Argomento: Tecniche di regressione per la modellazione ecologica Corso: Machine Learning per applicazioni ambientali

Descrizione della lezione:

Questa lezione introduce la regressione come approccio fondamentale del machine learning per modellare e prevedere variabili ambientali continue. Vengono affrontati sia gli aspetti teorici sia quelli applicativi, distinguendo le tecniche di regressione lineare e non lineare rilevanti per gli studi ambientali ed ecologici. Gli studenti esploreranno l'intero processo di regressione: preparazione dei dati, costruzione delle caratteristiche, addestramento e valutazione dei modelli mediante RMSE, nonché strategie per limitare overfitting e underfitting.
Il caso di studio riguarda la previsione della concentrazione di ossigeno disciolto a partire dalla temperatura dell'acqua e da altre variabili ambientali in un'area estuarina destinata alla molluschicoltura. L'esempio mostra come i modelli di regressione possano contribuire al monitoraggio ecologico e ai sistemi di allerta precoce per condizioni di stress degli organismi acquatici.

Obiettivi di apprendimento:

Al termine di questa unità, gli studenti saranno in grado di:
1. Definire i problemi di regressione e distinguerli dagli altri compiti di machine learning.
2. Spiegare i principi della regressione lineare semplice e multipla e la logica della valutazione mediante RMSE.
3. Sviluppare e valutare modelli di regressione utilizzando dataset ambientali.
4. Individuare e affrontare problemi comuni, quali rumore, multicollinearità, underfitting e overfitting, nella regressione di dati ecologici.
5. Interpretare i risultati dei modelli in contesti ecologici e ambientali.

Contenuti della lezione

Introduzione

Le tecniche di regressione sono strumenti essenziali per modellare variabili ambientali continue, come temperatura, salinità e concentrazioni di inquinanti. Questa sezione illustra la differenza tra regressione e classificazione, introduce l'errore quadratico medio radice (RMSE) e discute l'importanza dei modelli di regressione per comprendere e prevedere i fenomeni ecologici.

Contesto e motivazioni

Le tecniche di regressione sono fondamentali per modellare molti fenomeni ambientali caratterizzati da una risposta continua. Nelle scienze ecologiche e ambientali si cerca spesso di prevedere una variabile continua (per esempio la concentrazione di un inquinante, la biomassa di una specie, la portata di un fiume o un'anomalia di

temperatura) a partire da variabili predittive (meteorologia, caratteristiche del paesaggio e del bacino idrografico, uso del suolo e così via). La regressione collega quindi i dati osservativi e ambientali alla modellazione esplicativa e predittiva, collocandosi all'intersezione tra ecologia statistica, idrologia, climatologia e machine learning.
Esempi di problemi di regressione nelle scienze ecologiche e ambientali:

- Stimare la concentrazione di biossido di azoto (NO_2) o di particolato ($PM_{2,5}$) in aree urbane o industriali sulla base di dati meteorologici, traffico, uso del suolo e condizioni atmosferiche.
- Modellare l'abbondanza o la biomassa di una specie nel tempo o nello spazio in funzione di fattori ambientali quali temperatura, precipitazioni, nutrienti del suolo, frammentazione degli habitat e disturbo antropico.
- Prevedere risposte idrologiche, come la portata giornaliera di un fiume, a partire da precipitazioni, temperatura dell'aria, umidità, umidità antecedente del suolo e caratteristiche del bacino.
- Prevedere anomalie climatiche, come le anomalie di temperatura, o risposte degli ecosistemi, come flussi di carbonio ed emissioni di gas serra, mediante modelli di regressione continua.

Compiti di regressione e classificazione

È essenziale distinguere la regressione dalla classificazione: entrambe appartengono all'apprendimento supervisionato, ma perseguono obiettivi predittivi differenti.
Nella classificazione, la variabile obiettivo y è categorica e rappresenta etichette o classi discrete (per esempio campioni d'acqua «inquinati» o «puliti», oppure «presenza» o «assenza» di una specie). Nella regressione, invece, y è continua: può assumere valori su una scala numerica senza confini di classe intrinseci (per esempio la concentrazione esatta di nitrati nelle acque sotterranee o la densità di biomassa in un'area forestale).
Nella classificazione interessano le frontiere decisionali, cioè la categoria alla quale assegnare un nuovo campione. Nella regressione interessa la precisione numerica, ossia quanto la previsione si avvicina al valore continuo effettivamente misurato.
Mettere in evidenza questa distinzione fin dall'inizio aiuta a comprendere perché la regressione sia essenziale in ambito ambientale: molte variabili da prevedere sono intrinsecamente continue e richiedono sia la stima delle tendenze sia la comprensione dell'incertezza e della variabilità delle previsioni.

Perché la regressione è un problema di apprendimento supervisionato

La regressione appartiene ai metodi di apprendimento supervisionato perché l'addestramento richiede dati etichettati: coppie di input e valori obiettivo noti (Xi, yi).
Il modello «apprende» modificando i propri parametri per ridurre la discrepanza tra i valori previsti e quelli effettivamente osservati, yi. $\hat{y}_i = f(X_i)$
La «supervisione» deriva da questi risultati noti: il processo di apprendimento è guidato dal segnale di errore, cioè dalla differenza tra previsioni e realtà. L'obiettivo è ridurre progressivamente l'errore di previsione attraverso una funzione obiettivo, spesso l'errore quadratico medio.

Utilità dell'RMSE nella regressione ambientale ed ecologica

Un aspetto metodologico cruciale della regressione è scegliere una metrica di valutazione, ossia una misura quantitativa dell'accordo tra le previsioni del modello e i dati osservati.
L'errore quadratico medio radice (Root Mean Squared Error, RMSE) è una delle metriche più utilizzate nella modellazione ambientale ed ecologica ed è definito come segue:

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

Per comprendere questa metrica si può procedere dall'interno della formula verso l'esterno. Si calcola innanzitutto la differenza tra valore reale e previsto e la si eleva al quadrato, evitando che errori positivi e negativi si annullino. Si sommano quindi tutti gli errori quadratici e si divide per il numero di osservazioni per ottenere la media. Infine, la radice quadrata riporta il risultato all'unità di misura originaria.

- L'RMSE fornisce una misura interpretabile nella stessa unità della variabile obiettivo, fondamentale per comunicare i risultati agli esperti ambientali, ai portatori di interesse e ai decisori pubblici.
- Riflette sia la componente di errore sistematico (bias) sia la dispersione dei residui, descrivendo l'affidabilità predittiva complessiva del modello.
- L'RMSE può essere facilmente utilizzato negli algoritmi di ottimizzazione del machine learning, facilitando la regolazione automatica dei parametri.

Principali vantaggi dell'RMSE:

- L'RMSE è intuitivo: valori più bassi indicano una maggiore accuratezza delle previsioni.
- La penalizzazione quadratica mostra perché gli errori di previsione molto grandi pesano in modo sproporzionato, aspetto importante nella valutazione del rischio ecologico.
- Il confronto tra RMSE e metriche alternative, come errore assoluto medio (MAE) o errore medio con segno, stimola una riflessione critica sulla scelta e interpretazione delle metriche negli studi ambientali.

Panoramica delle tecnologie

La modellazione mediante regressione nelle scienze ecologiche e ambientali non è soltanto una procedura matematica: comprende anche una componente tecnologica che trasforma dati ambientali eterogenei in modelli interpretabili e verificabili.

