Tema: Regresijske tehnike za ekološko modeliranje

Kolegij: Strojno učenje za okolišne primjene

Opis lekcije:

Ova lekcija uvodi studente u regresiju kao temeljni pristup strojnog učenja za modeliranje i predviđanje kontinuiranih okolišnih varijabli. Lekcija naglašava i teorijsku i primijenjenu perspektivu, razlikujući linearne i nelinearne regresijske tehnike relevantne za okolišne i ekološke studije. Studenti će istražiti cjelovit regresijski radni tijek, uključujući pripremu podataka, inženjering značajki, treniranje modela, evaluaciju primjenom RMSE-a te ublažavanje podnaučenosti i prenaučenosti.
Odabrani primjer primjene usmjeren je na predviđanje koncentracije otopljenog kisika na temelju temperature vode i povezanih okolišnih varijabli u estuarijskom području uzgoja školjkaša. Ovaj primjer pokazuje kako regresijski modeli mogu poduprijeti stvarno ekološko praćenje i sustave ranog upozoravanja na uvjete stresa u vodenim ekosustavima.

Ishodi učenja:

Nakon završetka ove cjeline studenti će moći:
1. Definirati regresijske probleme i razlikovati ih od drugih zadataka strojnog učenja.
2. Objasniti načela jednostavne i višestruke linearne regresije te logiku RMSE evaluacije.
3. Razviti i evaluirati regresijske modele primjenom okolišnih skupova podataka.
4. Prepoznati i obraditi uobičajene probleme kao što su šum, multikolinearnost, podnaučenost i prenaučenost u regresijskim problemima s ekološkim podacima.
5. Interpretirati izlaze modela u ekološkom i okolišnom kontekstu.

Sadržaj predavanja

Uvod

Regresijske tehnike ključni su alati za modeliranje kontinuiranih okolišnih varijabli, kao što su temperatura, salinitet i koncentracije onečišćivala. Ovaj odjeljak objašnjava razliku između regresijskih i klasifikacijskih zadataka, uvodi koncept korijena srednje kvadratne pogreške (RMSE) te raspravlja o važnosti regresijskih modela za razumijevanje i predviđanje ekoloških pojava.

Kontekst i motivacija

Regresijske tehnike temeljne su za modeliranje brojnih okolišnih pojava s kontinuiranim odzivom. U ekološkim i okolišnim znanostima često se nastoji predvidjeti kontinuirana varijabla (npr. koncentracija onečišćivala, biomasa vrste, riječni protok, temperaturna anomalija) na temelju skupa prediktorskih varijabli (meteorologija, krajobrazne varijable, obilježja sliva, uporaba zemljišta itd.). Regresija stoga povezuje opažačke/okolišne

podatke s objašnjavajućim i prediktivnim modeliranjem te se nalazi na sjecištu statističke ekologije, hidrologije, klimatologije i strojnog učenja.
Primjeri regresijskih problema u ekologiji i okolišnim znanostima:

- Procjena koncentracije dušikova dioksida (NO_2) ili lebdećih čestica ($PM_{2.5}$) u urbanim/industrijskim okruženjima na temelju meteoroloških ulaza, varijabli prometa/uporabe zemljišta i atmosferskih uvjeta.
- Modeliranje brojnosti vrsta (ili biomase) kroz vrijeme ili prostor kao odgovora na okolišne pokretače kao što su temperatura, oborina, hranjive tvari u tlu, fragmentacija staništa i antropogeni poremećaji.
- Predviđanje hidroloških odziva, poput dnevnog riječnog protoka, na temelju oborine, temperature zraka, vlažnosti, prethodne vlažnosti tla i obilježja sliva.
- Prognoziranje klimatski povezanih anomalija (npr. temperaturnih anomalija) ili odziva ekosustava (tokova ugljika, emisija stakleničkih plinova) uporabom kontinuiranih regresijskih okvira.

Regresijski i klasifikacijski zadaci

Važno je razlikovati regresiju od klasifikacije jer obje pripadaju skupini problema nadziranog učenja, ali služe različitim prediktivnim ciljevima.
U klasifikaciji je ciljna varijabla y kategorijska — predstavlja diskretne oznake ili klase (npr. „onečišćeni" nasuprot „čistim" uzorcima vode ili „prisutnost" nasuprot „odsutnosti" određene vrste). Nasuprot tomu, u regresiji je y kontinuirana, što znači da njezine vrijednosti mogu varirati na numeričkoj skali bez inherentnih granica klasa (npr. predviđanje točne koncentracije nitrata u podzemnoj vodi ili gustoće biomase na šumskoj plohi).
U klasifikacijskom okruženju zanimaju nas granice odlučivanja — gdje novi uzorak treba smjestiti među poznate kategorije. U regresiji nas zanima numerička preciznost — koliko je naša predviđena vrijednost bliska stvarnom kontinuiranom mjerenju.
Isticanjem te razlike na početku predavanja studenti mogu bolje razumjeti zašto je regresijska analiza ključna u okolišnim domenama, u kojima su mnoge ciljne veličine inherentno kontinuirane i zahtijevaju ne samo ispravnu procjenu trenda nego i razumijevanje nesigurnosti i varijance u predviđanjima.

Zašto je regresija problem nadziranog učenja

Regresija pripada skupini metoda nadziranog učenja jer treniranje modela zahtijeva označene podatke — parove ulaza i poznatih ciljnih vrijednosti (Xi,yi)
$\hat{y}_i = f(X_i)$Model „uči" prilagođavanjem svojih parametara kako bi minimizirao razliku između predviđenih vrijednosti i stvarnih opaženih ishoda, yi.
„Nadzor" proizlazi iz tih poznatih ishoda: proces učenja vođen je povratnom informacijom iz signala pogreške, tj. time koliko su predviđanja modela pogrešna u odnosu na stvarnost. Cilj je smanjiti pogrešku predviđanja uzastopnim aproksimacijama, vođenima ciljnom funkcijom (često srednjom kvadratnom pogreškom).

Korisnost RMSE-a u okolišnoj i ekološkoj regresiji

Ključan metodološki aspekt regresijskog modeliranja jest odabir evaluacijske metrike — kvantitativne mjere koja pokazuje koliko se predviđanja modela podudaraju s opaženim podacima.
Korijen srednje kvadratne pogreške (RMSE) jedna je od najčešće korištenih metrika u okolišnom i ekološkom modeliranju, definirana kao:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

Ova je evaluacijska metrika gotovo samoobjašnjiva — u tom smislu krenut ćemo iznutra prema van. Kvadratne pogreške znače da se pogreška definira kao jednostavno oduzimanje stvarnih i predviđenih vrijednosti, koje se zatim kvadrira, čime se izbjegava poništavanje pozitivnih i negativnih odstupanja. Kako bi se dobio ukupni učinak svih predviđenih vrijednosti, ti se kvadratni rezultati zbrajaju. Konačno, zanima nas prosječan doprinos svake instance podataka, koji se izračunava dijeljenjem ukupnog doprinosa brojem instanci. Na kraju, operacija korijena vraća izvorne mjerne jedinice.

- RMSE pruža interpretabilnu mjeru u istim jedinicama kao ciljna varijabla, što je ključno za okolišne znanstvenike koji rezultate moraju komunicirati dionicima i donositeljima politika.
- Odražava i pristranost (sustavnu pogrešku) i varijancu (raspršenje reziduala), obuhvaćajući ukupnu prediktivnu pouzdanost modela.
- Evaluacijsku metriku RMSE lako je uključiti u optimizacijske algoritme koji se koriste u strojnom učenju, čime se olakšava automatizirano podešavanje parametara.

Isticanje glavnih prednosti RMSE-a:

- RMSE studentima nudi jasan i intuitivan početni uvid: manje vrijednosti upućuju na bolju točnost predviđanja.
- Kvadratna kazna pruža prirodnu ilustraciju zašto se ekstremne pogreške predviđanja nerazmjerno kažnjavaju — ideju koja se dobro uklapa u ekološko promišljanje rizika.
- Usporedba RMSE-a s alternativnim mjerama (npr. srednjom apsolutnom pogreškom ili srednjom pristranom pogreškom) potiče kritičko promišljanje o odabiru i interpretaciji metrika u okolišnim studijama.

Pregled tehnologije

Regresijsko modeliranje u ekološkim i okolišnim znanostima ne može se promatrati isključivo kao čisto matematički postupak, nego kao matematički utemeljen pristup

proširen snažnom tehnološkom komponentom koja heterogene okolišne podatke pretvara u interpretabilne i provjerljive modele.

