Tema: Regresijske tehnike za ekološko modeliranje Predmet: Strojno učenje za okoljske aplikacije

Opis učne enote:

Ta učna enota študente seznani z regresijo kot osrednjim pristopom strojnega učenja za modeliranje in napovedovanje zveznih okoljskih spremenljivk. Poudarja tako teoretične kot uporabne vidike ter razlikuje med linearnimi in nelinearnimi regresijskimi tehnikami, pomembnimi za okoljske in ekološke raziskave. Študenti bodo spoznali celoten potek regresijskega modeliranja, vključno s pripravo podatkov, inženirstvom značilk, učenjem modela, vrednotenjem z RMSE ter omejevanjem prenaučenja in podprileganja.
Izbrani primer uporabe se osredotoča na napovedovanje koncentracije raztopljenega kisika na podlagi temperature vode in povezanih okoljskih spremenljivk na območju gojenja školjk v estuariju. Primer ponazarja, kako lahko regresijski modeli podpirajo ekološko spremljanje v praksi in sisteme zgodnjega opozarjanja na stresne razmere v vodnem okolju.

Učni izidi:

Po zaključku te enote bodo študenti znali:
1. Opredeliti regresijske probleme in jih razlikovati od drugih nalog strojnega učenja.
2. Pojasniti načela enostavne in večkratne linearne regresije ter logiko vrednotenja z RMSE.
3. Razviti in ovrednotiti regresijske modele z uporabo okoljskih podatkovnih naborov.
4. Prepoznati in obravnavati pogoste težave, kot so šum, multikolinearnost, podprileganje in prenaučenje, pri regresijskih problemih z ekološkimi podatki.
5. Interpretirati rezultate modela v ekološkem in okoljskem kontekstu.

Vsebina predavanja

Uvod

Regresijske tehnike so bistvena orodja za modeliranje zveznih okoljskih spremenljivk, kot so temperatura, slanost in koncentracije onesnaževal. Ta razdelek pojasnjuje razliko med regresijskimi in klasifikacijskimi nalogami, predstavi pojem korena povprečne kvadratne napake (RMSE) ter obravnava pomen regresijskih modelov za razumevanje in napovedovanje ekoloških pojavov.

Kontekst in motivacija

Regresijske tehnike so temeljne pri modeliranju številnih okoljskih pojavov z zveznim odzivom. V ekoloških in okoljskih znanostih pogosto želimo napovedati zvezno spremenljivko (npr. koncentracijo onesnaževala, biomaso vrste, rečni pretok ali temperaturno anomalijo) na podlagi množice napovednih spremenljivk (meteoroloških spremenljivk, značilnosti krajine, značilnosti porečja, rabe tal itd.). Regresija tako povezuje opazovalne oziroma okoljske podatke s pojasnjevalnim in napovednim

modeliranjem — ter je na stičišču statistične ekologije, hidrologije, klimatologije in strojnega učenja.
Primeri regresijskih problemov v ekologiji in okoljskih znanostih:

- Ocenjevanje koncentracije dušikovega dioksida (NO_2) ali trdnih delcev ($PM_{2.5}$) v mestnih oziroma industrijskih okoljih na podlagi meteoroloških vhodnih podatkov, spremenljivk prometa in rabe tal ter atmosferskih razmer.
- Modeliranje številčnosti vrste (ali biomase) skozi čas ali prostor glede na okoljske dejavnike, kot so temperatura, padavine, hranila v tleh, drobljenje habitatov in motnje zaradi človekovega delovanja.
- Napovedovanje hidroloških odzivov, kot je dnevni rečni pretok, na podlagi padavin, temperature zraka, vlažnosti, predhodne vlažnosti tal in značilnosti porečja.
- Napovedovanje podnebnih anomalij (npr. temperaturnih anomalij) ali odzivov ekosistemov (tokov ogljika, emisij toplogrednih plinov) z uporabo zveznih regresijskih okvirov.

Regresijske naloge v primerjavi s klasifikacijskimi nalogami

Bistveno je razlikovati regresijo od klasifikacije, saj obe spadata v družino problemov nadzorovanega učenja, vendar imata različne napovedne cilje.
Pri klasifikaciji je ciljna spremenljivka y kategorična — predstavlja diskretne oznake ali razrede (npr. »onesnaženi« v primerjavi s »čistimi« vzorci vode ali »prisotnost« v primerjavi z »odsotnostjo« določene vrste). Nasprotno pa je pri regresiji y zvezna spremenljivka, kar pomeni, da se lahko njene vrednosti spreminjajo na številčni lestvici brez vnaprej določenih razrednih meja (npr. pri napovedovanju natančne koncentracije nitrata v podzemni vodi ali gostote biomase na gozdni ploskvi).
Pri klasifikaciji nas zanimajo odločitvene meje — kam med znane kategorije uvrstiti nov vzorec. Pri regresiji nas zanima številčna natančnost — kako blizu je napovedana vrednost dejanski zvezni meritvi.
Če to razliko poudarimo že na začetku predavanja, študenti lažje razumejo, zakaj je regresijska analiza bistvena na okoljskih področjih, kjer so številne ciljne spremenljivke po naravi zvezne in zahtevajo ne le pravilno oceno trenda, temveč tudi razumevanje negotovosti in variance napovedi.

Zakaj je regresija problem nadzorovanega učenja

Regresija spada v družino metod nadzorovanega učenja, ker učenje modela zahteva označene podatke — pare vhodov in znanih ciljnih vrednosti (X_i,y_i)
Model se »uči« s prilagajanjem svojih parametrov, da zmanjša neskladje med napovedanimi vrednostmi $\hat{y}_i = f(X_i)$ in dejanskimi opaženimi izidi, y_i.
»Nadzor« izhaja iz teh znanih izidov: učni proces usmerjajo povratne informacije iz signala napake, tj. podatek o tem, kako napačne so napovedi modela v primerjavi z dejanskim stanjem. Cilj je z zaporednimi približki zmanjšati napako napovedi, pri čemer proces usmerja ciljna funkcija (pogosto povprečna kvadratna napaka).

Uporabnost RMSE pri okoljski in ekološki regresiji

Ključni metodološki vidik regresijskega modeliranja je izbira metrike vrednotenja — kvantitativnega merila, kako dobro se napovedi modela ujemajo z opazovanimi podatki. Koren povprečne kvadratne napake (RMSE) je ena najpogosteje uporabljenih metrik pri okoljskem in ekološkem modeliranju ter je opredeljen kot:

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

Ta metrika vrednotenja je skoraj samoumevna; v tem smislu bomo postopek obravnavali od znotraj navzven. Kvadrirane napake pomenijo, da je napaka opredeljena kot preprosto odštevanje prave in napovedane vrednosti, rezultat pa kvadriramo, s čimer preprečimo izničenje pozitivnih in negativnih odklonov. Za določitev skupnega učinka vseh napovedanih vrednosti te kvadrirane rezultate seštejemo. Nazadnje nas zanima povprečni prispevek vsakega podatkovnega primera, ki ga izračunamo tako, da skupni prispevek delimo s številom primerov. Na koncu korenjenje povrne prvotne merske enote.

- RMSE zagotavlja razumljivo merilo v istih enotah kot ciljna spremenljivka, kar je ključnega pomena za okoljske znanstvenike, ki morajo rezultate sporočati deležnikom in oblikovalcem politik.
- Odraža tako pristranskost (sistematično napako) kot varianco (razpršenost ostankov) in s tem zajema splošno napovedno zanesljivost modela.
- Metriko vrednotenja RMSE je preprosto vključiti v optimizacijske algoritme, ki se uporabljajo pri strojnem učenju, s čimer je omogočeno samodejno prilagajanje parametrov.