Per rendere il processo accessibile agli studenti si utilizzeranno diversi modelli di regressione (regressione lineare, alberi di regressione, SVM per la regressione e reti neurali artificiali) su dataset rappresentativi delle scienze ecologiche e ambientali.
Integrando fonti di dati aperti, moderne librerie di elaborazione e tecniche di visualizzazione, saranno affrontati i principi tecnologici e metodologici della regressione nel machine learning ambientale.

Integrazione tecnologica e dataset

In primo luogo viene utilizzato un dataset sintetico per mostrare i principi della regressione lineare semplice e multipla in un contesto controllato e facilmente interpretabile. Gli studenti possono così concentrarsi sul comportamento dei modelli senza le complicazioni dovute al rumore e ai valori mancanti nei dati reali.

Nella seconda parte viene utilizzato un dataset marino reale del Copernicus Marine Environment Monitoring Service (CMEMS), con dati di salinità superficiale, concentrazione di clorofilla e velocità delle correnti, per applicare lo stesso processo di regressione a osservazioni ambientali reali e valutarne prestazioni e rilevanza ecologica.

Tutte le analisi vengono svolte in Google Colab con librerie Python standard:

- **pandas** e **NumPy** per il caricamento, la pulizia e l'elaborazione numerica dei dati;
- **scikit-learn** per implementare la regressione lineare semplice e multipla e calcolare le metriche di valutazione (RMSE e R^2);
- **Matplotlib** per visualizzare distribuzioni dei dati, adattamenti dei modelli di regressione e grafici diagnostici.

La regressione lineare semplice è impiegata come modello di riferimento per descrivere la relazione tra una singola variabile predittiva ambientale (per esempio la temperatura) e una variabile risposta continua (l'ossigeno disciolto). Offre il vantaggio dell'interpretabilità e illustra i concetti fondamentali della regressione.

La regressione lineare multipla estende questo approccio includendo ulteriori predittori (per esempio profondità e conducibilità elettrica) per rappresentare meglio i processi fisici ed ecologici responsabili della variabilità dell'ossigeno disciolto. Mostra come variabili ambientali pertinenti possano migliorare la capacità esplicativa mantenendo la trasparenza del modello.

Regressione lineare semplice e multipla: concetti, metodo e dimostrazione

La regressione lineare semplice mette in relazione un predittore e una variabile risposta:

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

- Dove il termine indicato rappresenta la variabilità non spiegata o il rumore di misura.ε

Nella regressione multipla si considerano più variabili indipendenti Xi e una variabile dipendente.
L'obiettivo consiste nel determinare β0 e β1 (o βk nella regressione multipla) in modo da minimizzare l'errore quadratico medio tra valori osservati e previsti.
La lezione illustra la regressione lineare semplice e multipla attraverso misure di sensori ambientali provenienti da un'area estuarina di molluschicoltura. Nel caso applicativo, la concentrazione di ossigeno disciolto (mg/L) è modellata in funzione della temperatura dell'acqua e di ulteriori predittori (per esempio profondità e conducibilità). Si mostra così come modelli lineari offrano relazioni di riferimento interpretabili e come l'impiego di più predittori possa migliorare la capacità esplicativa nei sistemi ecologici.

Fasi:

Preparazione dei dati:
caricare il dataset dei sensori ambientali con misure di temperatura dell'acqua, ossigeno disciolto e, se disponibili, profondità e conducibilità elettrica. Convertire i valori scalati in unità fisiche, controllare valori mancanti e anomali e suddividere i dati in insiemi di addestramento e test.

Analisi:
adattare due modelli:
(i) regressione lineare semplice con la temperatura come unico predittore dell'ossigeno disciolto (Temp → DO);
(ii) (ii) regressione lineare multipla con temperatura, profondità e conducibilità (Temp + Depth + Conductivity → DO).

Valutazione:
valutare entrambi i modelli con l'errore quadratico medio radice (RMSE) e il coefficiente di determinazione (R^2). Visualizzare valori previsti e osservati e analizzare i residui per individuare errori sistematici e verificare le ipotesi del modello.

Perfezionamento:
esaminare la multicollinearità tra predittori e discuterne l'impatto sulla stabilità del modello. Come estensione, introdurre tecniche di regolarizzazione, quali Ridge e Lasso, per ridurre l'overfitting e migliorare la robustezza.

Interpretazione:

interpretare coefficienti e distribuzione dei residui in termini ecologici e ambientali, considerando meccanismi fisici e biologici quali solubilità dell'ossigeno, fotosintesi e mescolamento delle acque.

Regressione non lineare: concetti, metodo e dimostrazione

La regressione non lineare estende la regressione lineare alle relazioni che non possono essere descritte adeguatamente da una retta. In numerosi processi ecologici e ambientali la variabile risposta non varia linearmente con i predittori, ma può mostrare soglie, saturazione, crescita esponenziale o andamenti ciclici.

Una definizione formale della regressione non lineare può essere espressa come segue:

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

- Dove f(·) è una funzione non lineare dei predittori Xi e dei parametri β, mentre εi rappresenta l'errore casuale.

Tra le forme funzionali più comuni figurano trasformazioni polinomiali, esponenziali, logaritmiche o di potenza, nonché modelli basati su kernel, alberi o reti neurali artificiali quando è necessaria maggiore flessibilità non parametrica.

L'obiettivo rimane lo stesso della regressione lineare: minimizzare la differenza tra valori osservati (yi) e previsti, generalmente tramite RMSE. $\hat{y}_i$

Nella regressione non lineare, tuttavia, la curva adattata può assumere forme differenti secondo la teoria ecologica, il comportamento dei dati e le trasformazioni delle caratteristiche.

Questa dimostrazione modella la concentrazione di ossigeno disciolto (DO) in funzione della temperatura dell'acqua usando un dataset sintetico relativo a un'area estuarina di molluschicoltura. La regressione polinomiale illustra come i modelli non lineari descrivano andamenti curvilinei quando le ipotesi di linearità non sono sufficienti.

Fasi:

Preparazione dei dati:

recuperare misure continue di temperatura dell'acqua e concentrazione di ossigeno disciolto dal sistema di monitoraggio dell'area estuarina. Integrare il dataset reale con dati sintetici che accentuino i pattern non lineari a scopo didattico. Gestire i valori mancanti, normalizzare i predittori e suddividere i dati in addestramento, validazione e test.

Analisi:

applicare la regressione polinomiale per descrivere relazioni non lineari tra temperatura e ossigeno disciolto. Confrontarla con Support Vector Regression (SVR), Random Forest Regression e reti neurali artificiali (ANN) per esplorare strutture non lineari più complesse.

Valutazione:

calcolare l'RMSE di tutti i modelli e visualizzare valori previsti e osservati con grafici bidimensionali (temperatura rispetto a ossigeno disciolto) e, facoltativamente, tridimensionali se sono presenti ulteriori predittori. Esaminare le curve di risposta non lineari per identificare soglie ecologiche, per esempio temperature oltre le quali l'ossigeno disciolto diminuisce rapidamente.

Ottimizzazione del modello:

utilizzare la validazione incrociata a 10 fold per scegliere il grado del polinomio e ottimizzare gli iperparametri dei kernel o delle reti neurali. Confrontare le prestazioni sui dati di test valutando errore di generalizzazione e interpretabilità.

Interpretazione:

collegare i risultati ai processi degli ecosistemi acquatici e discuterne le implicazioni per la gestione della molluschicoltura, l'allerta precoce di condizioni ipossiche e lo stress climatico negli estuari.