Kako bi se taj proces studentima učinio pristupačnijim, različiti regresijski modeli (linearna regresija, regresijska stabla, SVM za regresiju i ANN) primijenit će se na reprezentativnim skupovima podataka iz ekologije i okolišnih znanosti.
Integracijom otvorenih izvora podataka, suvremenih biblioteka za obradu podataka i tehnika vizualizacije obradit će se temeljna načela tehnoloških i metodoloških aspekata regresijskog modeliranja u ML-u za okolišne znanosti.

Tehnološka integracija i skupovi podataka

Najprije se koristi sintetički skup podataka kako bi se osnovna načela jednostavne i višestruke linearne regresije prikazala u kontroliranom i lako interpretabilnom okruženju. To studentima omogućuje da se usredotoče na razumijevanje ponašanja modela bez složenosti stvarnog šuma i nedostajućih podataka.

U drugom dijelu koristi se stvarni morski skup podataka iz Copernicus Marine Environment Monitoring Service (CMEMS), koji uključuje salinitet površine mora, koncentraciju klorofila i brzinu struje, kako bi se isti regresijski radni tijek primijenio na stvarne okolišne podatke te evaluirala izvedba modela i ekološka relevantnost.

Sve se analize provode u Google Colabu uporabom standardnih Python biblioteka:

- **pandas i NumPy za učitavanje podataka, čišćenje i numeričke operacije,**
- **scikit-learn za implementaciju regresijskih modela (jednostavna linearna regresija i višestruka linearna regresija) te izračun evaluacijskih metrika (RMSE i R^2),**
- **Matplotlib za vizualizaciju distribucija podataka, regresijskih prilagodbi i dijagnostike modela.**

Jednostavna linearna regresija koristi se kao osnovni model za opis odnosa između jednog okolišnog prediktora (npr. temperature) i kontinuirane odzivne varijable (otopljenog kisika). Njezina je glavna prednost interpretabilnost i sposobnost ilustriranja temeljnih regresijskih koncepata.

Višestruka linearna regresija proširuje ovaj pristup uključivanjem dodatnih prediktora (npr. dubine i električne vodljivosti) kako bi bolje predstavila fizičke i ekološke procese koji utječu na varijabilnost otopljenog kisika. Ovaj model pokazuje kako dodavanje relevantnih okolišnih varijabli može poboljšati objašnjavajuću snagu uz zadržavanje transparentnosti modela.

Jednostavna linearna regresija i višestruka linearna regresija: koncept, metoda i demonstracija

Jednostavna linearna regresija povezuje jedan prediktor s jednom odzivnom varijablom:

$$y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

- εGdje predstavlja neobjašnjenu varijabilnost ili mjerni šum.

U višestrukoj regresiji imamo više neovisnih varijabli, Xi, i jednu zavisnu varijablu.
Cilj je odrediti β0 i β1 (ili βk za višestruku regresiju) tako da model minimizira srednju kvadratnu pogrešku između opaženih i predviđenih vrijednosti.
Ova lekcija prikazuje jednostavnu i višestruku linearnu regresiju primjenom mjerenja okolišnih senzora iz estuarijskog područja uzgoja školjkaša. U primijenjenom slučaju koncentracija otopljenog kisika (mg/L) modelira se kao funkcija temperature vode i dodatnih prediktora (npr. dubine i vodljivosti). Takva postavka pokazuje kako linearni modeli mogu pružiti interpretabilne osnovne odnose i kako modeli s više prediktora mogu poboljšati objašnjavajuću snagu u ekološkim sustavima.

Koraci:

Priprema podataka:
Učitajte skup podataka okolišnih senzora koji sadrži mjerenja temperature vode, koncentracije otopljenog kisika i opcionalnih prediktora kao što su dubina i električna vodljivost. Pretvorite sve skalirane vrijednosti u fizičke jedinice, pregledajte skup podataka radi nedostajućih vrijednosti i odstupajućih vrijednosti te podijelite podatke na skup za treniranje i testni skup.
Analiza:
Prilagodite dva regresijska modela:
(i) jednostavni linearni regresijski model koji koristi temperaturu kao jedini prediktor otopljenog kisika (Temp → DO) i
(ii) višestruki linearni regresijski model koji koristi temperaturu, dubinu i vodljivost kao prediktore otopljenog kisika (Temp + Depth + Conductivity → DO).
Evaluacija:
Evaluirajte oba modela primjenom korijena srednje kvadratne pogreške (RMSE) i koeficijenta determinacije (R^2). Vizualizirajte predviđene nasuprot opaženim vrijednostima i analizirajte dijagrame reziduala kako biste procijenili sustavne pogreške i pretpostavke modela.
Poboljšavanje modela:
Ispitajte multikolinearnost među prediktorima i raspravite njezin utjecaj na stabilnost modela. Kao proširenje uvedite tehnike regularizacije, kao što su Ridge i Lasso regresija, kako biste smanjili prenaučenost i poboljšali robusnost modela.
Interpretacija:
Interpretirajte regresijske koeficijente i obrasce reziduala u ekološkim i okolišnim terminima, uzimajući u obzir fizičke i biološke mehanizme poput topljivosti kisika, fotosintetske aktivnosti i procesa miješanja vode.

Nelinearna regresija: koncept, metoda i demonstracija

Nelinearna regresija proširuje ideju linearne regresije na odnose koje pravac ne može primjereno opisati. U mnogim ekološkim i okolišnim procesima odzivna varijabla ne mijenja se linearno s prediktorima — umjesto toga može pokazivati pragove, zasićenje, eksponencijalni rast ili cikličke obrasce.
Formalna definicija nelinearne regresije može se izraziti kao:

$$y_i = f\left(X_i, \beta\right) + \varepsilon_i$$

- Gdje je f(·) nelinearna funkcija prediktora Xi i parametara β, a εi predstavlja slučajnu pogrešku.

Uobičajeni funkcijski oblici uključuju polinomne, eksponencijalne, logaritamske ili potencijske transformacije, kao i modele temeljene na jezgrama, stablima ili ANN-u kada je potrebna neparametarska fleksibilnost.

$\hat{y}_i$Cilj ostaje isti kao u linearnoj regresiji: minimizirati razliku između opaženih (yi) i predviđenih () vrijednosti, obično primjenom RMSE-a.
Međutim, u nelinearnoj regresiji prilagođena krivulja može poprimiti mnogo različitih oblika, ovisno o ekološkoj teoriji, ponašanju podataka i transformacijama značajki.

Ova demonstracija usredotočena je na modeliranje koncentracije otopljenog kisika (DO) kao funkcije temperature vode uporabom sintetičkog skupa podataka iz estuarijskog područja uzgoja školjkaša. Polinomna regresija može se koristiti za ilustraciju načina na koji nelinearni modeli zahvaćaju zakrivljene obrasce odziva kada linearne pretpostavke nisu dovoljne.

Koraci:

Priprema podataka:

Dohvatite kontinuirana mjerenja temperature vode i koncentracije otopljenog kisika iz sustava praćenja u estuarijskom uzgajalištu školjkaša. Dopunite stvarni skup podataka sintetičkim skupovima podataka osmišljenima za naglašavanje nelinearnih obrazaca u didaktičke svrhe. Obradite nedostajuće vrijednosti, normalizirajte prediktore i podijelite skup podataka na skup za treniranje, validacijski skup i testni skup.

Analiza:

Primijenite polinomnu regresiju kako biste zahvatili nelinearne odnose između temperature vode i koncentracije otopljenog kisika. Usporedite polinomni model s regresijom potpornih vektora (SVR), regresijom slučajnih

šuma i umjetnim neuronskim mrežama (ANN) radi istraživanja složenijih nelinearnih struktura.

Evaluacija:

Izračunajte RMSE za sve modele i vizualizirajte predviđene nasuprot opaženim vrijednostima uporabom dvodimenzijskih dijagrama (temperatura nasuprot otopljenom kisiku) i opcionalnih trodimenzijskih dijagrama kada su uključeni dodatni prediktori. Ispitajte nelinearne krivulje odziva kako biste otkrili ekološke pragove, primjerice razine temperature iznad kojih otopljeni kisik naglo opada.

Optimizacija modela:

Primijenite deseterostruku unakrsnu validaciju za odabir stupnja polinoma i podešavanje hiperparametara jezgre ili ANN-a. Usporedite izvedbu modela na testnom skupu procjenjujući i pogrešku generalizacije i interpretabilnost modela.

Interpretacija:

Povežite nalaze modela s procesima u vodenim ekosustavima i raspravite o implikacijama za upravljanje uzgojem školjkaša, rano upozoravanje na hipoksične uvjete i klimatski potaknut stres u estuarijskim okolišima.

