Relazione dataset n.11

**Introduzione**

Un ente locale ha monitorato la concentrazione in mare di una specifica alga tossica. Ci è stato fornito un dataset di osservazioni spaziali distinte in N stazioni. Abbiamo realizzato uno strumento previsionale che sulla base del meteo in una certa stazione, stima la presenza dell’alga nociva.

Abbiamo studiato il dataset per cercare di capire se la presenza della stazione influisse sulla previsione.

**1. Analisi dei dati**

Di seguito riportiamo le prime righe del dataset per mostrarne la struttura:

Immagine che contiene testo, Carattere, schermata, nero

Descrizione generata automaticamente

Abbiamo analizzato il dataset mediante la funzione “describe()” e abbiamo notato come la deviazione standard del dato di nostro interesse, ovvero la concentrazione, sia molto elevata rispetto al range di valori che assume.

Si nota inoltre che in seguito a piccole variazioni sugli altri parametri si hanno grosse variazioni su di essa.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

**1.1 Pulizia dei dati**

Abbiamo verificato la presenza di valori nulli mediante la funzione “isnull().sum()”,e abbiamo rimosso eventuali duplicati mediante “drop\_duplicates()”. Il dataset era già pulito e non presentava dati da rimuovere.

**1.2 Feature understanding**

Segue un’analisi dei dati in nostro possesso: come si può notare la temperatura dell’acqua, dell’aria e la pressione atmosferica seguono una distribuzione Gaussiana, seppur non completamente simmetrica.

Immagine che contiene schermata, diagramma, testo, Diagramma

Descrizione generata automaticamenteImmagine che contiene diagramma, schermata, Diagramma

Descrizione generata automaticamenteImmagine che contiene diagramma, schermata, Diagramma, testo

Descrizione generata automaticamenteImmagine che contiene diagramma, schermata, Diagramma

Descrizione generata automaticamenteSi noti che è stata normalizzata la colonna della densità in modo che l'area totale dell'istogramma sia pari a 1, in tal modo è possibile mostrare la curva di densità sovrapposta all’istogramma.

Immagine che contiene testo, schermata, Diagramma, linea

Descrizione generata automaticamente**Distribuzioni osservate**:

* **Tacqua**: La distribuzione è concentrata tra 20 e 30, con un picco intorno a 25.
* **Salinita**: La distribuzione è piuttosto concentrata intorno a 27-30.
* **Taria**: La distribuzione è più diffusa, ma c'è un picco principale intorno a 23-25.
* **Patm**: La distribuzione è concentrata intorno a 101500 con una variazione minore.

**1.3 Individuazione e gestione outliers**

Abbiamo sfruttato **pairplot** per avere una visione d’insieme riguardo alla distribuzione delle coppie di dati.

Diagonalmente abbiamo le distribuzioni delle singole features già analizzate in precedenza, mentre al di fuori di esse vediamo gli scatter plot tra le coppie.

Fornendo un parametro “hue” siamo andati a distinguere per colore i valori di concentrazione, fornendo quindi una visualizzazione supervisionata.

I punti colorati in rosso rappresentano le concentrazioni più alte; si può notare come compaiono in quantità ridotta rispetto al totale dei dati.

Immagine che contiene testo, diagramma

Descrizione generata automaticamente

Abbiamo dunque deciso di rimuovere da i plot i valori considerabili come outliers, per ottenere dei grafici più facilmente interpretabili.

Andando a rianalizzare i dati si ottiene una deviazione dei valori di concentrazione notevolmente inferiore (circa 2900), e si riescono anche a interpretare meglio i grafici.

Per eliminare gli outliers abbiamo sfruttato lo zscore, che indica quanti deviazioni standard quel valore è distante dalla media.

Si è deciso di considerare qualunque valore oltre le 5 deviazioni standard come un outlier in quanto ci risultava più facile analizzare ed utilizzare i dati.

Nel modello finale si è deciso di mantenerli in quanto i valori di concentrazione misurati sono per lo più bassi, sotto i 10.000.

Abbiamo trovato solo 21 valori sopra i 10.000, si ritiene dunque che tali valori di concentrazione siano importanti ai fini della previsione.

**1.4 Analisi dei dati e correlazione**

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

Immagine che contiene testo, diagramma

Descrizione generata automaticamente

Essendo comunque i grafici ancora poco leggibili abbiamo preferito categorizzare la concentrazione in dei range di valori che vanno da: 0 a 0-100, 100-1000, 1000-10000, 10000+.

Possiamo iniziare a trarre alcune conclusioni: la maggior parte delle misurazioni effettuate corrispondono a valori di concentrazione relativamente bassi, principalmente tra 0 e 1000, con molti valori corrispondenti a 0 esatto.

I restanti valori al di sopra di 1000 sono presenti in numero minore.

