu klasifikaciju zasniva se na pretpostavci o podjednakoj zastupljenosti klasa u skupu za treniranje stoga klasifikacija na nebalansiranom skupu predstavlja veliki izazov. Uobičajeno modeli daju loše rezultate, a posebno loše predviđaju manjinsku klasu zbog malog broja primjera na kojima bi algoritam naučio generalizirati ponašanje te klase. U praksi, manjinska klasa je obično važnija. Takvi primjeri su otkrivanje bolesti ili prijevara kreditnim karticama te otkrivanje ili predviđanje rijetkih događaja kao što je upravo pojava onečišćenja u moru za kupanje.
Glavni je problem što se algoritmi za klasifikaciju baziraju na optimizaciji cjelokupne točnosti, odnosno pokušavaju grešku klasifikacije svesti na minimum. U slučaju jake nebalansiranosti podataka, algoritmi "zanemare" manjinsku klasu. Konkretno, za klasifikaciju morskih uzoraka prema količini e. coli, gdje nezadovoljavajućih ima 3.2%, model koji sve klasificira kao zadovoljavajuće ima točnost od 96.8%. Takva točnost zvuči izvrsno iako je model očito potpuno beskoristan i neinformativan.

Prema [7] postoje tri osnovna pristupa poboljšanju predikcije za nebalansirane podatke. To su:

1. **Popravljanje na razini algoritma** prilagođava postojeće klasifikacijske algoritme da budu pristraniji manjinskoj klasi. Ovakav pristup zahtijeva dobro poznavanje klasifikatora, ali i domensko znanje odnosno

razumijevanje uzroka zbog kojih model ima loše rezultate. Nekad su podaci manjinske klase raspršeni, ili razdvojeni u manje grupe oko većinske klase što dodatno otežava klasifikaciju.

2. **Popravljanje na razini podataka** ponovnim uzorkovanjem mijenja distribuciju klasa u skupu za treniranje čime se umanjuje efekt nebalansiranosti na klasifikator. To može uključivati uzimanje povoljnog podskupa dostupnih podataka, multipliciranje nekih uzoraka ili čak generiranje sintetičkih podataka koji bi po nekom pravilu trebali pripadati određenoj kategoriji.

3. **Hibridni pristup** djeluje na razini podataka(uzorcima iz manjinske klase daje veću težinu) i na razini algoritma (modificira algoritam tako da prihvaća težine). Nedostatak ovog pristupa je potreba za definiranjem težine, odnosno cijene pojedine greške što nije informacija koja se vidi iz podataka. Obično je na istraživaču da ocijeni težinu neke greške i prilagodi je eksperimentalnim putem.

U ovoj cjelini bit će opisane tehnike popravljanja na razini podataka. Ovakav se pristup obično naziva balansiranje i možemo ga podijeliti u tri grupe. Undersampling metode eliminiraju neke uzorke iz većinske klase i time stvaraju podskup skupa za trenirajne. Oversampling metode stvaraju nadskup skupa za treniranje replicirajući podatke manjinske klase po nekom odabranom kriteriju. Hibridne metode spajaju ova dva pristupa - smanjuju broj uzoraka iz većinske klase i istovremeno povećavaju broj uzoraka iz manjinske. Isprobana su dva načina balansiranja podataka: ROSE i random undersample.

ROSE - Random Over-Sampling Examples je tehnika predložena u [10] koja generira sintetičke podatke. Za razliku od sličnih metoda koje također

generiraju sintetičke podatke ali pritom čuvaju i originalne podatke, ROSE stvara potpuno novi skup podataka koji uopće ne sadrži izvorne podatke. Funkcija `ROSE` iz paketa `ROSE` promatra odvojene prostore koje razapinju podaci iz dviju klasa i u te prostore smješta nove točke.
Funkcija kao parametre prima željeni omjer klasa u novom skupu podataka kao i željenu veličinu novoizgrađenog skupa. Ova funkcija, u ovisnosti o odabranim parametrima, može biti iz grupe koja radi oversampling, undersampling ili hibrid ova dva pristupa. U nekim se radovima čak ističe kao prednost ove metode to da se izvorni podaci mogu koristiti kao holdout podaci za evaluaciju konačnog modela.
Tijekom izrade ovog rada, isprobane su različite kombinacije ovih parametara uz različite algoritme za klasifikaciju i utvrđeno je da ova metoda balansiranja vodi u pretreniranje klasifikatora te je kao prikladnija metoda odabrana metoda random undersample.

Random undersample jednostavnija je metoda koja nasumičnim odabirom eliminira uzorke iz većinske klase. Koriste se originalni podaci i to svi iz manjinske klase i određeni dio iz većinske. Najbitnija je odluka koliko podataka odbaciti. Nije poželjno odbaciti previše podataka, pogotovo jer se odbacuju nasumično i može se lako dogoditi gubitak važnih informacija. S druge strane, potrebno je postići omjer među kategorijama s kojim možemo izgraditi funkcionalan klasifikator.
Eksperimentalnim putem - metodom pokušaja i pogrešaka, na dostupnom skupu podataka, za najbolji omjer odabrano je 1:9. Dakle, u skupu za treniranje preostaje 10% pozitivnih uzoraka i 90% negativnih.

Poglavlje 3

Gradnja i odabir modela

3.1 Metrika

Pri ocjenjivanju točnosti, prva stvar o kojoj je potrebno dobro razmisliti je metrika odnosno kriterij ocjenjivanja rezultata različitih modela. Da bismo lakše raspravljali o metrikama, upotrebljavat ćemo izraze kao u tablici 3.1.