Metriche di valutazione dei problemi di regressione

Le metriche di valutazione quantificano le prestazioni predittive, mentre le analisi diagnostiche verificano quanto la struttura del modello sia coerente con il processo ecologico sottostante. L'obiettivo è capire non solo quanto il modello sia accurato, ma anche perché ottenga certi risultati, dove funzioni o fallisca e cosa implichino gli errori.
Le misure maggiormente utilizzate nella regressione ecologica sono:

Errore quadratico medio radice (RMSE), già descritto:

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

Sensibile agli errori grandi: particolarmente utile quando deviazioni estreme hanno conseguenze ecologiche importanti, per esempio sottostimare la portata di una piena o sovrastimare i livelli di inquinanti.

Errore assoluto medio (MAE):

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

Meno sensibile ai valori anomali e utile per misurare lo scostamento medio.

Coefficiente di determinazione (R^2):
Questa misura non è affidabile per dati temporali o spaziali. Negli altri casi misura la proporzione della varianza di y spiegata dal modello, aiutando a valutarne la capacità esplicativa complessiva.

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

- SSres → somma dei quadrati dei residui (varianza non spiegata) → $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Somma totale dei quadrati (varianza totale dei dati osservati) → $SS_{tot}\sum_{i=1}^{n}(y_i - \overline{y})^2$
- Valore osservato→ y_i
- Previsione del modello → $\hat{y}_i$
- Media di tutti i valori osservati → $\acute{y}$
- Numero di campioni → n

Efficienza di Nash–Sutcliffe (NSE):
Comune in idrologia, esprime la capacità predittiva rispetto alla media delle osservazioni.

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

Dove Q_{obs} indica i valori osservati (in questo caso della portata Q), Q_{mod} i valori simulati della stessa variabile e il termine $\overline{Q}$ rappresenta la media delle osservazioni della variabile selezionata.

Caso di studio applicativo: regressione dell'ossigeno disciolto in un'area estuarina di molluschicoltura

Contesto e motivazioni

In un sito di molluschicoltura situato in un ambiente estuarino o fluviale, temperatura dell'acqua e ossigeno disciolto (DO, mg/L) vengono misurati continuamente mediante sensori in situ. L'ossigeno disciolto influenza direttamente la respirazione e lo stress dei molluschi e costituisce quindi un indicatore fondamentale dello stato di salute dell'ecosistema e della sostenibilità dell'acquacoltura.

L'obiettivo del caso di studio è costruire un modello di regressione che stimi l'ossigeno disciolto a partire dalla temperatura dell'acqua. Il modello fornisce un'indicazione rapida di possibili condizioni sfavorevoli: all'aumentare della temperatura ci si attende una diminuzione della concentrazione di ossigeno. Individuare precocemente queste condizioni permette agli allevatori di monitorare e ridurre più efficacemente i rischi di stress ecologico.

Gli studenti confrontano le previsioni con i valori misurati e analizzano le situazioni in cui il modello fallisce. Questo introduce intuitivamente l'idea che la regressione possa migliorare includendo ulteriori variabili ambientali, come saturazione dell'ossigeno (%), profondità, conducibilità e torbidità.

Regressione lineare semplice

Fase 1 – Comprensione dei dati

Descrizione del dataset e delle variabili

Il dataset di questo caso di studio deriva dal monitoraggio ambientale continuo di un sito estuarino di molluschicoltura. Le misure raccolte con sensori in situ comprendono temperatura dell'acqua, concentrazione di ossigeno disciolto e altre variabili ambientali impiegate per estendere il modello.
Il dataset contiene diverse migliaia di osservazioni e non presenta valori mancanti nelle variabili principali utilizzate nella modellazione. Tutti i valori, inizialmente espressi in unità scalate, sono stati convertiti durante il preprocessing in unità fisiche (°C e mg/L).

Nome del dataset:
Sistema di monitoraggio ambientale in situ (dataset sintetico preparato per finalità didattiche)

Il problema di regressione è definito come segue:

Variabile obiettivo (y):
Concentrazione di ossigeno disciolto (DO, mg/L)
Predittore principale (X):
Temperatura dell'acqua (Temp, °C)

Predittori aggiuntivi (per la regressione multipla):

- Profondità (m)
- Conducibilità elettrica (µS/cm)
- (facoltativo) saturazione dell'ossigeno (%)
- (facoltativo) torbidità

Queste variabili sono state selezionate per la loro rilevanza fisica ed ecologica. La temperatura influisce direttamente sulla solubilità dell'ossigeno, mentre profondità e conducibilità riflettono la struttura della colonna d'acqua e i processi di mescolamento che influenzano la disponibilità di ossigeno.
Non tutti gli attributi disponibili sono stati utilizzati nei modelli. Identificativi delle registrazioni e codici privi di significato fisico sono stati esclusi per evitare rumore e overfitting e preservare l'interpretabilità dei risultati.

Questo è un problema di **regressione supervisionata con variabile obiettivo continua.**

Fase 2 – Esplorazione dei dati

Il dataset dei sensori ambientali è stato caricato con la libreria Pandas. Tutte le misure scalate sono state convertite in unità fisiche (°C per la temperatura e mg/L per l'ossigeno disciolto), così da rendere fisicamente interpretabili i modelli di regressione.

Il dataset è stato importato da un file CSV delimitato da punti e virgola e preprocessato convertendo i valori scalati in unità fisiche (Figura 1). L'anteprima delle prime osservazioni conferma il corretto caricamento e prepara i dati all'analisi di regressione (Figura 2).Figura 1. Codice Python per caricare il dataset dei sensori da un CSV separato da punti e virgola e convertire temperatura e ossigeno disciolto in unità fisiche (°C e mg/L).1Figura 2. Anteprima delle prime osservazioni dopo la conversione delle unità, che conferma il corretto caricamento e l'interpretabilità fisica dei valori di temperatura e ossigeno disciolto.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())
```

Figura 1. Codice Python per caricare il dataset dei sensori da un CSV separato da punti e virgola e convertire temperatura e ossigeno disciolto in unità fisiche (°C e mg/L).1

	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

Figura 2. Anteprima delle prime osservazioni dopo la conversione delle unità, che conferma il corretto caricamento e l'interpretabilità fisica dei valori di temperatura e ossigeno disciolto.2

Non sono stati rilevati valori mancanti nelle variabili principali usate per la regressione (Figure 3 e 4).Figura 3. Codice Python per controllare i valori mancanti nelle variabili temperatura e ossigeno disciolto.3Figura 4. Output che mostra l'assenza di valori mancanti nelle variabili principali.4

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

Figura 3. Codice Python per controllare i valori mancanti nelle variabili temperatura e ossigeno disciolto.3

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

Figura 4. Output che mostra l'assenza di valori mancanti nelle variabili principali.4

Le statistiche descrittive sono state utilizzate per esaminare l'intervallo e la variabilità di temperatura e ossigeno disciolto e identificare possibili valori anomali (Figure 5 e 6).Figura 5. Codice Python per calcolare statistiche descrittive (conteggio, media, deviazione standard e intervallo) di temperatura e ossigeno disciolto.5Figura 6. Statistiche riassuntive di temperatura e ossigeno disciolto per valutarne intervallo e variabilità.6

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

Figura 5. Codice Python per calcolare statistiche descrittive (conteggio, media, deviazione standard e intervallo) di temperatura e ossigeno disciolto.5

	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

Figura 6. Statistiche riassuntive di temperatura e ossigeno disciolto per valutarne intervallo e variabilità.6

Per esplorare la relazione tra temperatura dell'acqua e concentrazione di ossigeno disciolto è stato creato un diagramma a dispersione (Figure 7 e 8).Figura 7. Codice Python per rappresentare con un diagramma a dispersione la relazione tra temperatura (°C) e ossigeno disciolto (mg/L).7Figura 8. Diagramma a dispersione dell'ossigeno disciolto (mg/L) rispetto alla temperatura dell'acqua (°C).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()
```