Glavne prednosti RMSE so:

- RMSE je za študente jasno in intuitivno izhodišče: manjše vrednosti pomenijo večjo natančnost napovedi.
- Kvadratna kazen nazorno ponazarja, zakaj so skrajne napake napovedi nesorazmerno močneje kaznovane — to pa se dobro ujema z razmišljanjem o ekoloških tveganjih.
- Primerjava RMSE z drugimi merili (npr. povprečno absolutno napako ali povprečno napako pristranskosti) spodbuja kritični razmislek o izbiri in interpretaciji metrik v okoljskih raziskavah.

Pregled tehnologije

Regresijskega modeliranja v ekoloških in okoljskih znanostih ni mogoče obravnavati zgolj kot čisto matematični postopek, temveč kot matematično utemeljen pristop z izrazito

tehnološko komponento, ki raznovrstne okoljske podatke pretvori v razumljive in preverljive modele.

Da bi bil ta proces študentom dostopnejši, bodo na reprezentativnih podatkovnih naborih s področja ekologije in okoljskih znanosti uporabljeni različni regresijski modeli (linearna regresija, regresijska drevesa, SVM za regresijo in ANN).
Z vključevanjem odprtih podatkovnih virov, sodobnih knjižnic za obdelavo podatkov in tehnik vizualizacije bodo obravnavana temeljna načela tehnoloških in metodoloških vidikov regresijskega modeliranja v strojnem učenju za okoljske znanosti.

Tehnološko povezovanje in podatkovni nabori

Najprej se za prikaz osnovnih načel enostavne in večkratne linearne regresije v nadzorovanem in lahko razumljivem okolju uporabi sintetični podatkovni nabor. Tako se lahko študenti osredotočijo na razumevanje vedenja modela brez zapletenosti šuma in manjkajočih podatkov iz resničnega sveta.

V drugem delu se uporabi resnični morski podatkovni nabor storitve Copernicus Marine Environment Monitoring Service (CMEMS), ki vključuje slanost morske površine, koncentracijo klorofila in hitrost tokov. Na teh podatkih se uporabi isti potek regresijskega modeliranja ter ovrednotita uspešnost modela in njegov ekološki pomen.

Vse analize se izvajajo v okolju Google Colab z uporabo standardnih knjižnic za Python:

- **pandas** in **NumPy** za nalaganje in čiščenje podatkov ter numerične operacije,
- **scikit-learn** za izvedbo regresijskih modelov (enostavne linearne regresije in večkratne linearne regresije) ter izračun metrik vrednotenja (RMSE in R^2),
- **Matplotlib** za vizualizacijo porazdelitev podatkov, prileganja regresijskih modelov in diagnostike modelov.

Enostavna linearna regresija se uporablja kot izhodiščni model za opis odnosa med eno okoljsko napovedno spremenljivko (npr. temperaturo) in zvezno odzivno spremenljivko (raztopljenim kisikom). Njeni glavni prednosti sta razumljivost in zmožnost ponazoritve temeljnih regresijskih pojmov.

Večkratna linearna regresija ta pristop razširi z vključitvijo dodatnih napovednih spremenljivk (npr. globine in električne prevodnosti), da bi bolje predstavila fizikalne in ekološke procese, ki vplivajo na spremenljivost raztopljenega kisika. Ta model prikazuje, kako lahko dodajanje ustreznih okoljskih spremenljivk izboljša pojasnjevalno moč in obenem ohrani preglednost modela.

Enostavna linearna regresija in večkratna linearna regresija: koncept, metoda in prikaz

Enostavna linearna regresija povezuje eno napovedno spremenljivko z eno odzivno spremenljivko:

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

- Pri tem ε predstavlja nepojasnjeno spremenljivost ali merilni šum.

Pri večkratni regresiji imamo več neodvisnih spremenljivk, $X_{i,}$ in eno odvisno spremenljivko.
Cilj je določiti β_0 in β_1 (oziroma β_k pri večkratni regresiji) tako, da model čim bolj zmanjša povprečno kvadratno napako med opazovanimi in napovedanimi vrednostmi.
Ta učna enota prikazuje enostavno in večkratno linearno regresijo z uporabo meritev okoljskih senzorjev z območja gojenja školjk v estuariju. V praktičnem primeru je koncentracija raztopljenega kisika (mg/L) modelirana kot funkcija temperature vode in dodatnih napovednih spremenljivk (npr. globine in prevodnosti). Ta zasnova ponazarja, kako lahko linearni modeli zagotovijo razumljive izhodiščne odnose in kako lahko modeli z več napovednimi spremenljivkami izboljšajo pojasnjevalno moč v ekoloških sistemih.

Koraki:

Priprava podatkov:
Naložite podatkovni nabor okoljskih senzorjev z meritvami temperature vode, koncentracije raztopljenega kisika in neobveznimi napovednimi spremenljivkami, kot sta globina in električna prevodnost. Vse skalirane vrednosti pretvorite v fizikalne enote, preverite, ali podatkovni nabor vsebuje manjkajoče vrednosti in osamelce, ter podatke razdelite na učno in testno množico.

Analiza:
Prilagodite dva regresijska modela:
(i) model enostavne linearne regresije, ki temperaturo uporablja kot edino napovedno spremenljivko raztopljenega kisika (Temp → DO), in
(ii) model večkratne linearne regresije, ki temperaturo, globino in prevodnost uporablja kot napovedne spremenljivke raztopljenega kisika (Temp + Depth + Conductivity → DO).

Vrednotenje:
Oba modela ovrednotite z uporabo korena povprečne kvadratne napake (RMSE) in koeficienta determinacije (R^2). Vizualizirajte napovedane vrednosti v primerjavi z opazovanimi ter analizirajte grafikone ostankov, da ocenite sistematične napake in predpostavke modela.

Izpopolnjevanje modela:
Preučite multikolinearnost med napovednimi spremenljivkami in obravnavajte njen vpliv na stabilnost modela. Kot razširitev uvedite tehnike regularizacije, kot sta regresiji Ridge in Lasso, da zmanjšate prenaučenje in izboljšate robustnost modela.

Interpretacija:
Regresijske koeficiente in vzorce ostankov interpretirajte z ekološkega in okoljskega vidika ter upoštevajte fizikalne in biološke mehanizme, kot so topnost kisika, fotosintetska aktivnost in procesi mešanja vode.

Nelinearna regresija: koncept, metoda in prikaz

Nelinearna regresija razširja zamisel linearne regresije na odnose, ki jih premica ne more ustrezno zajeti. Pri številnih ekoloških in okoljskih procesih se odzivna spremenljivka glede na napovedne spremenljivke ne spreminja linearno — namesto tega se lahko pojavijo pragovi, nasičenje, eksponentna rast ali ciklični vzorci.
Formalno lahko nelinearno regresijo izrazimo kot:

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

- Pri tem je f(·) nelinearna funkcija napovednih spremenljivk Xi in parametrov β, ε_i pa predstavlja naključno napako.

Pogoste funkcijske oblike vključujejo polinomske, eksponentne, logaritemske ali potenčne transformacije ter modele, ki temeljijo na jedrih, drevesih ali ANN, kadar je potrebna neparametrična prilagodljivost.

Cilj ostaja enak kot pri linearni regresiji: čim bolj zmanjšati razliko med opazovanimi (y_i) in napovedanimi ($\hat{y}_i$) vrednostmi, običajno z uporabo RMSE.
Vendar ima lahko pri nelinearni regresiji prilagojena krivulja veliko možnih oblik, odvisno od ekološke teorije, vedenja podatkov in transformacij značilk.

Ta prikaz se osredotoča na modeliranje koncentracije raztopljenega kisika (DO) kot funkcije temperature vode z uporabo sintetičnega podatkovnega nabora z območja gojenja školjk v estuariju. Polinomsko regresijo je mogoče uporabiti za ponazoritev, kako nelinearni modeli zajamejo ukrivljene vzorce odziva, kadar linearne predpostavke ne zadostujejo.