Osobito je važan oprez kod nebalansiranih skupova za treniranje budući da upotreba standardnih metrika kao što je ukupna točnost može biti varljiva. Kako je ranije navedeno, promatranje ukupne točnosti klasifikacije može nas navesti na potpuno krive zaključke o rezultatima modela. Pritom, pod poj-

		Stvarna klasa	
		Pozitivno	Negativno
Predviđena klasa	Pozitivno	TP	LP
	Negativno	LN	TN

Tablica 3.1: TP - točno pozitivno, LP - lažno pozitivno, LN - lažno negativno, TN - točno negativno

mom ukupne točnosti podrazumijevamo omjer točno klasificiranih uzoraka i ukupnog broja uzoraka.

$$uk\_tocnost = \frac{TP + TN}{TP + LP + LN + TN}$$

Zato je potrebno razmotriti alternativne metode ocjene pogreške. U literaturi se, za upotrebu uz nebalansirane podatke, predlažu neke od sljedećih metrika:

1. **Sensitivity** (recall) je broj koji označava koliko je od svih pozitivnih uzoraka predviđeno kao pozitivno.

   $$Sensitivity = \frac{TP}{TP + LN}$$

   Drugim riječima, ova metrika izražava sposobnost modela da prepozna pozitivni slučaj, odnosno vjerojatnost modela da će pozitivan slučaj klasificirati upravo kao pozitivan.

2. **Specificity** je broj koji označava koliko je od svih negativnih uzoraka predviđeno kao negativno.

   $$Specificity = \frac{TN}{LP + TN}$$

   Model s visokim specificityjem dobro će odbaciti većinu negativnih uzoraka i neće imati velik broj lažnih uzbuna.

3. **Precision** je broj koji označava koji udio od svih pozitivno predviđenih je zapravo pozitivan.

   $$Precision = \frac{TP}{TP + LP}$$

   Za razliku od metrike sensitivity iz koje se ne vidi koliko je lažnih uzbuna, precision će biti nizak ako model često daje lažno pozitivne rezultate.

4. **Informedness** je metrika kojom se istovremeno maksimiziraju sensitivity i specificity.

$$Informedness = Sensitivity + Specificity - 1$$

Ova metrika pokušava spojiti prednosti metrika sensitivity i specificity i neutralizirati nedostatke tih metrika. Iako je moguće da u slučaju velikog broja lažno pozitivinih uzoraka sensitivity bude visok, u tom će slučaju specificity biti nizak. Obratno, ako model loše prepoznaje pozitivnu klasu, to se neće odraziti na specificity, ali hoće na sensitivity.

5. **F-score** je harmonijska sredina metrika Precision i Recall.

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

S obzirom na mali ukupan broj pozitivnih uzoraka u skupu za treniranje, čest je slučaj da klasifikator u tijeku unakrsne validacije nijedan slučaj ne predvidi kao pozitivan, zbog čega nije moguće izračunati $Precision$, a onda ni $F - score$. Zato je, za odabir i fino podešavanje modela strojnog učenja, korištena metrika $Informedness$. Ta metrika uvažava udio točno predviđenih pozitivnih među svim pozitivnima, ali i udio točno predviđenih negativnih među negativnima. Dakle, informedness se povećava s brojem točno predviđenih onečišćenja, ali smanjuje se s brojem lažnih uzbuna.

3.2 Metoda validacije

Mogućnost generalizacije nekog modela direktno utječe na rezultate predviđanja na neovisnim podacima, a procjena ovih rezultata iznimno je važna u praksi. Kada je za strojno učenje dostupno izobilje podataka, uobičajeno je podijeliti skup podataka na skup za treniranje, validaciju i testiranje. U tom

slučaju, na skupu za treniranje trenira se više modela koji se uspoređuju i fino podešavaju na skupu za validaciju, a konačno odabrani model evaluira se na skupu za testiranje, tj. holdout skupu.
Međutim, takva podjela podataka obično rezultira premalim skupom za treniranje pa je potrebno na drukčiji način procijeniti performanse modela.

Unakrsna validacija (engl. cross validation - CV) postupak je kojim se, koristeći podatke za treniranje, procjenjuje točnost koju će model imati na holdout skupu. Ovaj je postupak koristan kako za odabir algoritma, tako i za fino podešavanje hiperparametara odabranog algoritma.

Ideja unakrsne validacije je ponovljeno izdvajanje određene količine podataka iz skupa za treniranje te gradnja modela bez tih podataka kako bi ih se iskoristilo za validaciju. Ovaj se postupak ponavlja, izvlačeći u svakom navratu različite podatke da bi se točnost konačnog modela procijenila na temelju srednje vrijednosti svih ovako izgrađenih modela. Ova ideja prikazana je u pseudokodu 3.1.

Pseudokod 3.1: Ideja unakrsne validacije

```
repeat
    podijeli trainset na train i validation
    treniraj model na train
    val_tocnost = ocjena(validation)
return prosjek(val_tocnost)
```

Prema količini podataka koja se odvaja, razlikuje se nekoliko tipova unakrsne validacije. LOOCV (iz engleskog *leave one out cross validation*) podrazumijeva izdvajanje jednog uzorka pri svakoj gradnji modela. Iako greška

predviđanja izdvojenog uzorka predstavlja nepristranu procjenu za grešku testiranja jer model nije treniran na tom uzorku, ta procjena je loša jer je jako varijabilna budući da ovisi o samo jednom uzorku. Ponavljajući ovaj postupak za svaki uzorak iz skupa za treniranje dobivamo srednju vrijednost greške koja je dobar procjenitelj za grešku testiranja.

Mana ovog postupka je potreba za mnogostrukim treniranjem modela, stoga je za velike skupove za treniranje pogodnija računalno jeftinija inačica LOOCV, a to je k-fold CV. U ovom slučaju, postupak započinje nasumičnom particijom skupa za treniranje na k dijelova, takozvanih foldova. Sada se, umjesto jednog uzorka, izdvaja jedan fold na kojem se validira model izgrađen na preostalih $(k-1)$ foldova.