Figura 7. Codice Python per rappresentare con un diagramma a dispersione la relazione tra temperatura (°C) e ossigeno disciolto (mg/L).7

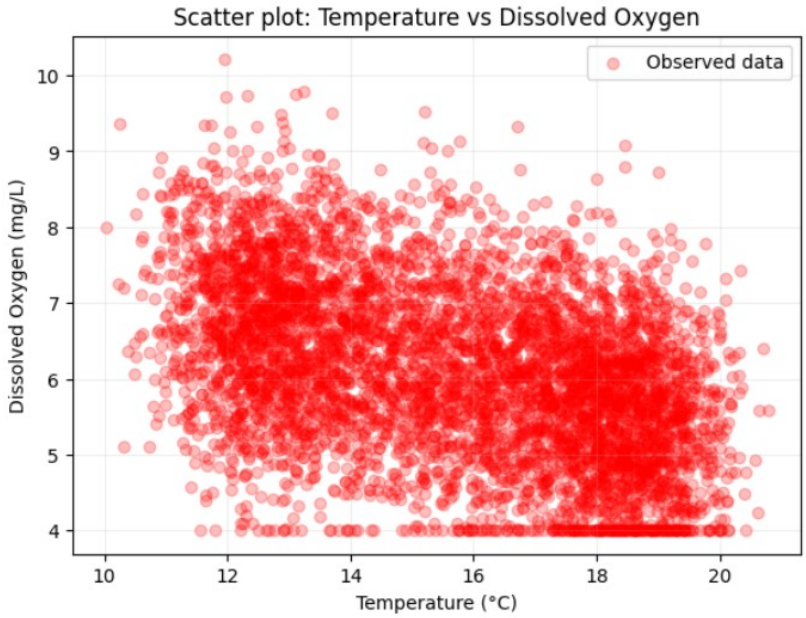

Figura 8. Diagramma a dispersione dell'ossigeno disciolto (mg/L) rispetto alla temperatura dell'acqua (°C).8

Il diagramma mostra una chiara tendenza negativa: l'ossigeno disciolto diminuisce all'aumentare della temperatura. Questo andamento è coerente con i principi fisici della solubilità dei gas nell'acqua. L'assenza di forti curvature o effetti soglia suggerisce che la regressione lineare possa essere una prima approssimazione adeguata. La relazione approssimativamente lineare osservata giustifica l'uso della regressione lineare come modello di riferimento.

Fase 3 – Preparazione del dataset per il modello di regressione

Sono stati selezionati dal dataset i predittori pertinenti e la variabile obiettivo.
La variabile obiettivo è la concentrazione di ossigeno disciolto (DO). I predittori comprendono la temperatura dell'acqua per la regressione semplice e temperatura, profondità e conducibilità elettrica per quella multipla.

Il dataset è stato suddiviso in addestramento e test con proporzione 80/20. L'insieme di addestramento serve per adattare i modelli, mentre quello di test è riservato alla valutazione indipendente delle prestazioni.
Questa separazione garantisce una valutazione non distorta della capacità di generalizzazione del modello.

Fase 4 – Modello lineare di riferimento (solo per confronto)

Addestramento del modello

Il dataset è stato suddiviso in addestramento e test (80/20). È stato adattato un modello di regressione lineare semplice usando la temperatura come unico predittore (Figura 9).Figura 9. Codice Python per addestrare una regressione lineare semplice che predice l'ossigeno disciolto dalla temperatura dell'acqua, con suddivisione addestramento–test e previsione.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)
```

Figura 9. Codice Python per addestrare una regressione lineare semplice che predice l'ossigeno disciolto dalla temperatura dell'acqua, con suddivisione addestramento–test e previsione.9

Equazione e prestazioni del modello

Le prestazioni sono state valutate mediante RMSE e R^2 (Figure 10 e 11).Figura 10. Codice Python per valutare il modello lineare e rappresentare la retta di regressione.10Figura 11. Output con equazione di regressione e metriche di prestazione (RMSE e R^2) del modello lineare semplice.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}")
```

Figura 10. Codice Python per valutare il modello lineare e rappresentare la retta di regressione.10

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

Figura 11. Output con equazione di regressione e metriche di prestazione (RMSE e R^2) del modello lineare semplice.11

Per valutare visivamente la relazione lineare adattata, è stata rappresentata la retta di regressione insieme ai punti osservati (Figure 12 e 13).Figura 12. Codice Python per rappresentare la retta di regressione lineare e i dati osservati.12Figura 13. Adattamento della regressione lineare tra temperatura dell'acqua e concentrazione di ossigeno disciolto.13

```
# 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()
```

Figura 12. Codice Python per rappresentare la retta di regressione lineare e i dati osservati.12

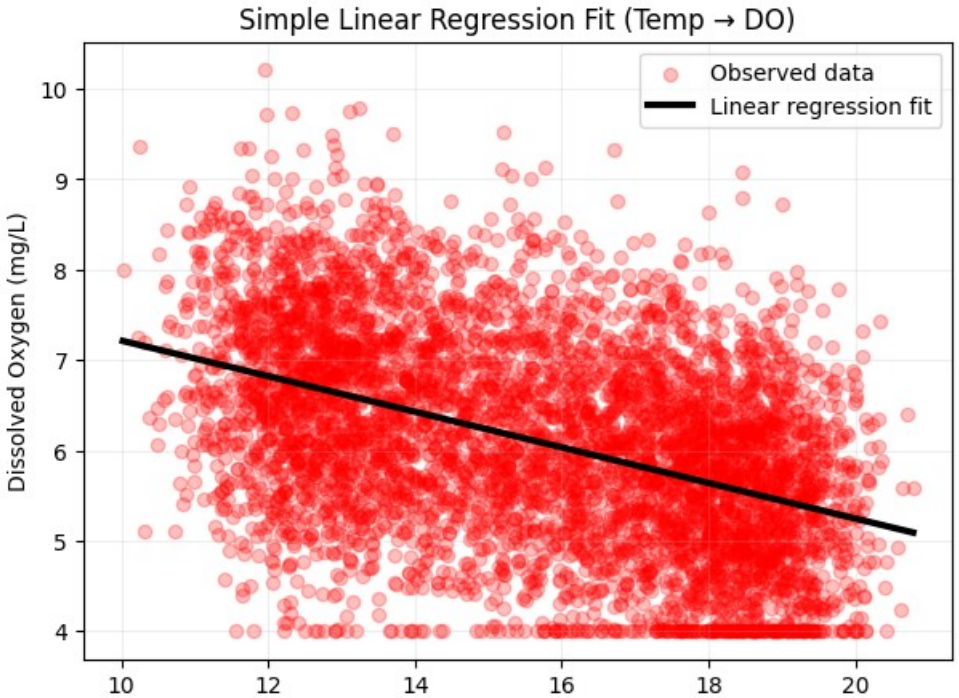

Figura 13. Adattamento della regressione lineare tra temperatura dell'acqua e concentrazione di ossigeno disciolto.13

Il coefficiente di regressione conferma una relazione negativa tra temperatura e ossigeno disciolto, coerente con il principio fisico secondo cui la solubilità dell'ossigeno diminuisce quando aumenta la temperatura dell'acqua.