Evaluacijske metrike za regresijske probleme

Evaluacijske metrike kvantificiraju prediktivnu izvedbu modela, dok dijagnostika procjenjuje koliko se struktura modela podudara s temeljnim ekološkim procesom. Cilj je razumjeti ne samo koliko je model točan, nego i zašto se ponaša tako kako se ponaša — gdje uspijeva, gdje zakazuje i što te pogreške znače.

Najčešće korištene mjere u ekološkoj regresiji jesu:

Korijen srednje kvadratne pogreške (RMSE) – već spomenuto:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

Osjetljiv na velike pogreške — idealan kada ekstremna odstupanja imaju znatne ekološke posljedice (npr. podcjenjivanje riječnog protoka tijekom poplave ili precjenjivanje razina onečišćivala).

Srednja apsolutna pogreška (MAE):

$$MAES = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Manje je osjetljiva na odstupajuće vrijednosti i pruža uvid u prosječno odstupanje.

Koeficijent determinacije (R^2):

Ova mjera nije pouzdana za vremenske nizove ni prostorne podatke. U drugim slučajevima mjeri udio varijance u y koji model objašnjava, što pomaže u procjeni ukupne objašnjavajuće snage.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ ¬ SSres → rezidualna suma kvadrata (neobjašnjena varijanca)

$SS_{tot}\sum_{i=1}^{n}(y_i - \bar{y})^2$ ¬ → ukupna suma kvadrata (ukupna varijanca u opaženim podacima)

y_i ¬ = opažena vrijednost

$\hat{y}_i$ ¬ = predviđanje modela

$\acute{y}$ ¬ = srednja vrijednost svih opaženih vrijednosti

n ¬ = broj uzoraka

Nash–Sutcliffeova učinkovitost (NSE):

Uobičajena u hidrologiji; izražava prediktivnu vještinu u odnosu na srednju vrijednost opažanja.

$$NSE = 1 - \frac{\sum (Q_{obs} - Q_{mod})^2}{\sum (Q_{obs} - \overline{Q})^2}$$

$\overline{Q}$pri čemu se Qobs odnosi na opažanja (u ovom slučaju protoka Q), Qmod na simulacije iste varijable kao i opažanja, a predstavlja prosječno opažanje odabrane varijable.

Primijenjeni slučaj: regresijsko modeliranje otopljenog kisika u estuarijskom području uzgoja školjkaša

Pozadina i motivacija

Na lokaciji uzgoja školjkaša smještenoj u estuarijskom/riječnom okolišu temperatura vode i otopljeni kisik (DO, mg/L) kontinuirano se mjere in situ senzorima. Otopljeni kisik izravno utječe na disanje i stres školjkaša te je stoga ključan pokazatelj zdravlja ekosustava i održivosti akvakulture.

Cilj ovog primjera primjene jest izgraditi regresijski model koji procjenjuje DO na temelju temperature vode. Taj model služi kao brzi pokazatelj potencijalno nepovoljnih uvjeta: kada temperatura poraste, očekuje se smanjenje koncentracije kisika. Rano otkrivanje takvih uvjeta omogućuje uzgajivačima učinkovitije praćenje i ublažavanje rizika od ekološkog stresa.

Studenti uspoređuju predviđanja modela s izmjerenim vrijednostima i analiziraju situacije u kojima model zakazuje. To pruža intuitivan uvod u ideju da se regresijski modeli mogu poboljšati uključivanjem dodatnih okolišnih prediktora kao što su zasićenje kisikom (%), dubina, vodljivost i zamućenost.

Jednostavna linearna regresija

Korak 1 – Razumijevanje podataka

Opis skupa podataka i varijable

Skup podataka korišten u ovom primijenjenom slučaju potječe iz kontinuiranog okolišnog praćenja na estuarijskoj lokaciji uzgoja školjkaša. Mjerenja su prikupljena in situ senzorima i uključuju temperaturu vode i koncentraciju otopljenog kisika, zajedno s dodatnim okolišnim varijablama koje se koriste za proširenje modela.
Skup podataka sadrži nekoliko tisuća opažanja i ne uključuje nedostajuće vrijednosti u primarnim varijablama koje se koriste za modeliranje. Sve vrijednosti bile su zadane u skaliranim jedinicama i tijekom predobrade pretvorene su u fizičke jedinice (°C i mg/L).

Naziv skupa podataka:

In situ sustav okolišnog praćenja (sintetički skup podataka pripremljen u obrazovne svrhe) Regresijski problem definira se na sljedeći način:

Ciljna varijabla (y):
Koncentracija otopljenog kisika (DO, mg/L)
Primarni prediktor (X):
Temperatura vode (Temp, °C)

Dodatni prediktori (za višestruku regresiju):

- Dubina (m)
- Električna vodljivost (µS/cm)
- (opcionalno) zasićenje kisikom (%)
- (opcionalno) zamućenost

Te su varijable odabrane na temelju njihove fizičke i ekološke relevantnosti. Temperatura izravno utječe na topljivost kisika, dok dubina i vodljivost odražavaju strukturu vodenog stupca i procese miješanja koji utječu na dostupnost kisika.
U regresijskim modelima nisu korišteni svi dostupni atributi skupa podataka. Varijable kao što su identifikatori zapisa i nefizički kodovi isključene su kako bi se izbjegli šum i prenaučenost te očuvala interpretabilnost regresijskih rezultata.

Ovaj problem predstavlja zadatak nadzirane regresije s kontinuiranom ciljnom varijablom.

Korak 2 – Pregled podataka

Skup podataka okolišnih senzora učitan je uporabom biblioteke Pandas. Sva skalirana mjerenja pretvorena su u fizičke jedinice (°C za temperaturu i mg/L za otopljeni kisik) kako bi se osigurala fizička interpretabilnost regresijskih modela.

Skup podataka uvezen je iz CSV datoteke odvojene točkom sa zarezom i predobrađen pretvaranjem skaliranih vrijednosti u fizičke jedinice (Slika 1). Pregled prvih opažanja potvrđuje ispravan uvoz podataka i priprema skup podataka za regresijsku analizu (Slika 2).

```python
# Load CSV (semicolon-separated)
df = pd.read_csv("/content/KrkaRiverDataset_synthetic_multi_R2_40pct.csv", sep=";")

# Convert scaled values to physical units
df["Temp_deg_C"] = df["Temp_deg_C"] / 100
df["HDO_mg/l"]   = df["HDO_mg/l"] / 100

# Quick inspection
display(df[["Temp_deg_C", "HDO_mg/l", "Depth_m", "SpCond_uS/cm"]].head())
```

Slika 1 Python kôd za učitavanje skupa podataka okolišnih senzora iz CSV datoteke odvojene točkom sa zarezom i pretvaranje skaliranih vrijednosti temperature i otopljenog kisika u fizičke jedinice (°C i mg/L).

	Temp_C	DO_mgL
0	14.81	6.32
1	15.26	4.96
2	16.43	7.09
3	15.68	7.14
4	16.21	5.73

Slika 2 Pregled prvih opažanja nakon pretvorbe jedinica, koji potvrđuje ispravan uvoz podataka i fizičku interpretabilnost vrijednosti temperature i otopljenog kisika.

U primarnim varijablama korištenima za regresijsko modeliranje nisu otkrivene nedostajuće vrijednosti (Slika 3, Slika 4).

```python
# Missing values check
print("Missing values:")
print(df[["Temp_deg_C", "HDO_mg/l"]].isnull().sum())
```

Slika 3 Python kôd korišten za provjeru nedostajućih vrijednosti u varijablama temperature i otopljenog kisika.

```
Missing values:
Temp_deg_C    0
HDO_mg/l      0
```

Slika 4 Izlaz koji pokazuje da u glavnim varijablama korištenima za regresijsko modeliranje nema nedostajućih vrijednosti.

Deskriptivna statistika korištena je za pregled raspona i varijabilnosti vrijednosti temperature i otopljenog kisika te za prepoznavanje potencijalnih odstupajućih vrijednosti (Slika 5, Slika 6).

```
# Descriptive stats
print("\nDescriptive statistics:")
display(df[["Temp_deg_C", "HDO_mg/l"]].describe())
```

Slika 5 Python kôd za izračun deskriptivne statistike (broj opažanja, srednja vrijednost, standardna devijacija i raspon) za varijable temperature i otopljenog kisika.