Koraki:

Priprava podatkov:

Pridobite zvezne meritve temperature vode in koncentracije raztopljenega kisika iz sistema za spremljanje gojenja školjk v estuariju. Resnični podatkovni nabor dopolnite s sintetičnimi podatkovnimi nabori, zasnovanimi tako, da za didaktične namene poudarijo nelinearne vzorce. Obravnavajte manjkajoče vrednosti, normalizirajte napovedne spremenljivke ter podatkovni nabor razdelite na učno, validacijsko in testno množico.

Analiza:

Uporabite polinomsko regresijo, da zajamete nelinearne odnose med temperaturo vode in koncentracijo raztopljenega kisika. Polinomski model primerjajte z regresijo s podpornimi vektorji (SVR), regresijo z naključnim gozdom in umetnimi nevronskimi mrežami (ANN), da raziščete zapletenejše nelinearne strukture.

Vrednotenje:

Izračunajte RMSE za vse modele ter z dvodimenzionalnimi grafikoni (temperatura glede na raztopljeni kisik) in neobveznimi tridimenzionalnimi grafikoni, kadar so vključene dodatne napovedne spremenljivke, vizualizirajte napovedane vrednosti v primerjavi z opazovanimi. Preučite nelinearne odzivne krivulje, da odkrijete ekološke pragove, kot so vrednosti temperature, nad katerimi raztopljeni kisik hitro upade.

Optimizacija modela:

Z 10-kratnim navzkrižnim preverjanjem izberite stopnjo polinoma in prilagodite hiperparametre jedra ali ANN. Uspešnost modelov na testni množici primerjajte tako, da ocenite napako posploševanja in razumljivost modela.

Interpretacija:

Ugotovitve modela povežite s procesi v vodnem ekosistemu in obravnavajte njihov pomen za upravljanje gojenja školjk, zgodnje opozarjanje na hipoksične razmere in stres v estuarijskih okoljih, ki ga povzročajo podnebne spremembe.

Metrike vrednotenja regresijskih problemov

Metrike vrednotenja količinsko opredelijo napovedno uspešnost modela, diagnostika pa ocenjuje, kako dobro se struktura modela ujema s temeljnim ekološkim procesom. Cilj je razumeti ne le, *kako natančen* je model, temveč tudi, *zakaj* deluje tako, kot deluje — kje je uspešen, kje odpove in kaj te napake pomenijo.

Najpogosteje uporabljena merila pri ekološki regresiji so:

Koren povprečne kvadratne napake (RMSE) – že omenjeno:

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

Občutljiva na velike napake — idealna, kadar imajo skrajni odkloni znatne ekološke posledice (npr. prenizka ocena poplavnega pretoka ali previsoka napoved ravni onesnaževal).

Povprečna absolutna napaka (MAE):

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

Manj občutljiva na osamelce in omogoča vpogled v povprečni odklon.

Koeficient determinacije (R^2):

Ta mera ni zanesljiva za podatke časovnih vrst ali prostorske podatke. V drugih primerih meri delež variance y, ki ga pojasni model, kar pomaga pri oceni njegove celotne pojasnjevalne moči.

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

- SSres → $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ → Vsota kvadratov ostankov (nepojasnjena varianca)
- SS_{tot} → $\sum_{i=1}^{n}(y_i - \overline{y})^2$ → Skupna vsota kvadratov (skupna varianca v opazovanih podatkih)
- y_i = opazovana vrednost
- $\hat{y}_i$ = napoved modela
- $\acute{y}$ = povprečje vseh opazovanih vrednosti
- n = število vzorcev

Koeficient učinkovitosti Nash–Sutcliffe (NSE):

Pogosta v hidrologiji; izraža napovedno uspešnost glede na povprečje opazovanj.

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

pri čemer Q_{obs} označuje opazovane vrednosti (v tem primeru pretoka Q), Q_{mod} simulirane vrednosti iste spremenljivke, $\overline{Q}$ pa predstavlja povprečje opazovanih vrednosti izbrane spremenljivke.

Praktični primer uporabe: regresijsko modeliranje raztopljenega kisika na območju gojenja školjk v estuariju

Ozadje in motivacija

Na gojišču školjk v estuarijskem oziroma rečnem okolju se z in situ senzorji neprekinjeno merita temperatura vode in raztopljeni kisik (DO, mg/L). Raztopljeni kisik neposredno vpliva na dihanje in stres školjk, zato je ključni kazalnik zdravja ekosistema in trajnosti akvakulture.

Cilj tega primera uporabe je izdelati **regresijski model**, ki oceni DO na podlagi temperature vode. Ta model je hiter kazalnik morebitnih neugodnih razmer: **ko se temperatura zviša, se pričakuje zmanjšanje koncentracije kisika.** Zgodnje odkrivanje takih razmer gojiteljem omogoča učinkovitejše spremljanje in zmanjševanje tveganj ekološkega stresa.

Študenti primerjajo napovedi modela z izmerjenimi vrednostmi in analizirajo primere, v katerih model odpove. Tako se intuitivno seznanijo z zamislijo, da je mogoče regresijske modele izboljšati z vključevanjem dodatnih okoljskih napovednih spremenljivk, kot so nasičenost s kisikom (%), globina, prevodnost in motnost.

Enostavna linearna regresija

1. korak – razumevanje podatkov

Opis podatkovnega nabora in spremenljivk

Podatkovni nabor, uporabljen v tej praktični študiji primera, izvira iz neprekinjenega okoljskega spremljanja na območju gojenja školjk v estuariju. Meritve so bile zbrane z in situ senzorji in vključujejo temperaturo vode ter koncentracijo raztopljenega kisika, skupaj z dodatnimi okoljskimi spremenljivkami za razširitev modela.
Podatkovni nabor vsebuje več tisoč opazovanj in v primarnih spremenljivkah, uporabljenih za modeliranje, nima manjkajočih vrednosti. Vse vrednosti so bile podane v skaliranih enotah in med predobdelavo pretvorjene v fizikalne enote (°C in mg/L).

Ime podatkovnega nabora:

Sistem okoljskega spremljanja in situ (sintetični podatkovni nabor, pripravljen za izobraževalne namene)
Regresijski problem je opredeljen tako:

Ciljna spremenljivka (y):
Koncentracija raztopljenega kisika (DO, mg/L)

Primarna napovedna spremenljivka (X):
Temperatura vode (Temp, °C)

Dodatne napovedne spremenljivke (za večkratno regresijo):

- Globina (m)
- Električna prevodnost (µS/cm)
- (neobvezno) nasičenost s kisikom (%)
- (neobvezno) motnost

Te spremenljivke so bile izbrane zaradi njihovega fizikalnega in ekološkega pomena. Temperatura neposredno vpliva na topnost kisika, globina in prevodnost pa odražata zgradbo vodnega stolpca in procese mešanja, ki vplivajo na razpoložljivost kisika.
V regresijskih modelih niso bili uporabljeni vsi razpoložljivi atributi podatkovnega nabora. Spremenljivke, kot so identifikatorji zapisov in nefizikalne kode, so bile izključene, da bi se izognili šumu in prenaučenju ter ohranili razumljivost rezultatov regresije.

Ta problem predstavlja **nalogo nadzorovane regresije z zvezno ciljno spremenljivko**.

2. korak – pregled podatkov

Podatkovni nabor okoljskih senzorjev je bil naložen s knjižnico Pandas. Vse skalirane meritve so bile pretvorjene v fizikalne enote (°C za temperaturo in mg/L za raztopljeni kisik), da bi omogočili fizikalno interpretacijo regresijskih modelov.