L'RMSE di 0,964 mg/L indica l'errore di previsione del modello, mentre R^2 = 0,204 suggerisce che la sola temperatura spieghi circa il 20% della varianza della concentrazione di ossigeno disciolto.

Diagnostica del modello

Valori previsti rispetto ai valori osservati

È stato creato un grafico dei valori previsti rispetto a quelli osservati per verificare quanto le previsioni coincidano con le misure di ossigeno disciolto (Figure 14 e 15).Figura 14. Codice Python per il grafico dei valori previsti rispetto a quelli osservati per il modello lineare semplice.14Figura 15. Concentrazioni di ossigeno disciolto previste rispetto a quelle osservate nel modello lineare semplice.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()
```

Figura 14. Codice Python per il grafico dei valori previsti rispetto a quelli osservati per il modello lineare semplice.14

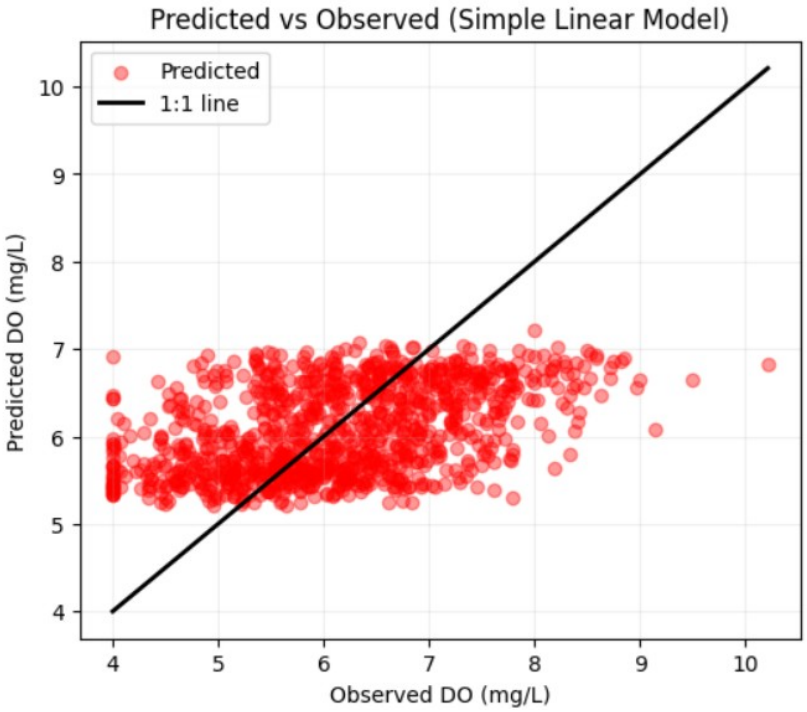

Figura 15. Concentrazioni di ossigeno disciolto previste rispetto a quelle osservate nel modello lineare semplice.15

I punti si concentrano attorno alla diagonale, indicando che il modello coglie la principale tendenza alla diminuzione dell'ossigeno disciolto all'aumentare della temperatura. Rimane tuttavia una dispersione evidente, a conferma che la sola temperatura non spiega completamente la variabilità della concentrazione di ossigeno disciolto.

Analisi dei residui

I residui sono stati calcolati come differenza tra valori osservati e previsti dell'ossigeno disciolto (Figure 16 e 17).Figura 16. Codice Python per calcolare e rappresentare i residui del modello lineare semplice.16Figura 17. Grafico dei residui del modello lineare semplice (ossigeno disciolto osservato meno previsto).17

```python
# 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()
```

Figura 16. Codice Python per calcolare e rappresentare i residui del modello lineare semplice.16

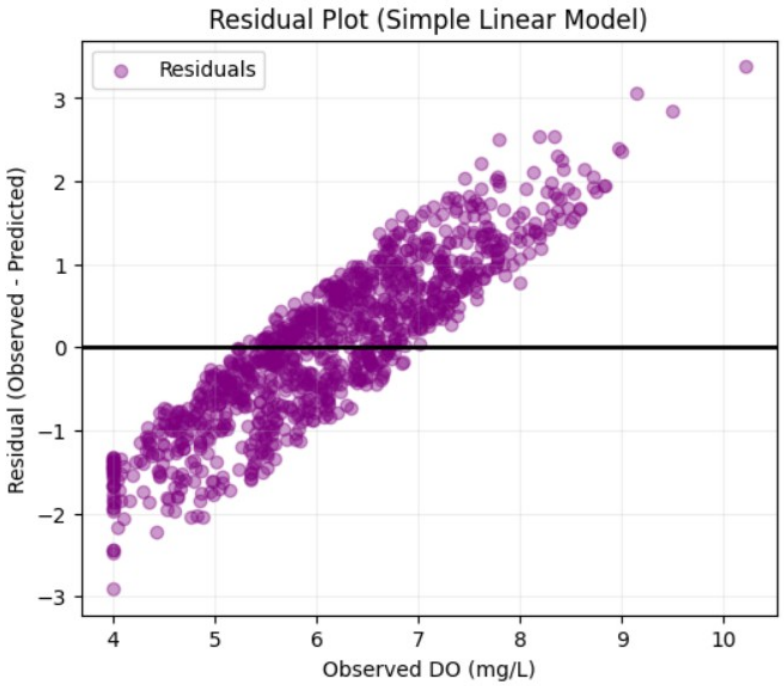

Figura 17. Grafico dei residui del modello lineare semplice (ossigeno disciolto osservato meno previsto).17

I residui sono distribuiti intorno allo zero senza una forte struttura sistematica: il modello lineare coglie quindi la tendenza dominante, ma lascia una variabilità non spiegata dovuta ad altri processi ecologici e fisici, quali il mescolamento delle acque e l'attività biologica.

Questi risultati evidenziano i limiti dell'uso della sola temperatura come predittore e motivano lo sviluppo di un modello di regressione lineare multipla con ulteriori variabili ambientali.

Modello di regressione lineare multipla (Temp + Depth + Conductivity → DO)

Addestramento del modello

Per migliorare le prestazioni predittive e tenere conto di ulteriori fattori fisici della variabilità dell'ossigeno disciolto, è stato sviluppato un modello lineare multiplo con temperatura, profondità e conducibilità elettrica come predittori.
I predittori sono stati scelti per la loro rilevanza ecologica. La temperatura controlla direttamente la solubilità dell'ossigeno, mentre profondità e conducibilità forniscono informazioni sulla struttura della colonna d'acqua e sui processi di mescolamento che influenzano la disponibilità di ossigeno.
Il dataset è stato suddiviso negli stessi insiemi di addestramento e test con proporzione 80/20 usati nel modello lineare semplice, per garantire un confronto equo. Il modello è stato addestrato sull'insieme di addestramento e utilizzato per effettuare previsioni sui dati di test.Figura 18. Codice Python per addestrare e valutare il modello di regressione lineare multipla con temperatura, profondità e conducibilità elettrica come predittori.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}")
```

Figura 18. Codice Python per addestrare e valutare il modello di regressione lineare multipla con temperatura, profondità e conducibilità elettrica come predittori.18

Equazione e prestazioni del modello

Le prestazioni del modello sono state valutate mediante errore quadratico medio radice (RMSE) e coefficiente di determinazione (R^2).