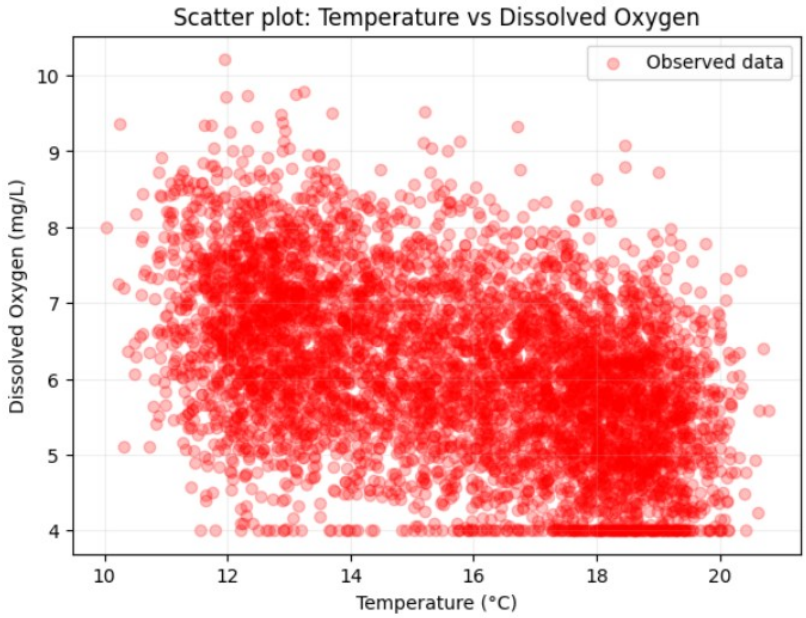
	Temp_deg_C	HDO_mg/l
count	5359.000000	5359.000000
mean	15.679737	6.104551
std	2.559164	1.119456
min	10.020000	4.000000
25%	13.285000	5.305000
50%	15.890000	6.070000
75%	17.990000	6.880000
max	20.790000	10.210000

Slika 6 Sažeta statistika vrijednosti temperature i otopljenog kisika korištena za procjenu raspona i varijabilnosti podataka.

Za istraživanje odnosa između temperature vode i koncentracije otopljenog kisika izrađen je dijagram raspršenja (Slika 7, Slika 8).

```
# Scatter plot
plt.figure(figsize=(7,5))
plt.scatter(df["Temp_deg_C"], df["HDO_mg/l"],
            color="red", alpha=0.25, label="Observed data")
plt.xlabel("Temperature (°C)")
plt.ylabel("Dissolved Oxygen (mg/L)")
plt.title("Scatter plot: Temperature vs Dissolved Oxygen")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Slika 7 Python kôd za vizualizaciju odnosa između temperature (°C) i otopljenog kisika (mg/L) uporabom dijagrama raspršenja.



Slika 8 Dijagram raspršenja otopljenog kisika (mg/L) u odnosu na temperaturu vode (°C)

Dijagram raspršenja otkriva jasan negativan trend: otopljeni kisik opada kako temperatura raste. Taj je obrazac u skladu s fizikalnim načelima topljivosti plinova u vodi. Izostanak snažne zakrivljenosti ili učinaka praga upućuje na to da linearni regresijski model može poslužiti kao prikladna prva aproksimacija. Opaženi približno linearni odnos između temperature i otopljenog kisika podupire uporabu linearne regresije kao osnovnog modela.

Korak 3 – Priprema skupa podataka za regresijski model

Za regresijsko modeliranje iz skupa podataka odabrani su relevantni prediktori i ciljna varijabla.
Ciljna varijabla definirana je kao koncentracija otopljenog kisika (DO), dok su prediktorske varijable uključivale temperaturu vode za jednostavni regresijski model te temperaturu vode, dubinu i električnu vodljivost za višestruki regresijski model.

Skup podataka podijeljen je na podskupove za treniranje i testiranje u omjeru 80/20. Skup za treniranje korišten je za prilagodbu regresijskih modela, dok je testni skup rezerviran za neovisnu evaluaciju izvedbe.
Takvo razdvajanje osigurava nepristranu procjenu sposobnosti generalizacije modela.

Korak 4 – Osnovni linearni model (samo za usporedbu)

Treniranje modela

Skup podataka podijeljen je na podskupove za treniranje i testiranje (podjela 80/20). Jednostavni linearni regresijski model prilagođen je uporabom temperature kao jedinog prediktora (Slika 9).

```
# Define predictor and target
X_simple = df[["Temp_deg_C"]]
y = df["HDO_mg/l"]

# Train/test split
X_train_s, X_test_s, y_train, y_test = train_test_split(
    X_simple, y, test_size=0.2, random_state=0
)

# Fit model
lin_model = LinearRegression()
lin_model.fit(X_train_s, y_train)

# Predict
y_pred_simple = lin_model.predict(X_test_s)
```

Slika 9 Python kôd za treniranje jednostavnog linearnog regresijskog modela uporabom temperature vode kao prediktora koncentracije otopljenog kisika, uključujući podjelu na skup za treniranje i testni skup te predviđanje.

Jednadžba modela i izvedba

Izvedba modela evaluirana je uporabom RMSE-a i R^2 (Slika 10, Slika 11).

```
# Evaluate
rmse_simple = np.sqrt(mean_squared_error(y_test, y_pred_simple))
r2_simple = r2_score(y_test, y_pred_simple)

print("SIMPLE LINEAR REGRESSION (Temp → DO)")
print(f"Equation: DO ≈ {lin_model.intercept_:.3f} + ({lin_model.coef_[0]:.3f}) * Temp")
print(f"RMSE = {rmse_simple:.4f} mg/L")
print(f"R²   = {r2_simple:.3f}")
```

Slika 10 Python kôd za evaluaciju linearnog regresijskog modela i vizualizaciju regresijskog pravca

```
LINEAR REGRESSION (Temp → DO)
Equation: DO ≈ 9.183 + (-0.197) * Temp
RMSE = 0.9644 mg/L
R²   = 0.204
```

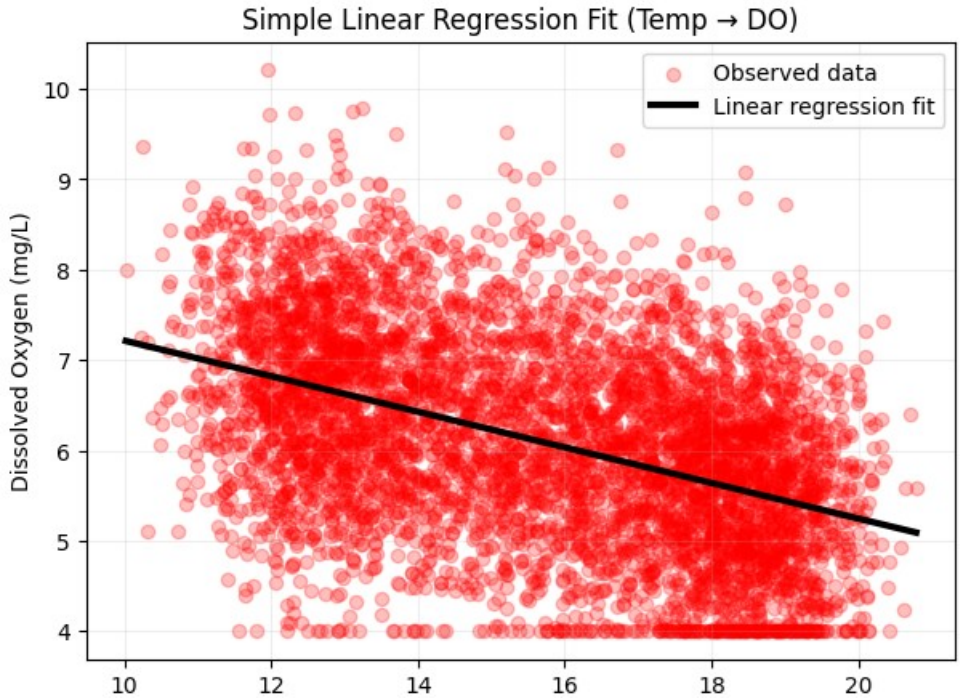
Slika 11 Izlaz koji prikazuje regresijsku jednadžbu i metrike izvedbe (RMSE i R^2) za jednostavni linearni regresijski model

Kako bi se vizualno procijenio prilagođeni linearni odnos, regresijski pravac prikazan je zajedno s opaženim podatkovnim točkama (Slika 12, Slika 13).

```python
# Plot Regression line
temp_grid = np.linspace(df["Temp_deg_C"].min(), df["Temp_deg_C"].max(), 300)
temp_grid_df = pd.DataFrame({"Temp_deg_C": temp_grid})
do_line = lin_model.predict(temp_grid_df)

plt.figure(figsize=(7,5))
plt.scatter(df["Temp_deg_C"], df["HDO_mg/l"],
            color="red", alpha=0.25, label="Observed data")
plt.plot(temp_grid, do_line,
         color="black", linewidth=3, label="Linear regression fit")
plt.xlabel("Temperature (°C)")
plt.ylabel("Dissolved Oxygen (mg/L)")
plt.title("Simple Linear Regression Fit (Temp → DO)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Slika 12 Python kôd za prikaz regresijskog pravca i opaženih podataka



Slika 13 Prilagodba linearne regresije između temperature vode i koncentracije otopljenog kisika

Regresijski koeficijent potvrđuje negativan odnos između temperature i koncentracije otopljenog kisika, u skladu s fizikalnim načelom prema kojem se topljivost kisika smanjuje s porastom temperature vode.