Budući da je cilj unakrsne validacije da što vjernije prikaže stvarne rezultate modela, potrebno je i što vjernije prikazati uvjete rada modela. S obzirom na osobitost ovdje razmatranih podataka i njihovu prostorno-vremensku komponentu, osmišljen je specijalizirani tip unakrsne validacije. Nazovimo je LODOCV - *Leave One Day Out Cross Validation*.

Promatrane plaže prostorno su bliske, a prikupljeni podaci uglavnom interpolirani po geografskoj komponenti te ne možemo zapravo tvrditi da su parametri koji opisuju uzorke uzete istog dana na različitim lokacijama posve neovisni. Interpoliraju se iz bliskih prostorno-vremenskih mjerenih podataka. Stoga je za očekivati da model može, na temelju poznavanja situacije na jednoj plaži, lakše predvidjeti situaciju na prostorno bliskoj plaži. Ovakav scenarij je u upotrebi ovog modela malo vjerojatan pa je bitno promatrati njegove performanse kad su mu nepoznati rezultati mikrobioloških pokazatelja na drugim plažama u istom danu. Zato je u ovoj prilagođenoj unakrsnoj

validaciji LODOCV podjela podataka na foldove napravljena na način da je pri svakom ponavljanju, odnosno validaciji, izbačen skup uzoraka koji su uzeti istog dana.
Na ovaj je način skup za treniranje podijeljen na 47 foldova.

3.3 Pregled klasifikacijskih algoritama

Iako je ranije navedeno kako se većina algoritama nadziranog strojnog učenja može upotrebljavati i za regresiju i za klasifikaciju, ipak postoje neki koji su pogodniji za klasifikaciju od drugih.
U literaturi je moguće pronaći velike količine različitih algoritama. Neki od najpoznatijih, isprobani su ovdje za predviđanje kakvoće mora za kupanje. To su random forest, support vector machines (SVM), linear discriminant analisys (LDA) i feed forward neuronske mreže.

Random forest algoritam je koji koristi stabla odluke kao gradivne blokove za dobivanje boljeg i stabilnijeg modela. Ideja ovog algoritma je istrenirati veliki broj stabala odluke te uprosječivanjem njihovih predviđanja poboljšati točnost modela i povećati sposobnost generalizacije. Pri gradnji stabala, pazi se da ona budu što manje korelirana, odnosno što različitija. To se čini na način da se svakom stablu, pri svakoj podjeli, na korištenje daje samo određen podskup od svih nezavisnih varijabli. Na ovaj način eliminira se mogućnost da sva stabla imaju podjele po istim varijablama. Nedostatak ovog algoritma je kompleksnost i sporost pri izgradnji modela u odnosu na neke druge algoritme.

Support vector machine je generalizacija jednostavnog klasifikatora - max margin classifier. To je klasifikator koji dijeli prostor uzoraka najširom mogućom

prugom takvom da su na svakoj strani pruge uzorci iz druge kategorije. Ovakav klasifikator može se koristiti samo za binarnu, a ne i višestruku klasifikaciju te može među kategorijama postavljati samo linearne granice. SVM je proširenje ove ideje koje može, upotrebom jezgri, postavljati granice proizvoljnih oblika. Vektorski prostor kojeg razapinju parametri ulaznih podataka proširuje se funkcijom f tih parametara na još jednu dimenziju, a zatim se novodobiveni, $(p+1)-$dimenzionalni vektorski prostor dijeli hiperravninom, tj. hiperprugom.

Linear discriminant analisys statistička je metoda zasnovana na Bayesovoj formuli $P(H_i|A) = \frac{P(H_i)P(A|H_i}{P(A)}$. Modelira se distribucija svake nezavisne varijable X odvojeno unutar svake klase zavisne varijable, dakle određuje se $P(X = x|Y = y)$, a zatim se koristi Bayesova formula da se odredi $P(Y = y|X = x)$. Dakle, ova metoda svaki uzorak smješta u onu kategoriju y zavisne varijable kojoj pripada s najvećom vjerojatnošću.

Neuronske mreže su zapravo brzorastuća grupa algoritama. Sve se zasnivaju na sličnoj ideji: grade linearne kombinacije nezavisnih varijabli kao nove, izvedene varijable, a zatim modeliraju nelinearnu funkciju koja povezuje ove varijable sa zavisnom varijablom. Spomenuta nelinearna funkcija je ono što stvara razliku među tipovima neuronskih mreža. U ovom radu iskorištena je feed forward neuronska mreža koja se pokazala dominantnom nad ostalim dosad spomenutim klasifikatorima.

3.4 Odabrani algoritam

U sljedećim cjelinama, detaljnije će se opisati neuronske mreže, poglavito algoritam `nnet` budući da su tim algoritmom dobiveni najbolji rezultati tije-

kom ovog istraživanja.
Termin neuronska mreža zapravo pokriva širok spektar različitih algoritama i metoda strojnog učenja. Ovdje će se koristiti i opisati jednostavna *feed forward* neuronska mreža s jednim skrivenim slojem koja se može prilagoditi da radi i regresiju i klasifikaciju.

Neka je $X = [X_1 \dots X_p]$ uzorak. Za K-klasnu klasifikaciju, neuronska mreža ima K izlaznih jedinica od kojih svaka izražava vjerojatnost da ulaz X pripada klasi k. Za $k = 1$ zapravo se vrši regresija. Izvedeni parametri Z_m dobivaju se iz linearnih kombinacija nezavisnih parametara, a zatim se izlazi $Y_k = f_k(X)$ dobivaju kao funkcije linearnih kombinacija od Z_m.