Rispetto alla regressione lineare semplice, quella multipla ha ottenuto un RMSE inferiore e un valore di R^2 superiore, indicando una maggiore accuratezza predittiva e una migliore

capacità di spiegare la variabilità dell'ossigeno disciolto.Figura 19. Output con i coefficienti di regressione e le metriche di prestazione (RMSE e R^2) del modello lineare multiplo.19

I coefficienti mostrano che la temperatura resta il predittore dominante, con coefficiente negativo che conferma la relazione inversa tra temperatura e concentrazione di ossigeno. Profondità e conducibilità elettrica contribuiscono ulteriormente alle previsioni, riflettendo le condizioni fisiche del sistema estuarino.

```
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
```

Figura 19. Output con i coefficienti di regressione e le metriche di prestazione (RMSE e R^2) del modello lineare multiplo.19

I risultati indicano che il modello lineare multiplo spiega circa il 42% della variabilità dell'ossigeno disciolto e migliora chiaramente il modello semplice basato soltanto sulla temperatura.

Valori previsti rispetto ai valori osservati

È stato creato un grafico dei valori previsti rispetto a quelli osservati per valutare visivamente l'accordo tra concentrazioni misurate e previste nel modello multiplo (Figura 20). I punti si dispongono più vicino alla retta 1:1 rispetto al modello semplice, segnalando un miglioramento delle prestazioni predittive.Figura 20. Codice Python per rappresentare valori previsti e osservati nel modello lineare multiplo.20

Gli scostamenti residui dalla diagonale indicano l'influenza di processi fisici e biologici non modellati, quali fotosintesi, respirazione e variabilità idrodinamica a breve termine (Figura 21).Figura 21. Concentrazioni di ossigeno disciolto previste rispetto a quelle osservate nel modello lineare multiplo.21

Complessivamente, il grafico conferma che l'aggiunta di profondità e conducibilità elettrica migliora la capacità di riprodurre le concentrazioni osservate di ossigeno disciolto.

```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()
```

Figura 20. Codice Python per rappresentare valori previsti e osservati nel modello lineare multiplo.20

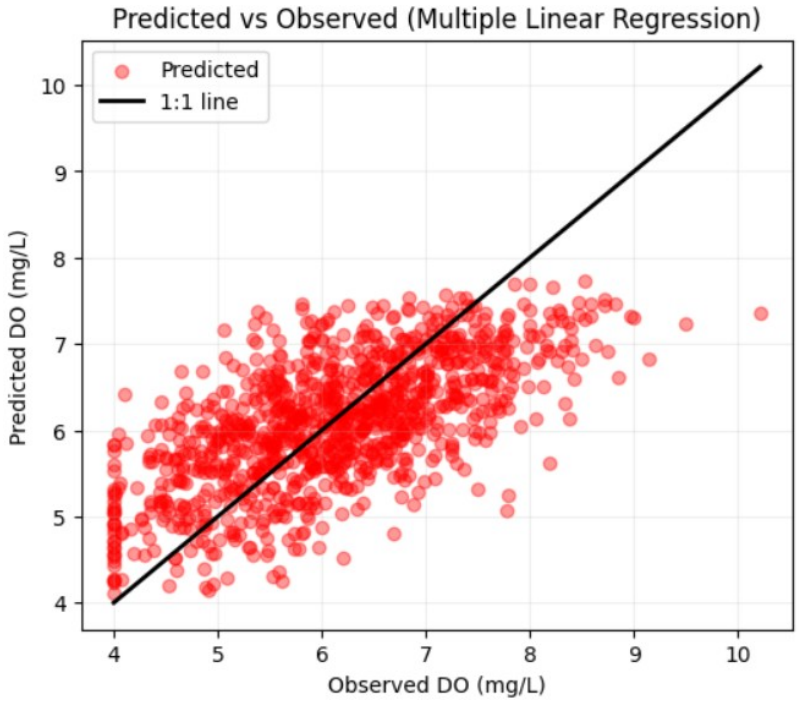

Figura 21. Concentrazioni di ossigeno disciolto previste rispetto a quelle osservate nel modello lineare multiplo.21

Analisi dei residui

Sono stati analizzati i residui per identificare errori sistematici e verificare la validità delle ipotesi del modello lineare (Figura 22).Figura 22. Codice Python per calcolare i residui e generarne il grafico per il modello lineare multiplo.22

```python
# 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()
```

Figura 22. Codice Python per calcolare i residui e generarne il grafico per il modello lineare multiplo.22

Il grafico mostra residui distribuiti intorno allo zero senza un andamento sistematico marcato, indicando che il modello lineare multiplo coglie le principali relazioni lineari tra ossigeno disciolto e predittori selezionati.Figura 23. Grafico dei residui del modello lineare multiplo: differenza tra concentrazioni di ossigeno disciolto osservate e previste.23

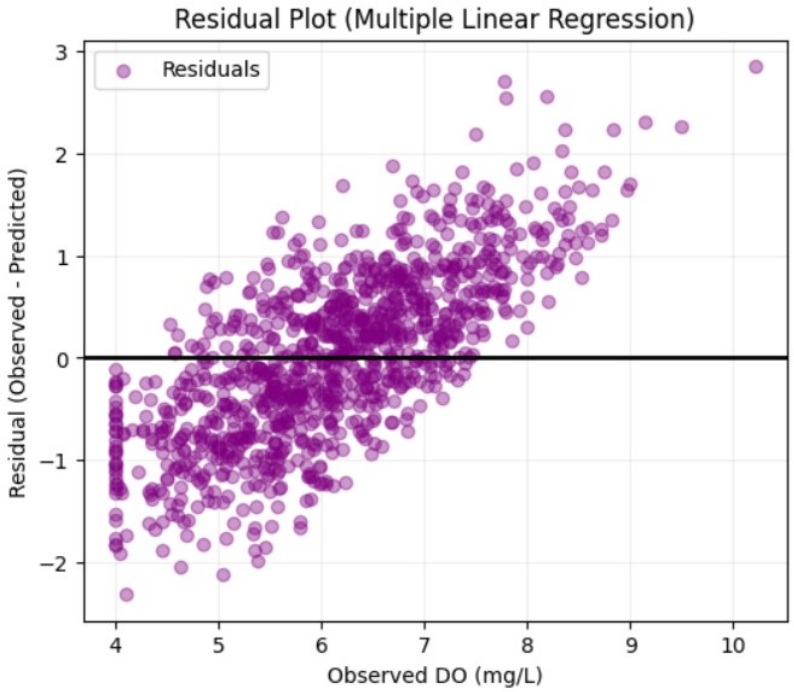

Figura 23. Grafico dei residui del modello lineare multiplo: differenza tra concentrazioni di ossigeno disciolto osservate e previste.23

Rispetto al modello semplice, i residui hanno una dispersione più omogenea, suggerendo una riduzione del bias e una maggiore stabilità quando vengono inclusi ulteriori predittori ambientali.

Rimane tuttavia una certa dispersione per temperature elevate e valori bassi di ossigeno, probabilmente dovuta a processi biologici e fisici non lineari, come fotosintesi,

respirazione e mescolamento idrodinamico di breve durata, non pienamente descritti dal modello lineare.

Nel complesso, l'analisi dei residui conferma che la regressione lineare multipla approssima ragionevolmente le dinamiche dell'ossigeno disciolto, evidenziando al tempo stesso la necessità di modelli più complessi e non lineari per spiegare la variabilità residua.