Vrijednost RMSE od 0,964 mg/L označava prosječnu pogrešku predviđanja modela, dok vrijednost R^2 od 0,204 sugerira da sama temperatura objašnjava približno 20 % varijance koncentracije otopljenog kisika.

Dijagnostika modela

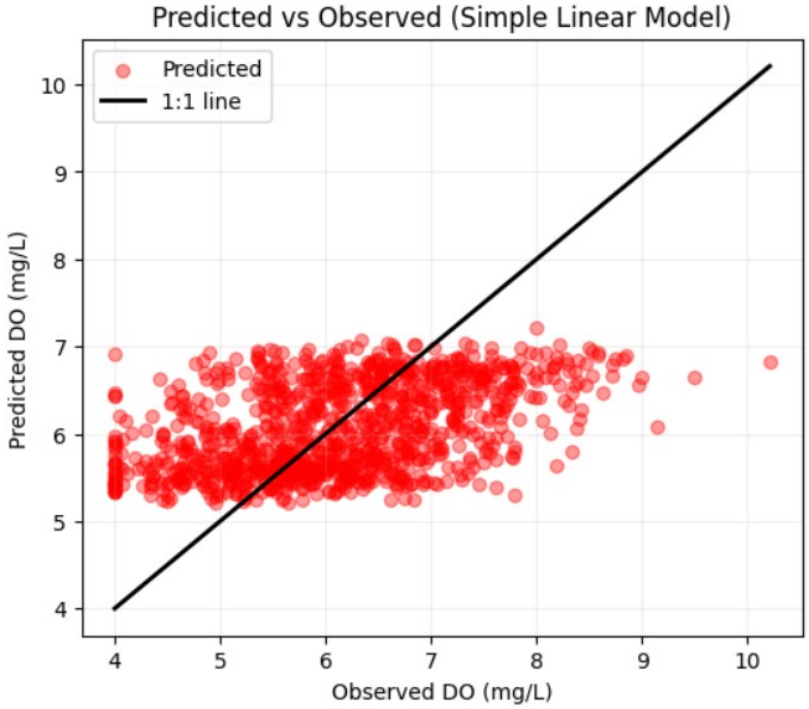
Predviđeno nasuprot opaženom

Izrađen je dijagram predviđenih nasuprot opaženim vrijednostima kako bi se ispitalo koliko se predviđanja modela podudaraju s izmjerenim vrijednostima otopljenog kisika (Slika 14, Slika 15).

```
# Predicted vs Observed
plt.figure(figsize=(6,5))
plt.scatter(y_test, y_pred_simple,
            color="red", alpha=0.4, label="Predicted")
plt.plot([y_test.min(), y_test.max()],
         [y_test.min(), y_test.max()],
         color="black", linewidth=2, label="1:1 line")

plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Predicted DO (mg/L)")
plt.title("Predicted vs Observed (Simple Linear Model)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Slika 14 Python kôd za izradu dijagrama predviđenih nasuprot opaženim vrijednostima za jednostavni linearni regresijski model



Slika 15 Predviđene nasuprot opaženim koncentracijama otopljenog kisika za jednostavni linearni regresijski model

Točke se grupiraju oko dijagonalne linije, što pokazuje da model zahvaća glavni opadajući trend otopljenog kisika s porastom temperature. Međutim, i dalje je prisutno uočljivo raspršenje, što upućuje na to da sama temperatura ne objašnjava u potpunosti varijabilnost koncentracije otopljenog kisika.

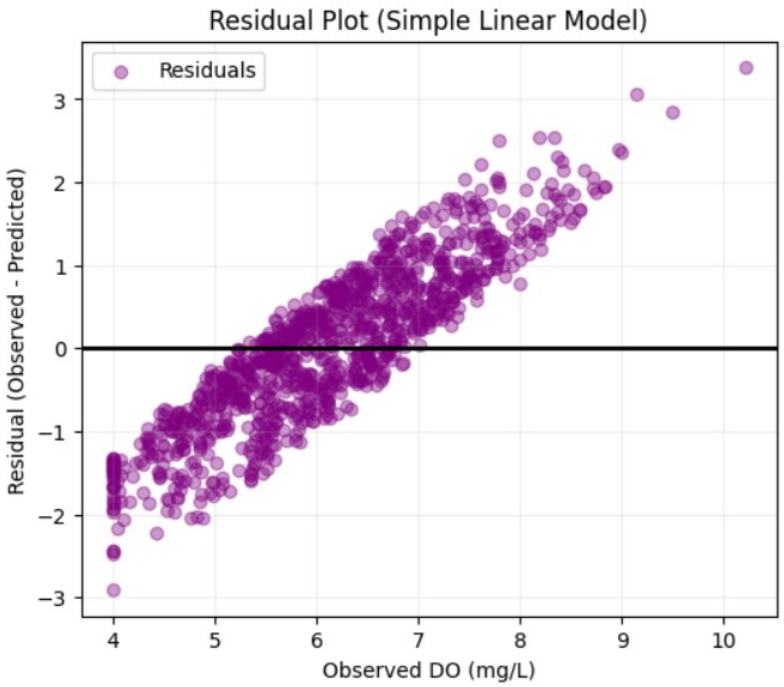
Analiza reziduala

Reziduali su izračunani kao razlika između opaženih i predviđenih vrijednosti otopljenog kisika (Slika 16, Slika 17).

```
# Residual plot
resid_simple = y_test - y_pred_simple

plt.figure(figsize=(6,5))
plt.scatter(y_test, resid_simple,
            color="purple", alpha=0.4, label="Residuals")
plt.axhline(0, color="black", linewidth=2)
plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Residual (Observed - Predicted)")
plt.title("Residual Plot (Simple Linear Model)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Slika 16 Python kôd za izračun i prikaz reziduala za jednostavni linearni regresijski model



Slika 17 Dijagram reziduala za jednostavni linearni regresijski model (opaženi minus predviđeni otopljeni kisik)

Reziduali su raspoređeni oko nule bez snažne sustavne strukture, što upućuje na to da linearni model zahvaća dominantan trend, ali ostavlja neobjašnjenu varijabilnost uzrokovanu dodatnim ekološkim i fizičkim procesima poput miješanja vode i biološke aktivnosti.

Ti rezultati ističu ograničenja uporabe temperature kao jedinog prediktora i motiviraju razvoj višestrukog linearnog regresijskog modela koji uključuje dodatne okolišne varijable.

Višestruki linearni regresijski model (Temp + Depth + Conductivity → DO)

Treniranje modela

Kako bi se poboljšala prediktivna izvedba i uzeli u obzir dodatni fizički pokretači varijabilnosti otopljenog kisika, razvijen je višestruki linearni regresijski model koji koristi temperaturu vode, dubinu i električnu vodljivost kao prediktorske varijable.
Ti su prediktori odabrani na temelju njihove ekološke relevantnosti. Temperatura vode izravno upravlja topljivošću kisika, dok dubina i električna vodljivost pružaju informacije o strukturi vodenog stupca i procesima miješanja koji utječu na raspodjelu kisika.
Skup podataka podijeljen je na podskupove za treniranje i testiranje uporabom istog omjera 80/20 kao u jednostavnom linearnom regresijskom modelu kako bi se osigurala poštena usporedba izvedbe modela. Model je treniran na skupu za treniranje, a zatim korišten za predviđanje koncentracija otopljenog kisika za testni skup (Slika 18).

```python
# Define predictors and target
X_multi = df[["Temp_deg_C", "Depth_m", "SpCond_uS/cm"]]
y = df["HDO_mg/l"]

# Train/test split
X_train, X_test, y_train, y_test = train_test_split(
    X_multi, y, test_size=0.2, random_state=0
)

# Fit multiple linear regression model
multi_model = LinearRegression()
multi_model.fit(X_train, y_train)

# Predictions
y_pred_multi = multi_model.predict(X_test)

# Evaluation metrics
rmse_multi = np.sqrt(mean_squared_error(y_test, y_pred_multi))
r2_multi = r2_score(y_test, y_pred_multi)

print("MULTIPLE LINEAR REGRESSION (Temp + Depth + Conductivity → DO)")
print("Intercept:", multi_model.intercept_)
print("Coefficients:", multi_model.coef_)
print(f"RMSE = {rmse_multi:.4f} mg/L")
print(f"R²   = {r2_multi:.3f}")
```

Slika 18 Python kôd za treniranje i evaluaciju višestrukog linearnog regresijskog modela uporabom temperature, dubine i električne vodljivosti kao prediktora

Jednadžba modela i izvedba

Izvedba modela evaluirana je uporabom korijena srednje kvadratne pogreške (RMSE) i koeficijenta determinacije (R^2).