Podatkovni nabor je bil uvožen iz datoteke CSV z vrednostmi, ločenimi s podpičjem, nato pa predobdelan s pretvorbo skaliranih vrednosti v fizikalne enote (Slika 1). Predogled prvih opazovanj potrjuje pravilen uvoz podatkov in pripravi podatkovni nabor za regresijsko analizo (Slika 2).

```
# 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 Koda v Pythonu za nalaganje podatkovnega nabora okoljskih senzorjev iz datoteke CSV z vrednostmi, ločenimi s podpičjem, in pretvorbo skaliranih vrednosti temperature ter raztopljenega kisika v fizikalne enote (°C in 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 Predogled prvih opazovanj po pretvorbi enot, ki potrjuje pravilen uvoz podatkov in fizikalno smiselnost vrednosti temperature in raztopljenega kisika.

V primarnih spremenljivkah, uporabljenih za regresijsko modeliranje, ni bilo zaznanih manjkajočih vrednosti (Slika 3, Slika 4).

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

Slika 3 Koda Python, uporabljena za preverjanje manjkajočih vrednosti v spremenljivkah temperature in raztopljenega kisika.

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

Slika 4 Izhod, ki kaže, da v glavnih spremenljivkah, uporabljenih za regresijsko modeliranje, ni manjkajočih vrednosti.

Opisna statistika je bila uporabljena za pregled razpona in spremenljivosti vrednosti temperature in raztopljenega kisika ter za prepoznavanje morebitnih osamelcev (Slika 5, Slika 6).

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

Slika 5 Koda Python za izračun opisne statistike (število, povprečje, standardni odklon in razpon) za spremenljivke temperature in raztopljenega kisika.

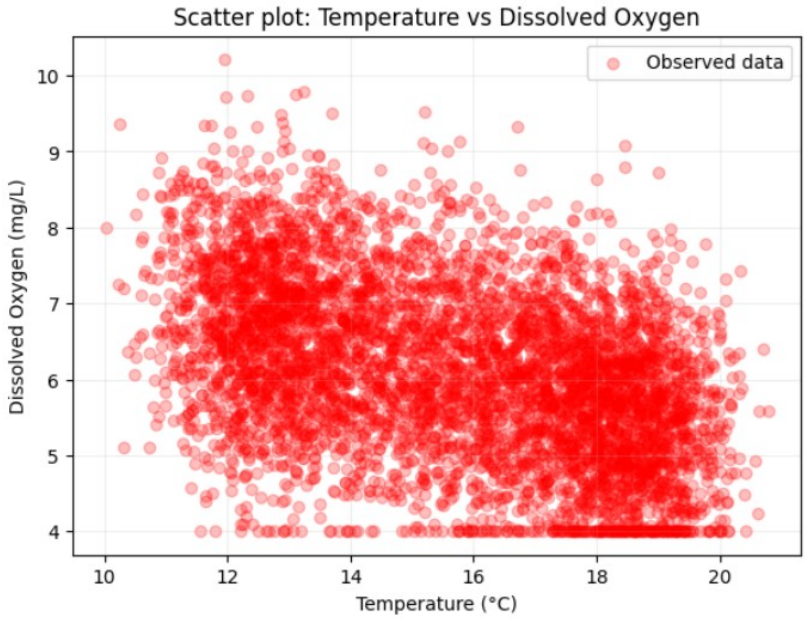
	Temp_deg_C	HDO_mg/l
count	5359.000000	5359.000000
mean	15.679737	6.104551
std	2.559164	1.119456
min	10.020000	4.000000
25%	13.285000	5.305000
50%	15.890000	6.070000
75%	17.990000	6.880000
max	20.790000	10.210000

Slika 6 Zbirna statistika vrednosti temperature in raztopljenega kisika, uporabljena za oceno razpona in spremenljivosti podatkov.

Za raziskovanje povezave med temperaturo vode in koncentracijo raztopljenega kisika je bil izdelan razsevni diagram (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 Koda Python za prikaz povezave med temperaturo (°C) in raztopljenim kisikom (mg/L) z razsevnim diagramom.



Slika 8 Razsevni diagram raztopljenega kisika (mg/L) glede na temperaturo vode (°C)

Razsevni diagram kaže jasen negativen trend: količina raztopljenega kisika se z naraščanjem temperature zmanjšuje. Ta vzorec je skladen s fizikalnimi načeli topnosti plinov v vodi. Odsotnost izrazite ukrivljenosti ali učinkov praga nakazuje, da je model linearne regresije primeren kot prvi približek. Opažena približno linearna povezava med temperaturo in raztopljenim kisikom podpira uporabo linearne regresije kot izhodiščnega modela.

3. korak – Priprava podatkovne množice za regresijski model

Za regresijsko modeliranje so bile iz podatkovne množice izbrane ustrezne napovedne spremenljivke in ciljna spremenljivka.
Kot ciljna spremenljivka je bila določena koncentracija raztopljenega kisika (DO), pri čemer je bila pri modelu enostavne linearne regresije napovedna spremenljivka temperatura vode, pri modelu večkratne linearne regresije pa temperatura vode, globina in električna prevodnost.

Podatkovna množica je bila razdeljena na učno in testno podmnožico v razmerju 80/20. Učna množica je bila uporabljena za prilagajanje regresijskih modelov, testna množica pa je bila rezervirana za neodvisno vrednotenje uspešnosti.
Ta ločitev zagotavlja nepristransko oceno sposobnosti posploševanja modela.

4. korak – Izhodiščni linearni model (samo za primerjavo)

Učenje modela

Podatkovna množica je bila razdeljena na učno in testno podmnožico (razdelitev 80/20). Model enostavne linearne regresije je bil prilagojen z uporabo temperature kot edine napovedne spremenljivke (Slika 9).

```python
# 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 Koda Python za učenje modela enostavne linearne regresije z uporabo temperature vode kot napovedne spremenljivke koncentracije raztopljenega kisika, vključno z razdelitvijo na učno in testno množico ter napovedovanjem.