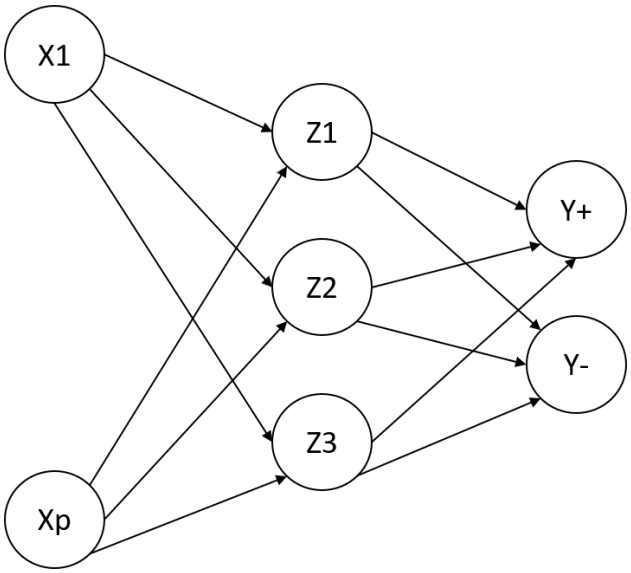
$$\begin{aligned} Z_m &= \sigma(\alpha_{m0} + \alpha_m X^T), m = 1, ..., M \\ T_k &= \beta_{k0} + \beta_k Z^T, k = 1, ..., K \\ f_k(X) &= g_k(T), k = 1, ...K \end{aligned}$$

Ili u raspisanom obliku iz kojeg je jasnije o kojim se linearnim kombinacijama radi:

$$Z_m = \sigma(\alpha_{m0} + [\alpha_{m1}...\alpha_{mp}] \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}) = \sigma(\alpha_{m0} + \alpha_{m1}X_1 + \ldots + \alpha_{mp}X_p)$$

$$T_k = \beta_{k0} + [\beta_{k1}, ..., \beta_{km}] \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix} = \beta_{k0} + \beta_{k1}Z_1 + \cdots + \beta_{km}Z_m$$

Korak u kojem se računaju Z_m nazivamo skriveni sloj jer te vrijednosti ne promatramo direktno. Općenito, ovakvih slojeva može biti više, a u svakom sloju može biti proizvoljan broj vrijednosti Z_m koje još nazivamo *neuronima*.



Slika 3.1: Prikaz arhitekture neuronske mreže

Dodavanjem slojeva i njihovim različitim povezivanjem dobivaju se kompleksnije arhitekture neuronskih mreža. Kod feed forward neuronskih mreža, svi neuroni ovise isključivo o neuronima u prethodnim slojevima, odatle i naziv feed forward jer sve informacije idu "prema naprijed". Na slici 3.1 prikazana je arhitektura neuronskih mreža kakve su trenirane u ovom radu. Sadrže jedan skriveni sloj sa tri neurona, a u izlaznom sloju sadrže dva neurona koja su označena sa $Y+$ i $Y-$ pri čemu $Y+$ predstavlja predviđenu vjerojatnost da ulaz pripada klasi P (onečišćeni uzorak morske vode), a $Y-$ predviđenu vjerojatnost da uzorka pripada klasi N.

Funkciju σ nazivamo aktivacijska funkcija i u ovom radu upotrijebljena je sigmoidna funkcija $\sigma(v) = 1/(1 + e^{-v})$ koja ima vrijednosti između 0 i 1. Funkcije $g_k(T)$ služe za završnu transformaciju vektora T i obično se izabire

identiteta ili softmax funkcija

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}.$$

Skup svih nepoznatih parametara neuronskih mreža

$$\{\alpha_{m0}, \alpha_m,\ m = 1, ..., M\}$$
$$\{\beta_{k0}, \beta_k,\ k = 1, ..., K\}$$

označit ćemo s θ, a parametre ćemo nazivati težinama.

Treniranje neuronske mreže u stvari je proces traženja težina koje najbolje opisuju podatke. Kao mjera prilagođenosti, odnosno funkcija greške koju minimiziramo upotrijebljena je formula za entropiju:

$$R(\theta) = -\sum_{i=1}^{n}\sum_{k=1}^{k} y_{ik} log f_k(x_i),$$

i optimizirana BFGS metodom.
Za klasifikaciju, algoritam `nnet` vraća vrijednost $Y+$, odnosno ako je ona veća od granice odluke, ulaz klasificira kao pozitivan, a inače kao negativan. Zadana granica odluke je 0.5.

Postupak minimizacije greške $R(\theta)$ vrši se algoritmom *back-propagation*, a sama optimizacija izvršena je BFGS metodom, što je kvazinjutnovska metoda koja koristi funkcijske vrijednosti i gradijente za pronalazak minimuma.

3.5 A posteriori odabir parametara

Kako bi se dodatno smanjio broj nezavisnih parametara, time i rizik od pretreniranja modela, napravljena je a posteriori analiza. To znači promatranje

značaja pojedinih parametara u već izgrađenom modelu za što je korištena metoda `importance` iz paketa `caret`. Ova metoda uzima niz podskupova svih parametara i za svaki takav podskup gradi model. Za izgrađeni model konstruira i računa površinu ispod ROC krivulje, a tu površinu interpretira kao mjeru značaja korištenih parametara u svakoj iteraciji. Sve mjere značaja skalirane su na vrijednosti između 0 i 100.