U usporedbi s jednostavnim linearnim regresijskim modelom, višestruki linearni regresijski model postigao je niži RMSE i višu vrijednost R^2, što upućuje na poboljšanu točnost predviđanja i povećanu sposobnost objašnjavanja varijabilnosti koncentracije otopljenog kisika (Slika 19).

Regresijski koeficijenti pokazuju da temperatura ostaje dominantan prediktor s negativnim koeficijentom, potvrđujući obrnut odnos između temperature i koncentracije otopljenog kisika. Dubina i električna vodljivost pridonose dodatnoj objašnjavajućoj snazi, odražavajući utjecaj stratifikacije i procesa miješanja na dostupnost kisika.

```
MULTIPLE LINEAR REGRESSION (Temp + Depth + Conductivity → DO)
Intercept: 21.181625209743284
Coefficients: [-0.1950301   0.57396502 -0.04874799]
RMSE = 0.8230 mg/L
R²   = 0.420
```

Slika 19 Izlaz koji prikazuje regresijske koeficijente i metrike izvedbe (RMSE i R^2) za višestruki linearni regresijski model

Ti rezultati pokazuju da višestruki linearni regresijski model objašnjava približno 42 % varijabilnosti koncentracije otopljenog kisika i daje jasno poboljšanje u odnosu na jednostavni regresijski model koji koristi samo temperaturu.

Predviđene nasuprot opaženim vrijednostima

Izrađen je dijagram predviđenih nasuprot opaženim vrijednostima kako bi se vizualno procijenilo slaganje između izmjerenih i predviđenih koncentracija otopljenog kisika za višestruki regresijski model (Slika 20). Podatkovne točke grupiraju se bliže dijagonalnoj liniji 1:1 u usporedbi s jednostavnim linearnim regresijskim modelom, što pokazuje poboljšanu prediktivnu izvedbu kada su uključeni dodatni okolišni prediktori.

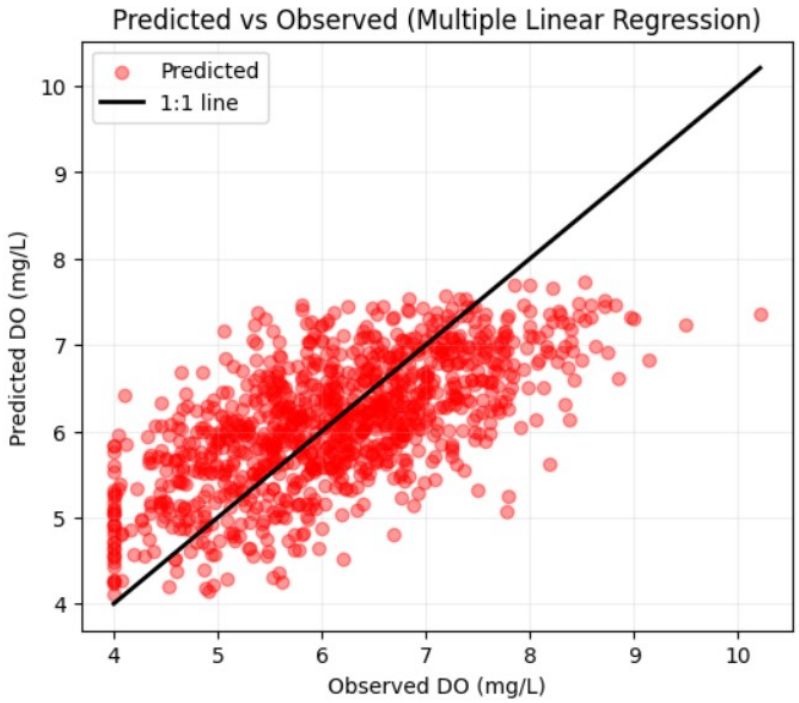
Preostala odstupanja od dijagonalne linije upućuju na utjecaj nemodeliranih bioloških i fizičkih procesa kao što su fotosinteza, respiracija i kratkoročna hidrodinamička varijabilnost (Slika 21).

Općenito, dijagram predviđenih nasuprot opaženim vrijednostima potvrđuje da uključivanje dubine i električne vodljivosti kao dodatnih prediktora povećava sposobnost modela da reproducira opažene koncentracije otopljenog kisika.

```python
# Predicted vs Observed (Multiple Linear Regression)
plt.figure(figsize=(6,5))
plt.scatter(y_test, y_pred_multi,
            color="red", alpha=0.4, label="Predicted")
plt.plot([y_test.min(), y_test.max()],
         [y_test.min(), y_test.max()],
         color="black", linewidth=2, label="1:1 line")

plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Predicted DO (mg/L)")
plt.title("Predicted vs Observed (Multiple Linear Regression)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Slika 20 Python kôd za izradu dijagrama predviđenih nasuprot opaženim vrijednostima za višestruki linearni regresijski model.



Slika 21 Predviđene nasuprot opaženim koncentracijama otopljenog kisika za višestruki linearni regresijski model

Analiza reziduala

Reziduali su analizirani kako bi se prepoznale sustavne pogreške i procijenila valjanost pretpostavki linearnog modela (Slika 22).

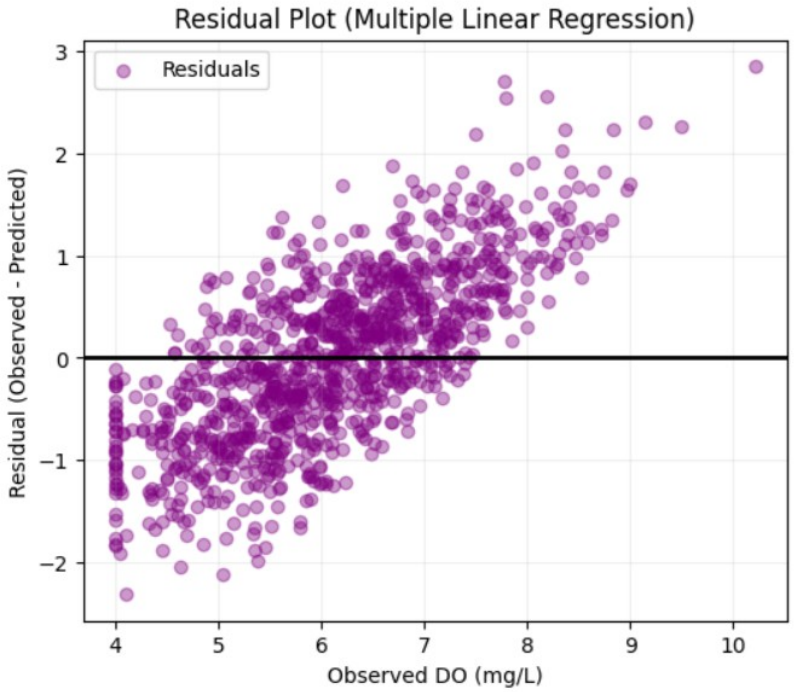
```
# Residual plot (Multiple Linear Regression)
resid_multi = y_test - y_pred_multi

plt.figure(figsize=(6,5))
plt.scatter(y_test, resid_multi, color="purple", alpha=0.4, label="Residuals")
plt.axhline(0, color="black", linewidth=2)

plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Residual (Observed - Predicted)")
plt.title("Residual Plot (Multiple Linear Regression)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Slika 22 Python kôd za izračun reziduala i izradu dijagrama reziduala za višestruki linearni regresijski model

Dijagram reziduala pokazuje da su reziduali raspoređeni oko nule bez snažnog sustavnog obrasca, što upućuje na to da višestruki linearni regresijski model zahvaća glavne linearne odnose između otopljenog kisika i odabranih prediktora (temperature, dubine i električne vodljivosti) (Slika 23).



Slika 23 Dijagram reziduala za višestruki linearni regresijski model koji prikazuje razliku između opaženih i predviđenih koncentracija otopljenog kisika

U usporedbi s jednostavnim linearnim regresijskim modelom, reziduali pokazuju homogenije raspršenje, što sugerira smanjenu pristranost i poboljšanu stabilnost modela kada su uključeni dodatni okolišni prediktori.