Enačba modela in uspešnost

Uspešnost modela je bila ovrednotena z RMSE in R^2 (Slika 10, Slika 11).

```python
# 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 Koda Python za vrednotenje modela linearne regresije in prikaz regresijske črte

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

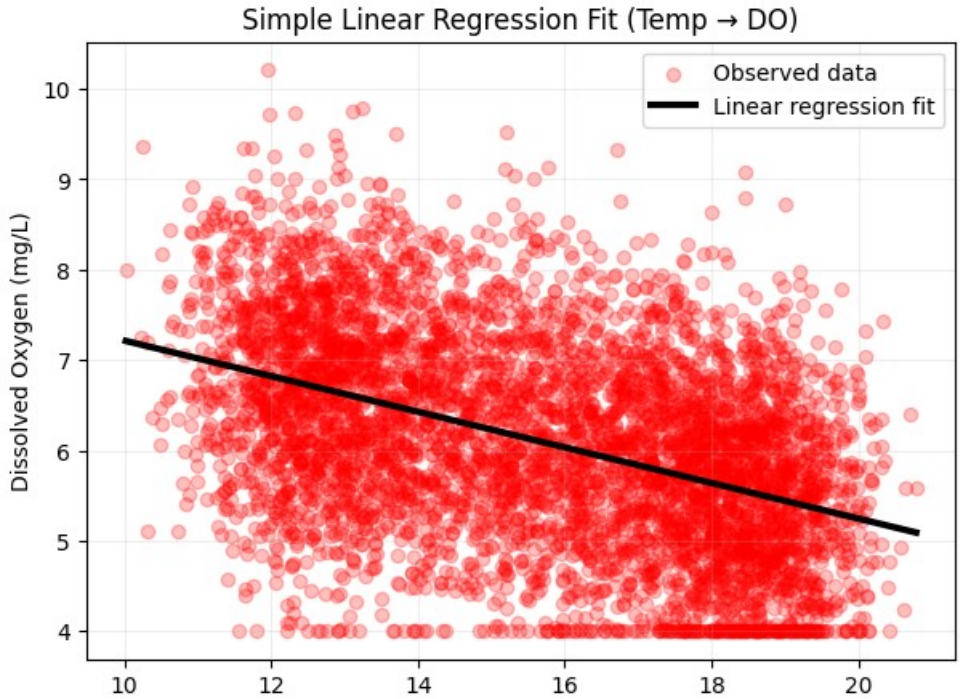
Slika 11 Izhod, ki prikazuje regresijsko enačbo in metrike uspešnosti (RMSE in R^2) za model enostavne linearne regresije

Za vizualno oceno prilagojene linearne povezave je bila regresijska črta narisana skupaj z opazovanimi podatkovnimi točkami (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 Koda Python za risanje črte linearne regresije in opazovanih podatkov



Slika 13 Prileganje linearne regresije med temperaturo vode in koncentracijo raztopljenega kisika

Regresijski koeficient potrjuje negativno povezavo med temperaturo in koncentracijo raztopljenega kisika, kar je skladno s fizikalnim načelom, da se topnost kisika z naraščanjem temperature vode zmanjšuje.

Vrednost RMSE 0,964 mg/L kaže povprečno napako napovedi modela, vrednost R^2 0,204 pa nakazuje, da temperatura sama pojasni približno 20 % variance koncentracije raztopljenega kisika.

Diagnostika modela

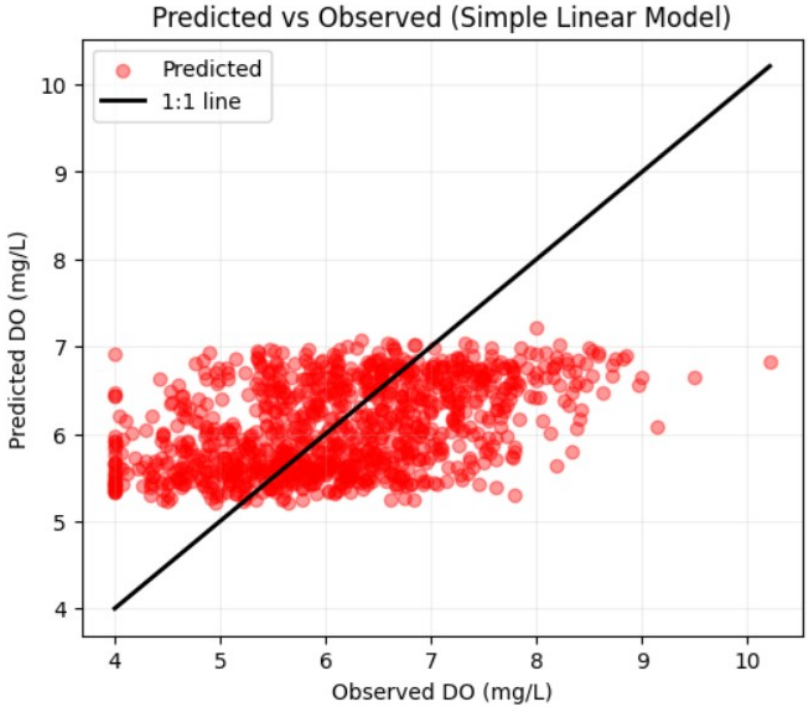
Napovedane vrednosti v primerjavi z opazovanimi

Izdelan je bil graf napovedanih vrednosti glede na opazovane, da bi preverili, kako dobro se napovedi modela ujemajo z izmerjenimi vrednostmi raztopljenega 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 Koda Python za izdelavo grafa napovedanih vrednosti glede na opazovane za model enostavne linearne regresije



Slika 15 Napovedane koncentracije raztopljenega kisika v primerjavi z opazovanimi za model enostavne linearne regresije

Točke se združujejo okoli diagonalne črte, kar kaže, da model zajame glavni padajoči trend raztopljenega kisika z naraščanjem temperature. Kljub temu ostaja opazen raztros, kar nakazuje, da temperatura sama ne pojasni v celoti spremenljivosti koncentracije raztopljenega kisika.

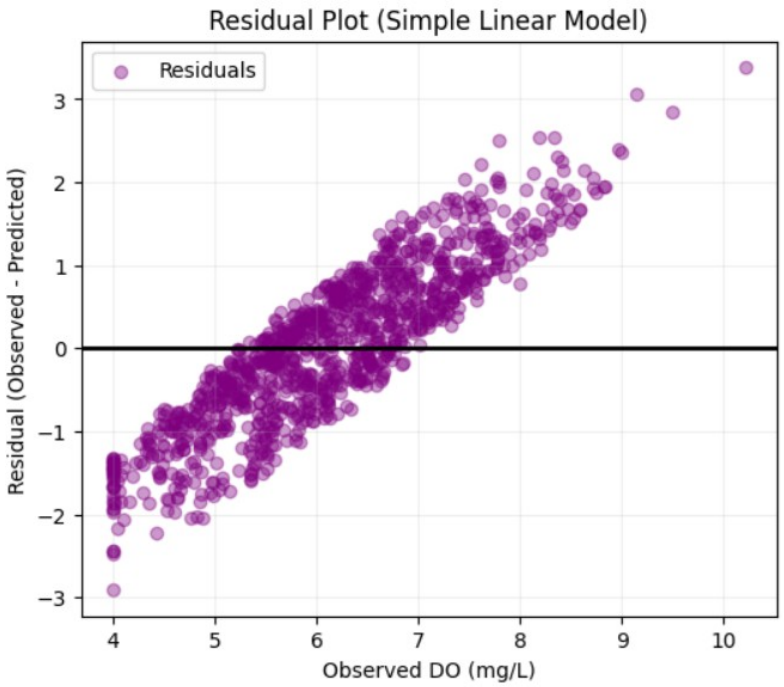
Analiza ostankov

Ostanki so bili izračunani kot razlika med opazovanimi in napovedanimi vrednostmi raztopljenega 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 Koda Python za izračun in risanje ostankov za model enostavne linearne regresije



Slika 17 Graf ostankov za model enostavne linearne regresije (opazovana vrednost raztopljenega kisika minus napovedana vrednost)

Ostanki so porazdeljeni okoli nič brez izrazite sistematične strukture, kar kaže, da linearni model zajame prevladujoči trend, vendar pušča nepojasnjeno spremenljivost, ki jo povzročajo dodatni ekološki in fizikalni procesi, kot sta mešanje vode in biološka aktivnost.

Ti rezultati poudarjajo omejitve uporabe temperature kot edine napovedne spremenljivke in utemeljujejo razvoj modela večkratne linearne regresije, ki vključuje dodatne okoljske spremenljivke.

Model večkratne linearne regresije (Temp + Depth + Conductivity → DO)

Učenje modela

Za izboljšanje napovedne uspešnosti in upoštevanje dodatnih fizikalnih dejavnikov spremenljivosti raztopljenega kisika **je bil razvit model večkratne linearne regresije z uporabo temperature vode, globine in električne prevodnosti kot napovednih spremenljivk**.
Te napovedne spremenljivke so bile izbrane na podlagi njihove ekološke pomembnosti. Temperatura vode neposredno uravnava topnost kisika, globina in električna prevodnost pa dajeta informacije o strukturi vodnega stolpca in procesih mešanja, ki vplivajo na porazdelitev kisika.
Podatkovna množica je bila razdeljena na učno in testno podmnožico v enakem razmerju 80/20 kot pri modelu enostavne linearne regresije, da bi omogočili nepristransko primerjavo uspešnosti modelov. Model je bil naučen na učni množici in nato uporabljen za napovedovanje koncentracij raztopljenega kisika v testni množici (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 Koda Python za učenje in vrednotenje modela večkratne linearne regresije z uporabo temperature, globine in električne prevodnosti kot napovednih spremenljivk

Enačba modela in uspešnost

Uspešnost modela je bila ovrednotena s korenom povprečne kvadratne napake (RMSE) in koeficientom determinacije (R^2).

V primerjavi z modelom enostavne linearne regresije je model večkratne linearne regresije dosegel nižji RMSE in višjo vrednost R^2, kar kaže na izboljšano natančnost napovedovanja in večjo sposobnost pojasnjevanja spremenljivosti koncentracije raztopljenega kisika (Slika 19).

Regresijski koeficienti kažejo, da temperatura ostaja prevladujoča napovedna spremenljivka z negativnim koeficientom, kar potrjuje obratno povezavo med temperaturo in koncentracijo raztopljenega kisika. Globina in električna prevodnost prispevata dodatno pojasnjevalno moč ter odražata vpliv stratifikacije in procesov mešanja na razpoložljivost 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 Izhod, ki prikazuje regresijske koeficiente in metrike uspešnosti (RMSE in R^2) za model večkratne linearne regresije

Ti rezultati kažejo, da model večkratne linearne regresije pojasni približno 42 % spremenljivosti koncentracije raztopljenega kisika in pomeni jasno izboljšanje v primerjavi z regresijskim modelom, ki uporablja samo temperaturo.

Napovedane vrednosti v primerjavi z opazovanimi

Izdelan je bil graf napovedanih vrednosti glede na opazovane za vizualno oceno ujemanja med izmerjenimi in napovedanimi koncentracijami raztopljenega kisika pri modelu večkratne regresije (Slika 20). Podatkovne točke se v primerjavi z modelom enostavne linearne regresije tesneje združujejo okoli diagonalne črte 1:1, kar kaže na izboljšano napovedno uspešnost ob vključitvi dodatnih okoljskih napovednih spremenljivk.

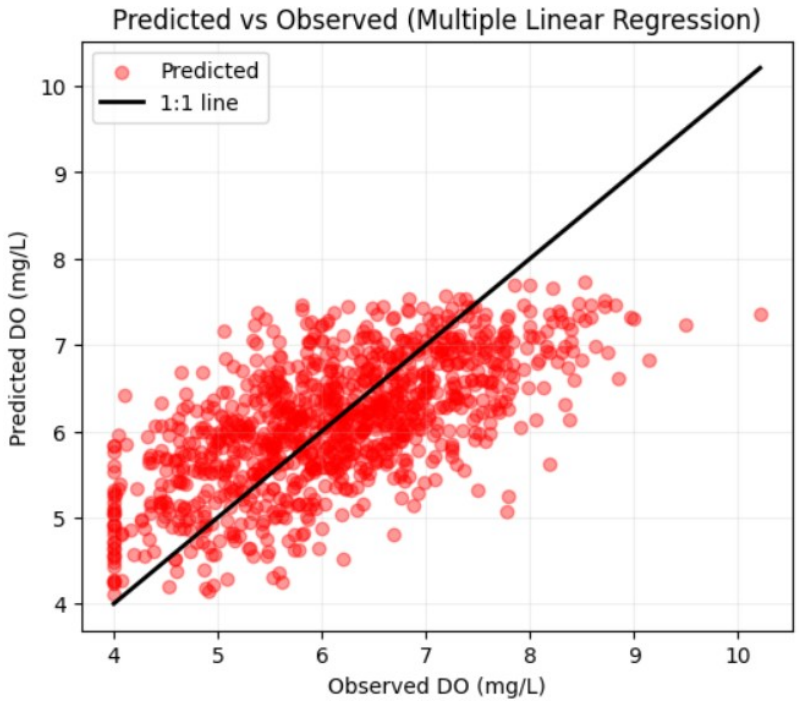
Preostala odstopanja od diagonalne črte kažejo vpliv nemodeliranih bioloških in fizikalnih procesov, kot so fotosinteza, dihanje in kratkotrajna hidrodinamična spremenljivost (Slika 21).

Na splošno graf napovedanih vrednosti glede na opazovane potrjuje, da vključitev globine in električne prevodnosti kot dodatnih napovednih spremenljivk izboljša zmožnost modela, da reproducira opazovane koncentracije raztopljenega 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 Koda Python za izdelavo grafa napovedanih vrednosti glede na opazovane za model večkratne linearne regresije.



Slika 21 Napovedane koncentracije raztopljenega kisika v primerjavi z opazovanimi za model večkratne linearne regresije

Analiza ostankov

Ostanki so bili analizirani za prepoznavanje sistematičnih napak in presojo veljavnosti predpostavk linearnega modela (Slika 22).

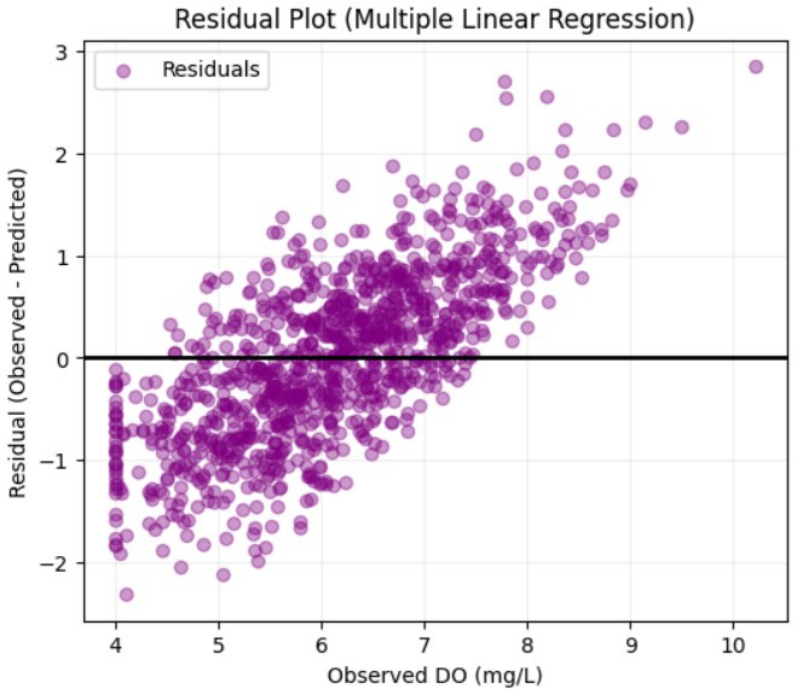
```
# Residual plot (Multiple Linear Regression)
resid_multi = y_test - y_pred_multi