ROC krivulja je graf koji prikazuje performanse klasifikatora za sve granice odluke. Na x-osi su vrijednosti $1 - specificity$, a na y-osi $sensitivity$ koje klasifikator ima za sve granice odluke iz segmenta $[0, 1]$. ROC krivulja nasumičnog klasifikatora je dijagonala, odnosno dužina $y = x$ na segmentu $[0, 1]$, a idealni klasifikator ima krivulju koja prolazi kroz točku $(0, 1)$. Za usporedbu više klasifikatora, često se koristi mjera AUC - area under curve. To je površina ispod ROC krivulje i nasumični klasifikator ima AUC= 0.5, dok idealni klasifikator ima AUC= 1.

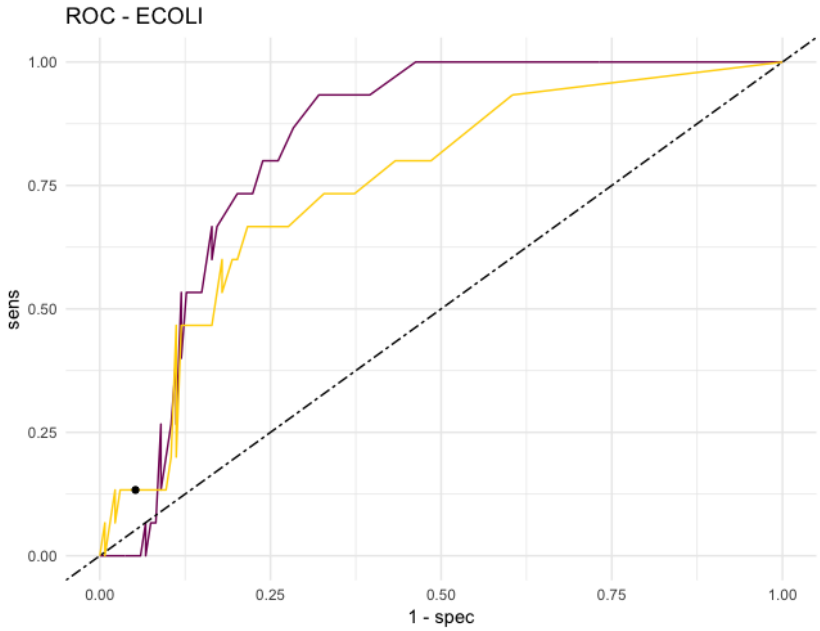
Izračunat je značaj svih parametara koji su odabrani a priori selekcijom za obje bakterije, a u ovom koraku odbacit će se svi parametri sa značajem manjim od 30 kao što je prikazano u tablici 3.2. Ta je granica odabrana eksperimentalnim putem. Odbacivanjem više varijabli dobiva se manje informativan model, a odbacivanjem manje varijabli, model je pretreniran i također ima manji AUC.

Na slici 3.2 prikazane su ROC krivulje dobivene LODOCV postupkom na modelima koji su građeni sa parametrima odabranim a priori selekcijom, te manjim skupom parametara koji su odabrani a posteriori selekcijom. Za obje bakterije, dobro se vidi da modeli s manje parametara imaju bolju ROC krivulju.

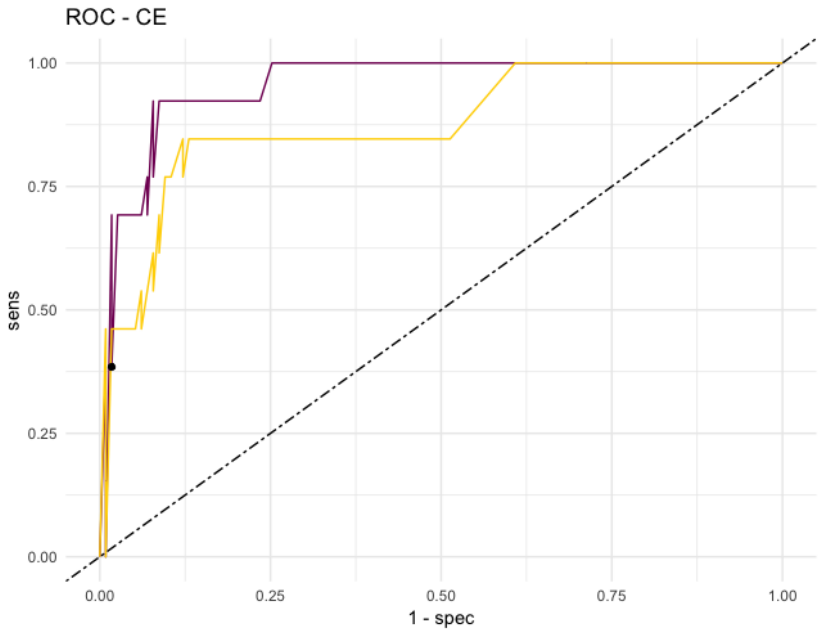
ECOLI	
Parametar	Značaj
osunčanost vrha atmosfere	100
osunčanost horiz. površine	88
vlažnost zraka	49
udaljenost najbližeg onečišćenja	38
vlažnost dan prije	35
temperatura	32
prošli e. coli	28
prošli c.e.	14
kiša dan prije	12
prosjek c.e.	8
tendencija tlaka dan prije	5
prosjek e. coli	0

CE	
Parametar	Značaj
prosjek e. coli	100
osunčanost vrha atmosfere	98
osunčanost horiz. površine	83
vlažnost dan prije	61
prošli c.e.	36
vlažnost	28
temperatura	13
prošli e. coli	6
kiša dan prije	0

Tablica 3.2: A posteriori odabir parametara



(a) ECOLI



(b) CE

Slika 3.2: ROC krivulje za modele prije i poslije a posteriori odabira parametara. Ljubičasta linija prikazuje model poslije odabira, a žuta model sa svim parametrima odabranim a priori selekcijom. Na oba grafa, crna točka predstavlja performanse modela za granicu odluke 0.5

3.6 Prilagodba granice odluke modela

Kao što je ranije navedeno, izlazna vrijednost neuronske mreže za svaki uzorak upravo je vjerojatnost da taj uzorak pripada klasi P, a neuronska mreža uzorke čija je vjerojatnost iznad neke granice odluke predviđa kao P, odnosno pozitivne na onečišćenje, a one koji su ispod te granice klasificira kao N. Zadana granica odluke je 0.5. Međutim, u procesu treniranja, svaki novi ulazni podatak prilagođava težine neuronske mreže i time mijenja vjerojatnosti i za sve prethodne ulaze. Kad je skup za treniranje nebalansiran, dogodi se da su težine bolje prilagođene većinskoj klasi pa je potrebno prilagoditi zadanu granicu odluke.