Međutim, određeno raspršenje i dalje je prisutno pri višim temperaturama i nižim razinama kisika, što vjerojatno odražava utjecaj nelinearnih bioloških i fizičkih procesa poput fotosinteze, respiracije i kratkoročnog hidrodinamičkog miješanja, koje linearni model s ograničenim brojem prediktora ne može u potpunosti prikazati.

Općenito, analiza reziduala potvrđuje da višestruki linearni regresijski model pruža razumnu aproksimaciju dinamike otopljenog kisika, istodobno naglašavajući potrebu za složenijim (nelinearnim) modelima kako bi se zahvatila preostala varijabilnost.

Usporedba s jednostavnim linearnim regresijskim modelom

Višestruki linearni regresijski model uspoređen je s jednostavnim linearnim regresijskim modelom koji je koristio samo temperaturu vode kao prediktor koncentracije otopljenog kisika.
Usporedba pokazuje da je višestruki linearni regresijski model postigao niži korijen srednje kvadratne pogreške (RMSE) i viši koeficijent determinacije (R^2) od jednostavnog linearnog regresijskog modela. To upućuje na poboljšanu točnost predviđanja i povećanu

sposobnost objašnjavanja varijabilnosti koncentracije otopljenog kisika kada su uključeni dodatni okolišni prediktori.

Iako jednostavni linearni regresijski model zahvaća dominantan obrnuti odnos između temperature i otopljenog kisika, ne uzima u obzir druge važne fizičke pokretače, poput strukture vodenog stupca i uvjeta miješanja. Uključivanjem dubine i električne vodljivosti višestruki linearni regresijski model bolje predstavlja složenost estuarijske dinamike kisika.

Ova usporedba pokazuje ograničenja regresijskih modela s jednim prediktorom za ekološke sustave i naglašava važnost uporabe više ekološki smislenih varijabli pri modeliranju okolišnih procesa.

Ekološka interpretacija

Negativni regresijski koeficijent za temperaturu odražava dobro poznato fizikalno načelo prema kojem se topljivost kisika smanjuje s porastom temperature vode. Toplija voda sadrži manje otopljenog kisika, što izravno utječe na disanje školjkaša i povećava fiziološki stres u akvakulturnim okolišima.

Doprinosi dubine i električne vodljivosti upućuju na ulogu stratifikacije, dotoka slatke vode i procesa miješanja u oblikovanju dostupnosti kisika unutar estuarijskog sustava. Varijacije dubine utječu na prodiranje svjetlosti i biološku aktivnost, dok vodljivost služi kao pokazatelj gradijenata saliniteta i izmjene vodenih masa između riječnih i morskih izvora.

Razdoblja s većim pogreškama predviđanja mogu odgovarati epizodama pojačane fotosintetske proizvodnje kisika tijekom dnevnih sati ili naglim događajima miješanja potaknutima plimom i osekom, vjetrom ili riječnim protokom.

Općenito, višestruki linearni regresijski model pruža i prediktivnu sposobnost i ekološki uvid povezivanjem dinamike otopljenog kisika s ključnim fizičkim i biološkim mehanizmima koji djeluju u estuarijskim područjima uzgoja školjkaša. To pokazuje kako regresijsko modeliranje može poduprijeti okolišno praćenje i sustave ranog upozoravanja na hipoksične stresne uvjete.

Praktična vježba: regresija u okolišnim znanostima

Tema: Predviđanje koncentracije klorofila-a na temelju fizičkih oceanografskih varijabli

Rad studenta: individualni

Alati:
Python (Jupyter Notebook / Google Colab), pandas, numpy, scikit-learn, matplotlib, seaborn

Skup podataka:
Copernicus Marine Service (CMEMS) – salinitet površine mora, klorofil i brzina struje.

Pozadina i motivacija

Biomasa fitoplanktona, koja se uobičajeno predstavlja koncentracijom klorofila-a (Chl-a), ključan je pokazatelj produktivnosti morskog ekosustava i kakvoće vode. Na varijacije klorofila-a utječu fizički pokretači poput saliniteta i oceanskih struja, koji utječu na dostupnost hranjivih tvari, stratifikaciju i procese miješanja.
U ovoj praktičnoj vježbi studenti primjenjuju regresijske tehnike za modeliranje koncentracije klorofila-a uporabom okolišnih prediktora izvedenih iz podataka Copernicus Marine Service (CMEMS). Zadatak pokazuje kako se jednostavna i višestruka linearna regresija mogu koristiti za kvantificiranje odnosa između bioloških i fizičkih varijabli u morskim sustavima te za potporu okolišnom praćenju i primjenama ranog upozoravanja.

Opis skupa podataka

Skup podataka izveden je iz morskih proizvoda CMEMS-a i sadrži vremenske nizove ili prostorne uzorke površinskih oceanskih uvjeta. Svaki redak odgovara jednoj točki opažanja (vrijeme i/ili lokacija) i uključuje sljedeće varijable:

Ciljna varijabla (y):
- Koncentracija klorofila-a (Chl-a, mg/m³)

Prediktorske varijable (X):
- Salinitet površine mora (PSU)
- Brzina površinske struje (m/s)

Opis zadatka

Od studenta se očekuje da izgradi i usporedi dva regresijska modela za predviđanje koncentracije klorofila-a:
1. **Jednostavna linearna regresija:**
   Chl-a = f(Salinity)
2. **Višestruka linearna regresija:**
   Chl-a = f(Salinity, Current Velocity)

Vodič korak po korak

Korak 1 – Postavljanje
- Otvorite Google Colab ili Jupyter Notebook.

- Uvezite potrebne Python biblioteke (pandas, numpy, scikit-learn, matplotlib).

Korak 2 – Učitavanje i pregled skupa podataka

- Učitajte CMEMS skup podataka iz priložene CSV datoteke.
- Prikažite prve retke skupa podataka.
- Izvijestite:
  - Broj opažanja (redaka)
  - Broj varijabli (stupaca)
  - Ciljna varijabla
  - Prediktorske varijable
- Provjerite nedostajuće vrijednosti.

Korak 3 – Priprema podataka

- Pretvorite skalirane vrijednosti u fizičke jedinice ako je potrebno.
- Uklonite nefizičke stupce ili stupce identifikatora.
- Pregledajte deskriptivnu statistiku i potencijalne odstupajuće vrijednosti.
- Podijelite skup podataka na skup za treniranje i testni skup (80/20).

Korak 4 – Model 1: jednostavna linearna regresija (osnovni model)

Trenirajte jednostavni linearni regresijski model koristeći salinitet kao jedini prediktor.
Formulacija modela:
Chl-a = f(Salinity)

- Prilagodite model na podacima za treniranje.
- Predvidite klorofil-a na testnim podacima.
- Izračunajte:
  - RMSE
  - R^2

Vizualizacije:

- Dijagram raspršenja: salinitet nasuprot klorofilu-a
- Regresijski pravac
- Dijagram predviđenih nasuprot opaženim vrijednostima
- Dijagram reziduala

Očekuje se od studenta:

Ukratko objasnite fizičko značenje regresijskog koeficijenta (kako je salinitet povezan s koncentracijom klorofila-a).

Korak 5 – Model 2: višestruka linearna regresija

Trenirajte višestruki linearni regresijski model uporabom oba prediktora:

Formulacija modela:
Chl-a = f(Salinity, Current Velocity)

- Prilagodite model na podacima za treniranje.
- Predvidite klorofil-a na testnim podacima.
- Izračunajte:
  - RMSE
  - R^2

Vizualizacije:

- Dijagram predviđenih nasuprot opaženim vrijednostima
- Dijagram reziduala
-

Očekuje se od studenta:

Interpretirajte koeficijente saliniteta i brzine struje u ekološkim terminima (miješanje, transport, stratifikacija).

Korak 6 – Usporedba modela

Usporedite dva modela:

- Koji model ima niži RMSE?
- Koji model ima viši R^2?
- Poboljšava li dodavanje brzine struje izvedbu predviđanja?

Očekuje se od studenta:

Navedite kratko objašnjenje (1–2 rečenice) zašto višestruki regresijski model ima bolju ili lošiju izvedbu od jednostavnog modela.

Korak 7 – Ekološka interpretacija

Studenti trebaju raspraviti:

- Kako salinitet utječe na fitoplankton kroz dotok slatke vode i opskrbu hranjivim tvarima.
- Kako brzina struje utječe na raspodjelu klorofila putem procesa miješanja i transporta.
- Zašto dio varijabilnosti ostaje neobjašnjen (dostupnost svjetlosti, hranjive tvari, biološki procesi).