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 Koda Python za izračun ostankov in izdelavo grafa ostankov za model večkratne linearne regresije

Graf ostankov kaže, da so ostanki porazdeljeni okoli nič brez izrazitega sistematičnega vzorca, kar nakazuje, da model večkratne linearne regresije zajame glavne linearne povezave med raztopljenim kisikom in izbranimi napovednimi spremenljivkami (temperaturo, globino in električno prevodnostjo) (Slika 23).



Slika 23 Graf ostankov za model večkratne linearne regresije, ki prikazuje razliko med opazovanimi in napovedanimi koncentracijami raztopljenega kisika

V primerjavi z modelom enostavne linearne regresije imajo ostanki bolj homogen raztros, kar nakazuje manjšo pristranskost in izboljšano stabilnost modela ob vključitvi dodatnih okoljskih napovednih spremenljivk.

Kljub temu pri višjih temperaturah in nižjih koncentracijah kisika ostaja nekaj razpršenosti, ki verjetno odraža vpliv nelinearnih bioloških in fizikalnih procesov, kot so fotosinteza, dihanje in kratkotrajno hidrodinamično mešanje, ki jih linearni model z omejenim številom napovednih spremenljivk ne more v celoti predstaviti.

Na splošno analiza ostankov potrjuje, da model večkratne linearne regresije zagotavlja razumen približek dinamike raztopljenega kisika, hkrati pa poudarja potrebo po kompleksnejših (nelinearnih) modelih za zajemanje preostale spremenljivosti.

Primerjava z modelom enostavne linearne regresije

Model večkratne linearne regresije je bil primerjan z modelom enostavne linearne regresije, ki je za napovedovanje koncentracije raztopljenega kisika uporabljal samo temperaturo vode.
Primerjava kaže, da je model večkratne linearne regresije dosegel nižji koren povprečne kvadratne napake (RMSE) in višji koeficient determinacije (R^2) kot model enostavne linearne regresije. To kaže na izboljšano natančnost napovedovanja in večjo sposobnost

pojasnjevanja spremenljivosti koncentracije raztopljenega kisika ob vključitvi dodatnih okoljskih napovednih spremenljivk.