Iako je rađeno balansiranje podataka, skup za treniranje i dalje sadrži puno više uzoraka iz klase N jer bi rigorozniji undersampling rezultirao velikim gubitkom bitnih informacija. Zbog toga je model pristran prema klasi N, odnosno ta granica odluke trebala bi biti puno niža od 0.5. Na slikama 3.2 crna točka označava sensitivity i specificity koji se dobije za granicu odluke 0.5 i tu je vidljivo koliko takav model ima mali sensitivity, to jest koliko mu je loša mogućnost predviđanja onečišćenja, odnosno koliko je pristran prema većinskoj klasi N.

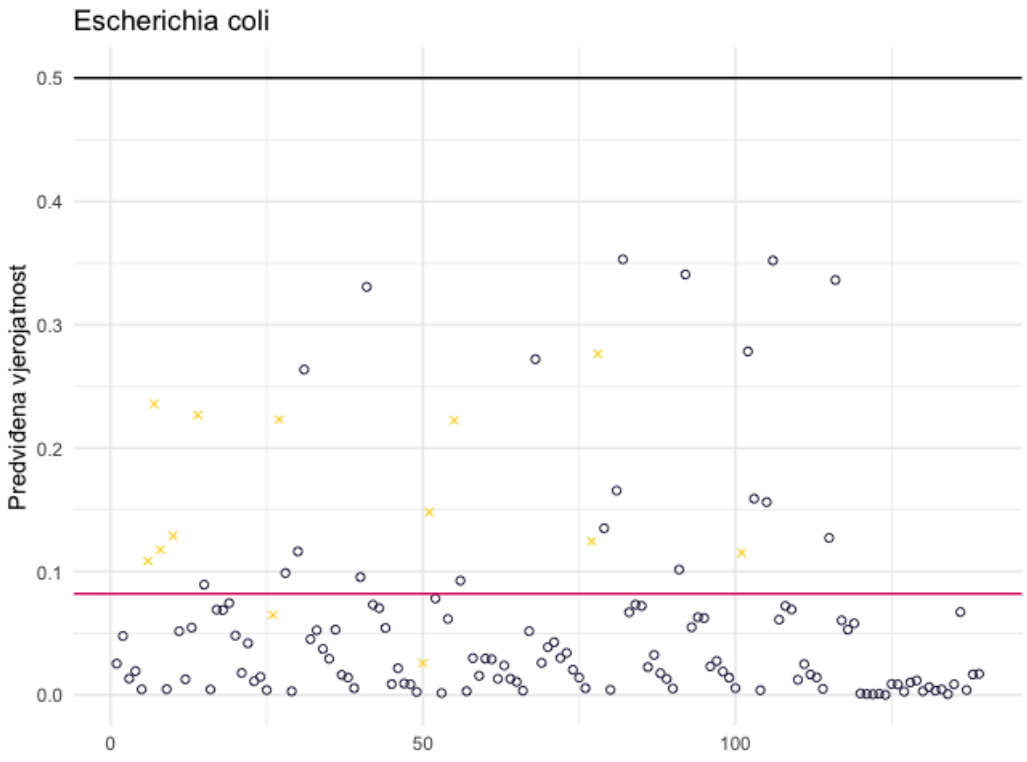
Kako bi se odredila optimalna vrijednost granice odluke, LODOCV postupak je ponavljan 10 puta i pri svakom ponavljanju bilježena je ona granica odluke za koju model ima najveći validacijski informedness. Aritmetička sredina ovako dobivenih granica odluke uzeta je za granicu odluke konačnog modela istreniranog na cijelom skupu za treniranje. Ovaj postupak proveden je odvojeno za bakterije e. coli i crijevni enterokok, dakle svaki model ima svoju graničnu vrijednost.

3.7 Evaluacija modela na holdout podacima

Kada je proveden cijeli dosad opisani postupak, preostalo je još ocijeniti dobivene modele, odnosno evaluirati njihove mogućnosti predviđanja na potpuno nepoznatim podacima. Ranije izdvojeni holdout podaci, sada su normalizirani, odnosno centrirani i skalirani kako bi bili valjan ulaz u neuronsku mrežu. Ovi su podaci centrirani i skalirani aritmetičkom sredinom i standardnom devijacijom skupa za treniranje, dakle istim parametrima kojima je normaliziran skup za treniranje, a zatim je na njima napravljeno predviđanje koje je uspoređeno sa stvarnom klasom.

3.7.1 Escherichia coli

Za bakteriju e. coli izgrađen je model `nnet` uz korištenje sljedećih šest parametara: osunčanost vrha atmosfere, osunčanost horizontalne površine, vlažnost zraka, udaljenost do najbližeg izvora onečišćenja, vlažnost zraka dan prije i temperatura zraka na dan uzorkovanja. Određena granica odluke za model je 0.082, a predviđene vjerojatnosti na holdout podacima prikazane su na slici 3.3. Osim prilagođene granice odluke 0.082, radi demonstracije na slici je prikazana i horizontalna linija koja predstavlja zadanu granicu odluke 0.5 ispod koje se nalaze sve predviđene vrijednosti. Rezultati ovako podešenog modela na holdout podacima prikazani su matricom zabune 3.3.