Confronto con la regressione lineare semplice

Il modello lineare multiplo è stato confrontato con il modello semplice che utilizza soltanto la temperatura dell'acqua per prevedere l'ossigeno disciolto.
Il confronto mostra un RMSE inferiore e un R^2 superiore per il modello multiplo rispetto a quello semplice. Ciò indica una maggiore accuratezza predittiva e una maggiore capacità di spiegare la variabilità dell'ossigeno disciolto.

Il modello semplice descrive la relazione inversa dominante tra temperatura e ossigeno disciolto, ma trascura altri fattori fisici importanti, quali struttura della colonna d'acqua e condizioni di mescolamento. L'inclusione di profondità e conducibilità aiuta invece a descrivere questi processi.

Il confronto mostra i limiti dei modelli con un solo predittore per i sistemi ecologici e sottolinea l'importanza di utilizzare più variabili dotate di significato ecologico nella modellazione ambientale.

Interpretazione ecologica

Il coefficiente negativo della temperatura riflette il principio fisico secondo cui la solubilità dell'ossigeno diminuisce all'aumentare della temperatura dell'acqua. L'acqua più calda contiene meno ossigeno disciolto, con conseguenze dirette per la respirazione e lo stress dei molluschi.

Il contributo di profondità e conducibilità elettrica evidenzia il ruolo della stratificazione, degli apporti di acqua dolce e del mescolamento nella disponibilità di ossigeno nell'estuario. Le variazioni di profondità influiscono sulla penetrazione della luce e sui processi biologici, mentre la conducibilità riflette le variazioni di salinità e il mescolamento tra acque dolci e marine.

I periodi con errori predittivi più elevati potrebbero corrispondere a episodi di intensa produzione fotosintetica di ossigeno durante il giorno o a eventi improvvisi di mescolamento causati da maree, vento o portata fluviale.

In sintesi, la regressione lineare multipla offre sia capacità predittiva sia informazioni ecologiche, collegando le dinamiche dell'ossigeno disciolto ai principali meccanismi fisici

e biologici delle aree estuarine destinate alla molluschicoltura. Ciò mostra come la regressione possa supportare il monitoraggio e la gestione ambientale.

Esercitazione pratica: regressione nelle scienze ambientali

Argomento: prevedere la concentrazione di clorofilla-a a partire da variabili oceanografiche fisiche

Lavoro: individuale

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

Dataset:
Copernicus Marine Service (CMEMS) – salinità superficiale del mare, clorofilla e velocità delle correnti.

Contesto e motivazioni

La biomassa del fitoplancton, generalmente rappresentata dalla concentrazione di clorofilla-a (Chl-a), è un indicatore fondamentale della produttività degli ecosistemi marini e della qualità dell'acqua. Le variazioni di clorofilla-a dipendono da fattori fisici come salinità e correnti oceaniche, che influenzano disponibilità di nutrienti, stratificazione e mescolamento.
In questa esercitazione gli studenti applicano tecniche di regressione per modellare la concentrazione di clorofilla-a mediante predittori ambientali ricavati dai dati del Copernicus Marine Service (CMEMS). L'attività mostra come la regressione lineare semplice e multipla quantifichi le relazioni tra variabili biologiche e fisiche nei sistemi marini e contribuisca al monitoraggio ambientale.

Descrizione del dataset

Il dataset deriva dai prodotti marini CMEMS e contiene serie temporali o campioni spaziali delle condizioni oceaniche superficiali. Ogni riga corrisponde a un punto di osservazione (tempo e/o posizione) e comprende le variabili seguenti:

Variabile obiettivo (y):
- Concentrazione di clorofilla-a (Chl-a, mg/m^3)

Variabili predittive (X):
- Salinità superficiale del mare (PSU)
- Velocità della corrente superficiale (m/s)

Descrizione dell'attività

Lo studente deve costruire e confrontare due modelli di regressione per prevedere la concentrazione di clorofilla-a:

1. **Regressione lineare semplice:**
   Chl-a = f(Salinità)
2. **Regressione lineare multipla:**
   Chl-a = f(Salinità, Velocità della corrente)

Guida passo per passo

Fase 1 – Configurazione

- Aprire Google Colab o Jupyter Notebook.
- Importare le librerie Python necessarie (pandas, numpy, scikit-learn, matplotlib).

Fase 2 – Caricamento ed esplorazione del dataset

- Caricare il dataset CMEMS dal file CSV fornito.
- Visualizzare le prime righe del dataset.
- Riportare:
  - Numero di osservazioni (righe)
  - Numero di variabili (colonne)
  - Variabile obiettivo
  - Variabili predittive
- Verificare la presenza di valori mancanti.

Fase 3 – Preparazione dei dati

- Convertire, se necessario, i valori scalati in unità fisiche.
- Rimuovere le colonne identificative o prive di significato fisico.
- Esaminare le statistiche descrittive e i possibili valori anomali.
- Suddividere il dataset in addestramento e test (80/20).

Fase 4 – Modello 1: regressione lineare semplice (modello di riferimento)

Addestrare una regressione lineare semplice usando la salinità come unico predittore.
Formulazione del modello:
Chl-a = f(Salinità)

- Adattare il modello ai dati di addestramento.
- Prevedere la clorofilla-a sui dati di test.
- Calcolare:
  - RMSE
  - R^2

Visualizzazioni:

- Diagramma a dispersione: salinità rispetto a clorofilla-a

- Retta di regressione
- Grafico dei valori previsti rispetto a quelli osservati
- Grafico dei residui

Cosa ci si aspetta dallo studente:
Spiegare brevemente il significato fisico del coefficiente di regressione (come la salinità sia associata alla concentrazione di clorofilla-a).

Fase 5 – Modello 2: regressione lineare multipla

Addestrare una regressione lineare multipla usando entrambi i predittori:
Formulazione del modello:
Chl-a = f(Salinità, Velocità della corrente)

- Adattare il modello ai dati di addestramento.
- Prevedere la clorofilla-a sui dati di test.
- Calcolare:
  - RMSE
  - R^2

Visualizzazioni:

- Grafico dei valori previsti rispetto a quelli osservati
- Grafico dei residui
-

Cosa ci si aspetta dallo studente:
Interpretare in termini ecologici i coefficienti della salinità e della velocità della corrente (mescolamento, trasporto, stratificazione).

Fase 6 – Confronto dei modelli

Confrontare i due modelli:

- Quale modello ha un RMSE inferiore?
- Quale modello ha un R^2 superiore?
- L'aggiunta della velocità della corrente migliora le prestazioni predittive?

Cosa ci si aspetta dallo studente:
Fornire una breve spiegazione (1–2 frasi) del perché la regressione multipla ottenga risultati migliori o peggiori rispetto a quella semplice.

Fase 7 – Interpretazione ecologica

Gli studenti dovrebbero discutere:

- Come la salinità influenzi il fitoplancton attraverso l'apporto di acqua dolce e di nutrienti.
- Come la velocità delle correnti influisca sulla distribuzione della clorofilla attraverso mescolamento e trasporto.

- Perché una parte della variabilità rimanga non spiegata (disponibilità di luce, nutrienti, processi biologici).