Model enostavne linearne regresije sicer zajame prevladujočo obratno povezavo med temperaturo in raztopljenim kisikom, vendar ne upošteva drugih pomembnih fizikalnih dejavnikov, kot sta struktura vodnega stolpca in razmere mešanja. **Z vključitvijo globine in električne prevodnosti model večkratne linearne regresije bolje predstavlja kompleksnost dinamike kisika v estuariju.**

Ta primerjava ponazarja omejitve regresijskih modelov z eno napovedno spremenljivko za ekološke sisteme in poudarja pomen uporabe več ekološko smiselnih spremenljivk pri modeliranju okoljskih procesov.

Ekološka interpretacija

Negativni regresijski koeficient za temperaturo odraža dobro znano fizikalno načelo, da se topnost kisika z naraščanjem temperature vode zmanjšuje. Toplejša voda vsebuje manj raztopljenega kisika, kar neposredno vpliva na dihanje školjk in povečuje fiziološki stres v okoljih akvakulture.

Prispevka globine in električne prevodnosti kažeta na vlogo stratifikacije, dotoka sladke vode in procesov mešanja pri oblikovanju razpoložljivosti kisika v estuarijskem sistemu. Spremembe globine vplivajo na prodiranje svetlobe in biološko aktivnost, prevodnost pa je kazalnik gradientov slanosti in izmenjave vodnih mas med rečnimi in morskimi viri.

Obdobja z večjimi napakami napovedi lahko sovpadajo z okrepljenim fotosintetskim nastajanjem kisika podnevi ali z nenadnim mešanjem zaradi plimovanja, vetra ali rečnega pretoka.

Na splošno model večkratne linearne regresije zagotavlja tako napovedno sposobnost kot ekološki vpogled, saj dinamiko raztopljenega kisika povezuje s ključnimi fizikalnimi in biološkimi mehanizmi, ki delujejo na območjih gojenja školjk v estuarijih. To ponazarja, kako lahko regresijsko modeliranje podpira spremljanje okolja in sisteme zgodnjega opozarjanja na stresne razmere zaradi hipoksije.

Praktična vaja: Regresija v okoljskih znanostih

Tema: Napovedovanje koncentracije klorofila a na podlagi fizikalnih oceanografskih spremenljivk

Delo študenta: samostojno

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

Podatkovna množica:
Copernicus Marine Service (CMEMS) – površinska slanost morja, klorofil in hitrost toka.

Ozadje in motivacija

Biomasa fitoplanktona, ki jo običajno izražamo s koncentracijo klorofila a (Chl-a), je ključni kazalnik produktivnosti morskega ekosistema in kakovosti vode. Na spreminjanje koncentracije klorofila a vplivajo fizikalni dejavniki, kot so slanost in morski tokovi, ki določajo razpoložljivost hranil, stratifikacijo in procese mešanja.
V tej praktični vaji študenti uporabijo regresijske tehnike za modeliranje koncentracije klorofila a z okoljskimi napovednimi spremenljivkami, pridobljenimi iz podatkov storitve Copernicus Marine Service (CMEMS). Naloga prikazuje, kako je mogoče enostavno in večkratno linearno regresijo uporabiti za kvantificiranje povezav med biološkimi in fizikalnimi spremenljivkami v morskih sistemih ter za podporo spremljanju okolja in zgodnjemu opozarjanju.

Opis podatkovne množice

Podatkovna množica izhaja iz morskih produktov CMEMS in vsebuje časovne vrste ali prostorske vzorce razmer na površju oceana. Vsaka vrstica ustreza eni opazovalni točki (čas in/ali lokacija) ter vključuje naslednje spremenljivke:

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

Napovedne spremenljivke (X):
- Površinska slanost morja (PSU)
- Hitrost površinskega toka (m/s)

Opis naloge

Študent mora izdelati in primerjati dva regresijska modela za napovedovanje koncentracije klorofila a:
1. **Enostavna linearna regresija:**
   Chl-a = f(Salinity)
2. **Večkratna linearna regresija:**
   Chl-a = f(Salinity, Current Velocity)

Navodila po korakih

1. korak – Priprava
- Odprite Google Colab ali Jupyter Notebook.

- Uvozite potrebne knjižnice Python (pandas, numpy, scikit-learn, matplotlib).

2. korak – Nalaganje in pregled podatkovne množice

- Naložite podatkovno množico CMEMS iz priložene datoteke CSV.
- Prikažite prve vrstice podatkovne množice.
- Navedite:
  - Število opazovanj (vrstic)
  - Število spremenljivk (stolpcev)
  - Ciljna spremenljivka
  - Napovedne spremenljivke
- Preverite manjkajoče vrednosti.

3. korak – Priprava podatkov

- Po potrebi pretvorite skalirane vrednosti v fizikalne enote.
- Odstranite nefizikalne ali identifikacijske stolpce.
- Preglejte opisno statistiko in morebitne osamelce.
- Podatkovno množico razdelite na učno in testno množico (80/20).

4. korak – Model 1: enostavna linearna regresija (izhodiščni model)

Naučite model enostavne linearne regresije s slanostjo kot edino napovedno spremenljivko.
Formulacija modela:
Chl-a = f(Salinity)

- Prilagodite model učnim podatkom.
- Napovejte koncentracijo klorofila a na testnih podatkih.
- Izračunajte:
  - RMSE
  - R^2

Vizualizacije:

- Razsevni diagram: slanost glede na klorofil a
- Regresijska črta
- Graf napovedanih vrednosti glede na opazovane
- Graf ostankov

Od študenta se pričakuje:

Na kratko pojasnite fizikalni pomen regresijskega koeficienta (kako je slanost povezana s koncentracijo klorofila a).

5. korak – Model 2: večkratna linearna regresija

Naučite model večkratne linearne regresije z uporabo obeh napovednih spremenljivk:
Formulacija modela:
Chl-a = f(Salinity, Current Velocity)

- Prilagodite model učnim podatkom.
- Napovejte koncentracijo klorofila a na testnih podatkih.
- Izračunajte:
  - RMSE
  - R^2

Vizualizacije:

- Graf napovedanih vrednosti glede na opazovane
- Graf ostankov
-

Od študenta se pričakuje:

Interpretirajte koeficienta slanosti in hitrosti toka z ekološkega vidika (mešanje, prenos, stratifikacija).

6. korak – Primerjava modelov

Primerjajte modela:

- Kateri model ima nižji RMSE?
- Kateri model ima višji R^2?
- Ali dodajanje hitrosti toka izboljša uspešnost napovedovanja?

Od študenta se pričakuje:

V 1–2 povedih na kratko pojasnite, zakaj je model večkratne regresije uspešnejši ali manj uspešen od enostavnega modela.

7. korak – Ekološka interpretacija

Študenti naj razpravljajo o naslednjem:

- Kako slanost prek dotoka sladke vode in vnosa hranil vpliva na fitoplankton.
- Kako hitrost toka s procesi mešanja in prenosa vpliva na porazdelitev klorofila.
- Zakaj del spremenljivosti ostaja nepojasnjen (razpoložljivost svetlobe, hranila, biološki procesi).

Kaj je treba oddati – Poročilo o vaji

Oblika:
Jupyter Notebook ali Google Colab z besedilnimi celicami Markdown in celicami s kodo.
Poročilo mora vsebovati:

a. Uvod

- Kaj je bilo modelirano?