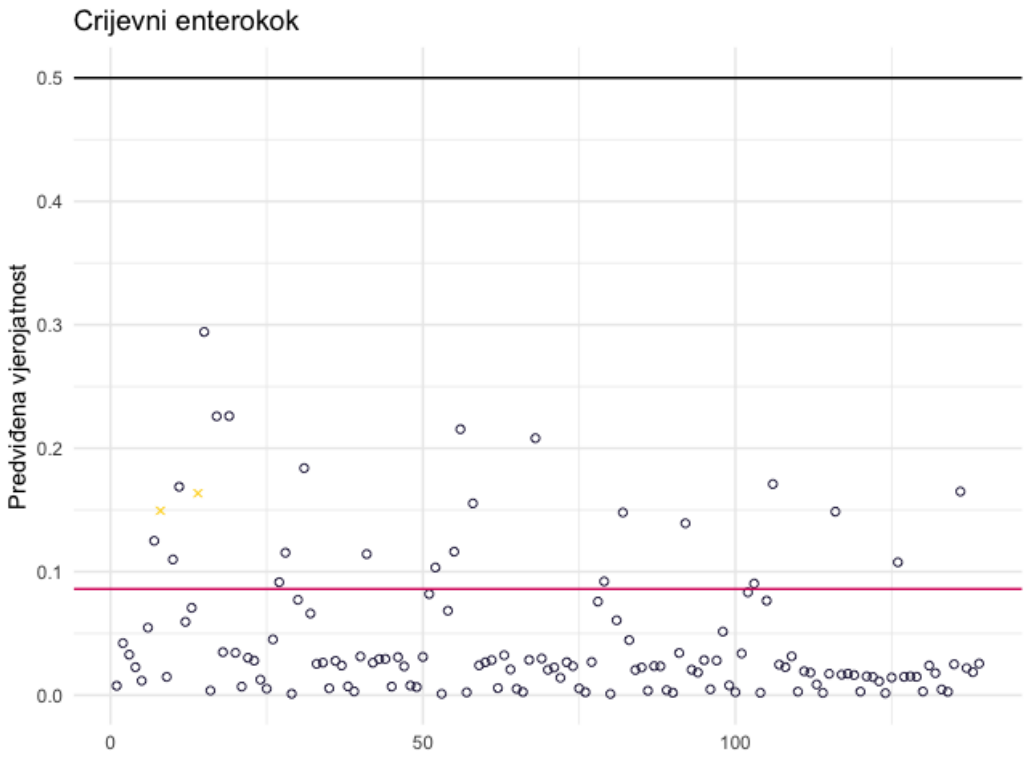
Slika 3.3: Prikazani su redom svi uzorci iz holdout skupa i predviđene vjerojatnosti da su pozitivni na onečišćenje, a različitim oblicima prikazane su njihove stvarne kategorije za e. coli

		Stvarno		
		Pozitivno	Negativno	
Predviđeno	Pozitivno	11	19	36.7%
	Negativno	2	107	98.2%
		Sensitivity = 84.6%	Specificity = 84.9%	Točnost = 84.9%
		Informedness = 69.5%		

Tablica 3.3: Matrica zabune za e. coli

3.7.2 Crijevni enterokok

Za bakteriju crijevni enterokok također je izgrađen model `nnet` ovaj put uz korištenje pet parametara: prosjek e. coli, osunčanost vrha atmosfere, osunčanost horizontalne površine, vlažnost zraka dan prije i prošla vrijednost crijevnog enterokoka. Primijenjena je granica odluke 0.086, a dobivena predviđanja prikazana su na slici 3.4. Uz predviđanja i prave klase, na slikama su prikazane i dvije linije koje predstavljaju granice odluke. Jedna predstavlja zadanu granicu 0.5, dok druga, znatno niža predstavlja granicu odluke određenu LODOCV postupkom. Odgovarajuća konfuzijska matrica za tu granicu odluke dana je tablicom 3.4.



Slika 3.4: Predviđene vjerojatnosti za onečišćenje crijevnim enterokokijem

		Stvarno		
		Pozitivno	Negativno	
Predviđeno	Pozitivno	2	23	8%
	Negativno	0	114	100%
		Sensitivity = 100%	Specificity = 83.2%	Točnost = 83.5%
		Informedness = 83.2%		

Tablica 3.4: Matrica zabune za crijevni enterokok

3.7.3 Združeni model

Cilj rada bio je stvoriti model koji uvažava onečišćenje bilo kojom od bakterija; e. coli ili crijevni enterokok. Spajanjem prethodno opisanih modela booleovom operacijom ili, dobivamo model koji odgovara zahtjevu. Dakle, uzorak će biti ocijenjen kao nezadovoljavajuć za kupanje ako predi đanje barem jednog modela bude pozitivno. Rezultati takvog modela, prikazani su konfuzijskom matricom 3.5

		Stvarno		
		Pozitivno	Negativno	
Predviđeno	Pozitivno	11	26	29.7%
	Negativno	2	100	98%
		Sensitivity = 84.6%	Specificity = 79.4%	Točnost = 79.9%
		Informedness = 64%		

Tablica 3.5: Matrica zabune za ocjenu uzorka

3.8 Zaključak

Iako je prikupljeno relativno malo podataka u kontekstu strojnog učenja, dobiveni rezultati su bolji od očekivanih. Većina iskorištenih podataka je na neki način interpolirana, bilo u vremenu, prostoru ili oboje. Osim toga, provedeno uzorkovanje je obavljano rijetko iako postoje znanstveni radovi koji tvrde da se situacija mijenja daleko brže [6]. Kada bi podaci bili precizniji (npr. meteorološke stanice na plažama...) i učestalije praćeni, i strojno učenje bi sigurno davalo pouzdanije rezultate.