Materiale da consegnare – Relazione sull'esercitazione

Formato:
Jupyter Notebook o Google Colab con celle Markdown e di codice. La relazione deve contenere:

a. Introduzione
- Che cosa è stato modellato?
- Perché la clorofilla-a è importante per gli ecosistemi marini?

b. Descrizione dei dati
- Numero di campioni
- Variabili predittive
- Variabile obiettivo
- Gestione dei valori mancanti

c. Metodi
- Regressione lineare semplice
- Regressione lineare multipla
- Proporzione della suddivisione addestramento/test
- Metriche di valutazione (RMSE, R^2)

d. Risultati
- RMSE e R^2 di entrambi i modelli
- Grafici (dispersione, previsti rispetto a osservati, residui)
- Breve confronto tra i modelli

e. Discussione
- Interpretazione dei coefficienti
- Significato ambientale degli errori
- Limiti della regressione lineare

f. Conclusioni
- Sintesi di quanto appreso sulla regressione mediante dati ambientali marini

Valutazione

ESERCITAZIONE PRATICA DI REGRESSIONE

L'esercitazione pratica consente agli studenti di applicare modelli di regressione lineare semplice e multipla a un dataset ambientale reale del Copernicus Marine Service

(CMEMS). L'obiettivo è valutare la capacità di implementare, convalidare e interpretare modelli di regressione a supporto delle decisioni ambientali.

QUIZ

Il quiz valuta la comprensione concettuale e il ragionamento applicativo sviluppati durante la lezione e l'esercitazione sulla regressione nelle scienze ambientali. Verifica la comprensione delle differenze tra regressione semplice e multipla e del ruolo dei modelli nel monitoraggio e nella previsione ambientale.

BREVE RELAZIONE SCRITTA

La breve relazione scritta conclude l'esercitazione pratica, consentendo agli studenti di sintetizzare, interpretare e comunicare i risultati della modellazione mediante regressione.

Criteri di valutazione
1. QUIZ

Domanda 1
Che tipo di problema di machine learning è la previsione della concentrazione di ossigeno disciolto da variabili ambientali quali temperatura e profondità?
a) Classificazione
b) Regressione
c) Clustering
d) Rilevamento delle anomalie

Domanda 2
Quale variabile rappresenta l'obiettivo (variabile dipendente) nel caso applicativo della lezione?
a) Temperatura dell'acqua
b) Conducibilità elettrica
c) Concentrazione di ossigeno disciolto
d) Profondità

Domanda 3
Quale delle seguenti definizioni descrive meglio un problema di regressione nel machine learning?
a) Prevedere un'etichetta categorica
b) Prevedere un valore numerico continuo
c) Raggruppare campioni simili
d) Ridurre il numero di variabili

Domanda 4
Qual è la differenza principale tra regressione lineare semplice e multipla?

a) La regressione semplice è non supervisionata
b) La regressione multipla utilizza più di una variabile predittiva
c) La regressione semplice è sempre più accurata
d) La regressione multipla non può essere valutata

Domanda 5
Perché la regressione è considerata un metodo di apprendimento supervisionato?
a) Perché non richiede dati etichettati
b) Perché il modello apprende senza variabile obiettivo
c) Perché il modello viene addestrato su coppie note di input e output
d) Perché funziona solo su dataset piccoli

Domanda 6
Che cosa misura l'errore quadratico medio radice (RMSE)?
a) La proporzione di varianza spiegata
b) L'errore di previsione espresso nelle stesse unità della variabile obiettivo
c) Il numero di previsioni corrette
d) La correlazione tra variabili

Domanda 7
Qual è lo scopo principale della suddivisione di un dataset in addestramento e test?
a) Aumentare la complessità del modello
b) Rimuovere i valori anomali
c) Valutare le prestazioni del modello su dati mai visti
d) Ridurre il tempo di calcolo

Domanda 8
Quale situazione suggerisce che la regressione non lineare possa essere più adatta di quella lineare?
a) La relazione tra le variabili è una retta
b) I dati mostrano soglie o andamenti curvilinei
c) I dati non contengono rumore
d) È disponibile un solo predittore

Domanda 9
Quali informazioni fornisce un grafico dei valori previsti rispetto a quelli osservati?
a) Mostra la distribuzione dei predittori
b) Visualizza quanto le previsioni si avvicinino ai valori misurati
c) Individua i valori mancanti
d) Seleziona automaticamente i migliori predittori

Domanda 10
Perché i modelli di regressione sono utili nelle scienze ambientali?
a) Sostituiscono le misure fisiche
b) Aiutano a prevedere e interpretare variabili ambientali continue

c) Eliminano l'incertezza nelle osservazioni
d) Funzionano solo con dati di laboratorio

2. ESERCITAZIONE PRATICA DI REGRESSIONE

Preparazione e preprocessing dei dati

• Caricamento e ispezione corretti del dataset (dimensioni, predittori, variabile obiettivo).
• Gestione appropriata dei valori mancanti e conversione delle unità, quando necessario.
• Procedura di preprocessing trasparente e riproducibile (codice pulito e passaggi spiegati chiaramente).

Implementazione dei modelli

- Corretta configurazione e addestramento di almeno due modelli:
  - regressione lineare semplice
  - regressione lineare multipla.
- Uso adeguato della suddivisione addestramento/test.
- Codice organizzato e parametri esplicitati.

Valutazione e diagnostica dei modelli

- Calcolo corretto e presentazione delle metriche
  - RMSE
  - R^2.
- Confronto logico dei modelli (per esempio «la regressione multipla migliora R^2 rispetto a quella semplice»).
- Uso di grafici diagnostici (valori previsti rispetto a osservati, residui).
- Consapevolezza dei problemi di overfitting e underfitting.

Selezione e giustificazione del modello

- Giustificazione chiara della scelta del modello preferito.
- Spiegazione ragionevole della capacità di generalizzazione in base ai risultati sul test.

Visualizzazione e presentazione dei risultati

- Figure (diagrammi a dispersione, adattamenti di regressione, grafici dei residui) chiaramente etichettate.
- Visualizzazioni utili a interpretare il comportamento del modello.

Collegamento al contesto ambientale

- Lo studente spiega come i modelli di regressione possano contribuire al monitoraggio ambientale e ai sistemi di allerta precoce.

3. BREVE RELAZIONE SCRITTA

Chiarezza e struttura della relazione

• Relazione organizzata logicamente (Introduzione → Dati → Metodi → Risultati → Discussione → Conclusioni).
• Passaggi del processo visibili nel notebook.
• Linguaggio preciso e conciso.

Interpretazione e discussione dei risultati

• Risultati della regressione (RMSE, R^2, residui) presentati accuratamente e interpretati in modo significativo.
• Riflessione sui motivi per cui la regressione multipla ha funzionato meglio (o meno bene) rispetto a quella semplice.

Collegamento al contesto ecologico e ambientale

• Risultati collegati al significato ambientale reale (per esempio previsione della variabilità della clorofilla nel monitoraggio degli ecosistemi marini).
• Comprensione di come la regressione basata sui dati possa contribuire alle decisioni ambientali.

Qualità della presentazione grafica

• Figure chiaramente etichettate e richiamate nel testo.
• Elementi visivi utili a spiegare il comportamento del modello, non soltanto decorativi.

Comunicazione e stile

• Scrittura chiara, oggettiva e professionale.
• Uso corretto della terminologia del machine learning e ambientale (predittore, obiettivo, RMSE, residui, coefficiente).
• Riconoscimento dei limiti (modello semplificato, ipotesi di linearità, pochi predittori).

Pesi suggeriti

• Quiz: 20%
• Esercitazione pratica di regressione (notebook/codice): 50%
• Breve relazione scritta: 30%

Riferimenti bibliografici e letture di approfondimento

- 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.
- Portale di accesso ai dati del Copernicus Marine Service (CMEMS)
- Articoli scientifici ad accesso aperto (elenco completo disponibile sulla piattaforma e-learning del corso)