- Zakaj je klorofil a pomemben za morske ekosisteme?

b. Opis podatkov
- Število vzorcev
- Napovedne spremenljivke
- Ciljna spremenljivka
- Obravnava manjkajočih vrednosti

c. Metode
- Enostavna linearna regresija
- Večkratna linearna regresija
- Razmerje delitve na učno/testno množico
- Metrike vrednotenja (RMSE, R^2)

d. Rezultati
- RMSE in R^2 za oba modela
- Grafi (razsevni graf, graf napovedanih vrednosti glede na opazovane, graf ostankov)
- Kratka primerjava modelov

e. Razprava
- Interpretacija koeficientov
- Okoljski pomen napak
- Omejitve linearne regresije

f. Zaključek
- Povzetek pridobljenega znanja o regresijskem modeliranju z uporabo morskih okoljskih podatkov

Ocenjevanje

PRAKTIČNA NALOGA IZ REGRESIJE

Praktična naloga iz regresije študentom omogoča, da z neposrednim praktičnim delom uporabijo regresijske modele (enostavno in večkratno linearno regresijo) na resničnem okoljskem naboru podatkov, pridobljenem iz storitve Copernicus Marine Service (CMEMS). Namenjena je ocenjevanju sposobnosti študentov za implementacijo, validacijo in interpretacijo regresijskih modelov v podporo okoljskemu odločanju.

KVIZ

Kviz bo zasnovan tako, da bo ocenjeval konceptualno razumevanje ter sposobnost uporabe znanja in sklepanja, ki so jih študenti razvili na predavanju in pri praktični vaji o regresijskih tehnikah v okoljskih znanostih. Preverjal bo, ali študenti razumejo razliko

med enostavno in večkratno linearno regresijo ter vlogo regresijskih modelov pri okoljskem spremljanju in napovedovanju.

KRATKO PISNO POROČILO

Kratko pisno poročilo je sklepni korak praktične naloge iz regresije, v katerem lahko študenti strnejo, interpretirajo in predstavijo svoje ugotovitve iz vaje regresijskega modeliranja.

Merila ocenjevanja
1. KVIZ

Vprašanje 1
Kakšna vrsta problema strojnega učenja je napovedovanje koncentracije raztopljenega kisika iz okoljskih spremenljivk, kot sta temperatura in globina?
a) Klasifikacija
b) Regresija
c) Razvrščanje v skupine
d) Odkrivanje anomalij

Vprašanje 2
Katera spremenljivka predstavlja ciljno (odvisno) spremenljivko v primeru uporabe, obravnavanem v tej učni enoti?
a) Temperatura vode
b) Električna prevodnost
c) Koncentracija raztopljenega kisika
d) Globina

Vprašanje 3
Katera od naslednjih možnosti najbolje opredeljuje regresijski problem pri strojnem učenju?
a) Napovedovanje kategorične oznake
b) Napovedovanje zvezne številske vrednosti
c) Združevanje podobnih vzorcev
d) Zmanjševanje števila spremenljivk

Vprašanje 4
Kakšna je glavna razlika med enostavno linearno regresijo in večkratno linearno regresijo?
a) Enostavna regresija je nenadzorovana
b) Večkratna regresija uporablja več kot eno napovedno spremenljivko
c) Enostavna regresija je vedno natančnejša
d) Večkratne regresije ni mogoče ovrednotiti

Vprašanje 5
Zakaj regresija velja za metodo nadzorovanega učenja?
a) Ker ne zahteva označenih podatkov
b) Ker se model uči brez ciljne spremenljivke
c) Ker se model uči z uporabo znanih parov vhod–izhod
d) Ker deluje samo na majhnih naborih podatkov

Vprašanje 6
Kaj meri koren povprečne kvadratne napake (RMSE)?
a) Delež pojasnjene variance
b) Povprečno napako napovedi v istih enotah kot ciljna spremenljivka
c) Število pravilnih napovedi
d) Korelacijo med spremenljivkami

Vprašanje 7
Kaj je glavni namen delitve nabora podatkov na učno in testno množico?
a) Povečanje kompleksnosti modela
b) Odstranjevanje osamelcev
c) Vrednotenje uspešnosti modela na še nevidenih podatkih
d) Skrajšanje časa računanja

Vprašanje 8
Katera situacija nakazuje, da bi bil nelinearni regresijski model lahko primernejši od linearnega?
a) Razmerje med spremenljivkami tvori ravno črto
b) Podatki kažejo pragove ali ukrivljene vzorce odziva
c) V podatkih ni šuma
d) Na voljo je le ena napovedna spremenljivka

Vprašanje 9
Katere informacije ponuja graf napovedanih vrednosti glede na opazovane?
a) Prikazuje porazdelitev napovednih spremenljivk
b) Prikazuje, kako dobro se napovedi ujemajo z izmerjenimi vrednostmi
c) Odkriva manjkajoče podatke
d) Samodejno izbere najboljše napovedne spremenljivke

Vprašanje 10
Zakaj so regresijski modeli uporabni v okoljskih znanostih?
a) Nadomeščajo fizikalne meritve
b) Pomagajo napovedovati in interpretirati zvezne okoljske spremenljivke
c) Odpravljajo negotovost v opazovanjih
d) Delujejo samo za laboratorijske podatke

2. PRAKTIČNA NALOGA IZ REGRESIJE

Priprava in predobdelava podatkov

• Pravilno nalaganje in pregled nabora podatkov (dimenzije, napovedne spremenljivke, ciljna spremenljivka).
• Ustrezna obravnava manjkajočih podatkov in pretvarjanje enot, kjer je potrebno.
• Pregleden in ponovljiv potek predobdelave (koda je čista, koraki pa so jasno pojasnjeni).

Implementacija modela

• Pravilna nastavitev in učenje vsaj dveh regresijskih modelov:
  Enostavna linearna regresija
  Večkratna linearna regresija
• Ustrezna uporaba delitve na učno/testno množico.
• Koda je urejena, parametri pa so izrecno navedeni.

Vrednotenje in diagnostika modela

- Pravilen izračun regresijskih metrik in pravilno poročanje o njih:
  RMSE
  R^2
  • Logična primerjava modelov (npr. »večkratna regresija je izboljšala R^2 v primerjavi z enostavno regresijo«).
  • Uporaba diagnostičnih grafov (grafi napovedanih vrednosti glede na opazovane in grafi ostankov).
  • Razumevanje prenaučenja in podprileganja.

Izbira modela in utemeljitev

• Jasna utemeljitev, kateremu modelu je treba dati prednost.
• Smiselna razlaga sposobnosti posploševanja modela na podlagi rezultatov testnih podatkov.

Vizualizacija in predstavitev rezultatov

• Slike (razsevni grafi, prilagojene regresijske krivulje, grafi ostankov) so jasno označene.
• Vizualizacije podpirajo interpretacijo obnašanja modela.

Povezava z okoljskim kontekstom

• Študent pojasni, kako lahko regresijski modeli podpirajo okoljsko spremljanje in sisteme zgodnjega opozarjanja.

3. KRATKO PISNO POROČILO

Jasnost in struktura poročila

- Zgradba: Uvod → Podatki → Metode → Rezultati → Razprava → Zaključek.
- Beležnica prikazuje korake dela.
- Jezik je natančen in jedrnat.

Interpretacija rezultatov in razprava

- Rezultati regresije (RMSE, R^2, ostanki) so pravilno navedeni in razloženi.
- Študent pojasni, zakaj je večkratna regresija uspešnejša (ali ne) od enostavne.

Povezava z ekološkim in okoljskim kontekstom

- Ugotovitve imajo okoljski pomen (npr. napoved klorofila pri spremljanju morskega ekosistema).
- Študent razume prispevek regresije k okoljskemu odločanju.

Kakovost vizualne predstavitve

- Slike so jasno označene in navedene v besedilu.
- Vizualizacije pomagajo razložiti obnašanje modela in niso le okras poročila.

Komunikacija in slog

- Pisanje je jasno, objektivno in strokovno.
- Izrazi so pravilni (napovedna/ciljna spremenljivka, RMSE, ostanki, koeficient).
- Študent navede omejitve modela, linearnosti in števila spremenljivk.

Predlagane uteži

- Kviz: **20 %**
- Praktična naloga iz regresije (beležnica/koda): **50 %**
- Kratko pisno poročilo: **30 %**

Viri / Nadaljnje branje

- 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.
- Portal za dostop do podatkov Copernicus Marine Service (CMEMS)
- Odprtodostopni raziskovalni članki (celoten seznam bo v e-učilnici)