Literatura

[1] Direktiva 2006/7/EZ Europskog parlamenta i vijeća od 15. veljače 2006. o upravljanju kvalitetom vode za kupanje i stavljanju izvan snage Direktive 76/160/EEZ

[2] Uredba o kakvoći mora za kupanje Vlade Republike Hrvatske

[3] D. Vukić Lušić, L. Kranjčević, S. Maćešić, D. Lušić, S. Jozić, Ž. Linšak, L. Bilajac, L. Grbčić, N. Bilajac, Temporal variations analyses and predictive modeling of microbiological seawater quality, Water Research

[4] A. Dal Pozzolo, O. Caelen, G. Bontempi, When is undersampling effective in unbalanced classification tasks?

[5] M. Gevrey, Review and comparison of methods to study the contribution of variables in artificial neural network models

[6] Mark D. Wyer, David Kay, Huw Morgan, Sam Naylor, Simon Clark, John Watkins, Cheryl M. Davies, Carol Francis, Hamish Osborn, Sarah Bennett, Within-day variability in microbial concentrations at a UK designated bathing water: Implications for regulatory monitoring and the application of predictive modelling based on historical compliance data

Literatura

[7] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A Review on Ensambles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches

[8] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction

[9] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R

[10] G. Menardi, N. Torelli, Training and assessing classiffication rules with unbalanced data

[11] N. Lunardon, G. Menardi, N. Torelli, ROSE: A Package for Binary Imbalanced Learning

[12] H. Wickham, G. Grolemund, R for Data Science

[13] G. Dreyfus, Neural Networks Methodology and Applications

[14] B.D.Ripley, Pattern Recognition and Neural Networks

[15] https://machinelearningmastery.com

TEMELJNA DOKUMENTACIJSKA KARTICA

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU
ODJEL ZA MATEMATIKU

DIPLOMSKI RAD

PRIMJENA STROJNOG UČENJA U OCJENJIVANJU KAKVOĆE MORA ZA KUPANJE

Daniela Džal

Sažetak:

Kakvoća mora za kupanje procjenjuje se na temelju količine bakterija escherichia coli i crijevni enterokok u uzorku. S obzirom da mikrobiološka analiza traje nekoliko dana, takva informacija nije upotrebljiva kod obavještavanja javnosti o trenutnoj kvaliteti vode. Stoga je cilj ovog rada primijeniti predikcijske modele na navedenim podacima. U okviru rada izgrađene su dvije feed forward neuronske mreže za klasifikaciju kakvoće mora, svaka za po jednu vrstu bakterije. Novi se uzorak klasificira kao nezadovoljavajuć ako barem jedna neuronska mreža predvidi prekomjernu količinu bakterija. Model je verificiran usporedbom sa stvarnim mjerenjima na skupu za testiranje. Dobiveni modeli klasificiraju onečišćenje prema bakteriji e. coli s informednessom 69.5%, a prema crijevnom enterokokiju s 83.2%. Opisanim spajanjem ovih modela dobije se model koji klasificira nove uzorke s informednessom 64%. S obzirom na specifičnost problema, tešku predvidljivost količine bakterija i nebalansiranost podataka, dobiveni su rezultati iznad očekivanja, a daljnje bi se poboljšanje dobilo češće uzorkovanim podacima.

Ključne riječi:

TEMELJNA DOKUMENTACIJSKA KARTICA

neuronske mreže, nebalansirani podaci, unakrsna validacija

Podatci o radu:

49 stranica, 25 slika i tablica, 15 literaturnih navoda, jezik izvornika: hrvatski

Mentor: *doc. dr. sc. Ivo Ugrina*

Neposredna voditeljica: *dr. sc. Ivana Nižetić Kosović*

Članovi povjerenstva:

prof. dr. sc. Damir Vukičević

izv. prof. dr. sc. Jurica Perić

Povjerenstvo za diplomski rad je prihvatilo ovaj rad *29. siječnja 2021.*

TEMELJNA DOKUMENTACIJSKA KARTICA

FACULTY OF SCIENCE, UNIVERSITY OF SPLIT
DEPARTMENT OF MATHEMATICS

MASTER'S THESIS

APPLICATION OF MACHINE LEARNING ON EVALUATION OF BEACH WATER QUALITY

Daniela Džal

Abstract:

Beach water quality assessment is based on the amount of e. coli and intestinal enterococci in a sample. Microbiological analysis lasts up to few days and gained information is not applicable to informing the public on present conditions. The goal of this thesis is to apply predictive models on available data. Two neural networks have been trained, each classifying beach water quality based on only one bacteria type. New sample is classified as non-satisfactory if a high level of bacteria is predicted by at least one neural network. Model is evaluated by comparing predictions with microbiological results on test data. Trained models classify beach water quality with informedness 69.5% for e. coli, and 83.2% for enterococci. Merged model has informedness 64%. Considering unpredictability of bacteria levels and data imbalance, acquired results are above expectations and further improvements could be achieved by training the model on more frequently sampled data.

Key words:

neural networks, imbalanced data, cross validation

Specifications:

49 pages, 25 figures and tables, 15 references, original in Croatian

TEMELJNA DOKUMENTACIJSKA KARTICA

Mentor: *assistant professor Ivo Ugrina*

Immediate mentor: *research fellow Ivana Nižetić Kosović*

Committee:

professor Damir Vukičević

associate professor Jurica Perić

This thesis was approved by a Thesis commettee on *January 29, 2021.*