Što predati – izvješće o vježbi

Format:
Jupyter Notebook ili Google Colab s markdown i kodnim ćelijama. Izvješće mora sadržavati:

a. Uvod

- Što se modeliralo?
- Zašto je klorofil-a važan za morske ekosustave?

b. Opis podataka

- Broj uzoraka
- Prediktorske varijable
- Ciljna varijabla
- Postupanje s nedostajućim vrijednostima

c. Metode

- Jednostavna linearna regresija
- Višestruka linearna regresija
- Omjer podjele na skup za treniranje i testni skup
- Evaluacijske metrike (RMSE, R^2)

d. Rezultati

- RMSE i R^2 za oba modela
- Grafički prikazi (dijagram raspršenja, predviđeno nasuprot opaženom, reziduali)
- Kratka usporedba modela

e. Rasprava

- Interpretacija koeficijenata
- Okolišno značenje pogrešaka
- Ograničenja linearne regresije

f. Zaključak

- Sažetak naučenoga o regresijskom modeliranju uporabom morskih okolišnih podataka

Vrednovanje

PRAKTIČNI REGRESIJSKI ZADATAK

Praktični regresijski zadatak studentima pruža priliku za praktičnu primjenu regresijskih modela (jednostavne i višestruke linearne regresije) na stvarnom okolišnom skupu podataka dobivenom iz Copernicus Marine Service (CMEMS). Namijenjen je procjeni sposobnosti studenata da implementiraju, validiraju i interpretiraju regresijske modele za potporu okolišnom odlučivanju.

KVIZ

Kviz će biti osmišljen za procjenu konceptualnog razumijevanja i primijenjenog zaključivanja stečenog kroz predavanje i praktičnu vježbu o regresijskim tehnikama u okolišnim znanostima. Provjeravat će razumiju li studenti jednostavnu nasuprot

višestrukoj linearnoj regresiji te ulogu regresijskih modela u okolišnom praćenju i predviđanju.

KRATKO PISANO IZVJEŠĆE

Kratko pisano izvješće završni je korak u praktičnom regresijskom zadatku, koji studentima omogućuje sintetiziranje, interpretiranje i priopćavanje nalaza iz vježbe regresijskog modeliranja.

Kriteriji vrednovanja
1. KVIZ

Pitanje 1
Koja je vrsta problema strojnog učenja predviđanje koncentracije otopljenog kisika iz okolišnih varijabli poput temperature i dubine?
a) Klasifikacija
b) Regresija
c) Klasteriranje
d) Otkrivanje anomalija

Pitanje 2
Koja varijabla predstavlja ciljnu (zavisnu) varijablu u primijenjenom primjeru ove lekcije?
a) Temperatura vode
b) Električna vodljivost
c) Koncentracija otopljenog kisika
d) Dubina

Pitanje 3
Koja od sljedećih tvrdnji najbolje definira regresijski problem u strojnom učenju?
a) Predviđanje kategorijske oznake
b) Predviđanje kontinuirane numeričke vrijednosti
c) Grupiranje sličnih uzoraka
d) Smanjenje broja varijabli

Pitanje 4
Koja je glavna razlika između jednostavne linearne regresije i višestruke linearne regresije?
a) Jednostavna regresija je nenadzirana
b) Višestruka regresija koristi više od jedne prediktorske varijable
c) Jednostavna regresija uvijek je točnija
d) Višestruka regresija ne može se evaluirati

Pitanje 5
Zašto se regresija smatra metodom nadziranog učenja?

a) Zato što ne zahtijeva označene podatke
b) Zato što model uči bez ciljne varijable
c) Zato što se model trenira uporabom poznatih parova ulaz–izlaz
d) Zato što radi samo na malim skupovima podataka

Pitanje 6
Što mjeri korijen srednje kvadratne pogreške (RMSE)?
a) Udio objašnjene varijance
b) Prosječnu pogrešku predviđanja u istim jedinicama kao ciljna varijabla
c) Broj točnih predviđanja
d) Korelaciju između varijabli

Pitanje 7
Koja je glavna svrha podjele skupa podataka na skup za treniranje i testni skup?
a) Povećati složenost modela
b) Ukloniti odstupajuće vrijednosti
c) Evaluirati izvedbu modela na neviđenim podacima
d) Smanjiti vrijeme računanja

Pitanje 8
Koja situacija upućuje na to da bi nelinearni regresijski model mogao biti prikladniji od linearnog?
a) Odnos između varijabli tvori pravac
b) Podaci pokazuju pragove ili zakrivljene obrasce odziva
c) U podacima nema šuma
d) Dostupan je samo jedan prediktor

Pitanje 9
Koje informacije pruža dijagram predviđenih nasuprot opaženim vrijednostima?
a) Prikazuje distribuciju prediktora
b) Vizualizira koliko se predviđanja podudaraju s izmjerenim vrijednostima
c) Otkriva nedostajuće podatke
d) Automatski odabire najbolje prediktore

Pitanje 10
Zašto su regresijski modeli korisni u okolišnim znanostima?
a) Zamjenjuju fizička mjerenja
b) Pomažu u predviđanju i interpretaciji kontinuiranih okolišnih varijabli
c) Uklanjaju nesigurnost u opažanjima
d) Rade samo za laboratorijske podatke

2. PRAKTIČNI REGRESIJSKI ZADATAK

Priprema i predobrada podataka

• Ispravno učitavanje i pregled skupa podataka (oblik, prediktori, ciljna varijabla).
• Primjereno postupanje s nedostajućim podacima i pretvorbom jedinica kada je potrebno.
• Transparentan i reproducibilan tijek predobrade (kôd je čist, a koraci jasno objašnjeni).

Implementacija modela

• Točno postavljanje i treniranje najmanje dvaju regresijskih modela:
  jednostavna linearna regresija
  višestruka linearna regresija
• Primjerena uporaba podjele na skup za treniranje i testni skup.
• Kôd je organiziran, a parametri su eksplicitno navedeni.

Evaluacija i dijagnostika modela

- Ispravan izračun i izvješćivanje o regresijskim metrikama:
  RMSE
  R^2
  • Logična usporedba modela (npr. „višestruka regresija poboljšala je R^2 u usporedbi s jednostavnom regresijom").
  • Uporaba dijagnostičkih grafova (predviđeno nasuprot opaženom, dijagrami reziduala).
  • Svijest o prenaučenosti i podnaučenosti.

Odabir modela i obrazloženje

• Jasno obrazloženje koji se model preferira.
• Razumno objašnjenje sposobnosti generalizacije modela na temelju rezultata na testnim podacima.

Vizualizacija i prikaz rezultata

• Slike (dijagrami raspršenja, regresijske prilagodbe, dijagrami reziduala) jasno su označene.
• Vizualni prikazi podupiru interpretaciju ponašanja modela.

Povezanost s okolišnim kontekstom

• Student objašnjava kako regresijski modeli mogu poduprijeti okolišno praćenje i sustave ranog upozoravanja.

3. KRATKO PISANO IZVJEŠĆE

Jasnoća i struktura izvješća

• Izvješće je logično organizirano (Uvod → Podaci → Metode → Rezultati → Rasprava → Zaključak).
• Koraci radnog tijeka vidljivi su u bilježnici.
• Jezik je precizan i sažet.

Interpretacija rezultata i rasprava

• Regresijski rezultati (RMSE, R^2, reziduali) točno su prikazani i smisleno interpretirani.
• Student promišlja zašto je višestruka regresija imala bolju izvedbu (ili nije) od jednostavne regresije.

Povezanost s ekološkim i okolišnim kontekstom

• Nalazi su povezani sa stvarnim okolišnim značenjem (npr. predviđanje varijabilnosti klorofila za praćenje morskih ekosustava).
• Student pokazuje razumijevanje načina na koji podatkovno vođena regresija može pridonijeti okolišnom odlučivanju.

Kvaliteta vizualnog prikaza

• Slike su jasno označene i na njih se upućuje u tekstu.
• Vizualni prikazi pomažu objasniti ponašanje modela, a ne služe samo kao ukras izvješća.

Komunikacija i stil

• Pisanje je jasno, objektivno i profesionalno.
• Terminologija strojnog učenja i okolišna terminologija koriste se ispravno (prediktor, cilj, RMSE, reziduali, koeficijent).
• Student navodi ograničenja (pojednostavljeni model, linearne pretpostavke, ograničen broj prediktora).

Predloženo ponderiranje

• Kviz: 20 %
• Praktični regresijski zadatak (bilježnica/kôd): 50 %
• Kratko pisano izvješće: 30 %

Reference / dodatna literatura

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer.
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
- Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.
- Copernicus Marine Service (CMEMS) Data Access Portal
- Članci otvorenog pristupa (potpuni popis bit će dostupan u e-sustavu kolegija)