Immagine che contiene testo, diagramma

Descrizione generata automaticamente

Analizziamo ora i diversi parametri in rapporto tra loro:

* **Tacqua vs Salinita**: Non sembra esserci una correlazione chiara. I dati sono dispersi con qualche cluster intorno a valori intermedi di Salinita.
* **Tacqua vs Taria**: I dati mostrano una distribuzione più ampia ma senza una correlazione evidente.
* **Tacqua vs Patm**: Anche qui, non c'è una correlazione chiara, ma i dati sono distribuiti in un range limitato di Patm.
* **Salinita vs Taria**: La relazione mostra una dispersione con alcune concentrazioni di dati. Non c'è una correlazione forte visibile.
* **Salinita vs Patm**: I dati sono abbastanza sparsi senza mostrare una relazione diretta.
* **Taria vs Patm**: Mostra una dispersione dei dati con concentrazioni in specifici range di Taria e Patm.

Non sembrano esserci pattern chiari che colleghino i livelli di concentrazione con le variabili specifiche, suggerendo che la concentrazione potrebbe non essere direttamente influenzata dalle singole variabili rappresentate.

**Correlazione**

Analizzando la correlazione si può notare che la temperatura dell’acqua e quella dell’aria sono le variabili che influenzano in maggior modo la concentrazione, seppur in modo poco significativo.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

**Conclusioni**

Non ci sono correlazioni evidenti tra le variabili rappresentate.

Le concentrazioni più alte di "Conc" non seguono un pattern specifico in relazione alle altre variabili.

**1.5 Normalizzazione dei dati**

Si è deciso di provare a normalizzare i dati, con particolare attenzione alla colonna della “Concentrazione”, la quale avendo una varianza molto elevata risultava difficile da analizzare.

Attraverso il logaritmo si riescono a comprimere i valori elevati, rendendo la distribuzione più simmetrica e meno sensibile alle variazioni.

Si può in seguito applicare la trasformazione di Box-Cox per stabilizzare la varianza e rendere la distribuzione più simile ad una normale.

È stato aggiunto un valore di 1.01 a ciascun valore di concentrazione prima di applicare il logaritmo, in questo modo si è certi di ottenere valori strettamente positivi, piccoli a tal punto da non influenzare i risultati in modo significativo.

Si può subito notare come la correlazione sia stata influenzata dalla normalizzazione sulla colonna della concentrazione:

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Per completezza abbiamo deciso di sfruttare anche un **Box Plot** per un ulteriore rappresentazione visiva della distribuzione dopo la normalizzazione:

Immagine che contiene diagramma, testo, schermata, Disegno tecnico

Descrizione generata automaticamente

Dal grafico possiamo notare che la salinità presenta un range interquartile molto stretto, indicando che i dati sono più concentrati vicino alla mediana, mentre le altre variabili presentano una maggior dispersione dei dati.

I dati dopo la normalizzazione risultano abbastanza simmetrici, tuttavia sono ancora presenti diversi outliers di minor entità rispetto ai precedenti già rimossi.

**2. Scelta del modello predittivo**

**2.1 Kernel lineare**

Il kernel lineare è stato scelto come primo approccio per la sua semplicità e velocità di calcolo. È adatto quando la relazione tra le caratteristiche è approssimativamente lineare, nel nostro caso sapevamo già che essendo le relazioni non lineari non avrebbe portato a risultati ottimali, abbiamo voluto comunque verificare.

Di fatto il modello non riesce a prevedere concentrazioni che si discostano dalla media, come si può notare nel grafico.

Immagine che contiene testo, schermata, linea, Diagramma

Descrizione generata automaticamente

**2.2 Kernel polinomiale**

Il kernel polinomiale trasforma i dati in uno spazio di caratteristiche ad alta dimensione dove può essere più facile modellare relazioni complesse.

Permette di considerare combinazioni non lineari delle caratteristiche originali, migliorando le capacità predittive del modello.

La predizione mediante kernel polinomiale risulta più efficace, in quanto i dati predetti hanno un errore medio assoluto che va da circa 1500 fino a 3000.

Tuttavia, come si evince dal grafico, l’errore è basso perché la maggior parte dei valori di concentrazione rientra nella fascia 0-1000, il modello non è in grado di prevedere valori più elevati che si discostano dalla media.

L’accuratezza rilevata è leggermente maggiore, in una fascia dal 40% al 70% in base all’addestramento casuale.

Essendo che i dati in nostro possesso sono per la maggiore parte nella fascia citata, l’accuratezza rilevata risulterà spesso alta, andando però a escludere tutti i dati al di fuori della fascia.

Fino a qui non ci siamo preoccupati di valutare l’impatto della variabile categorica essendo i modelli non ottimali per il nostro problema.

Immagine che contiene testo, schermata, Diagramma, diagramma

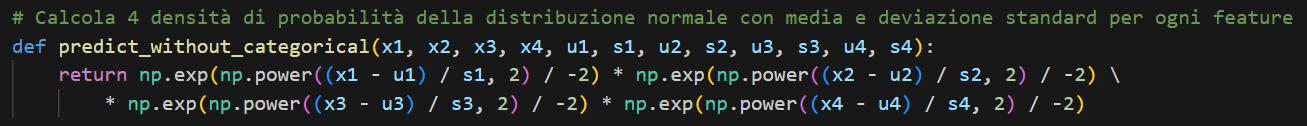
Descrizione generata automaticamente

**2.2 Discesa del gradiente**

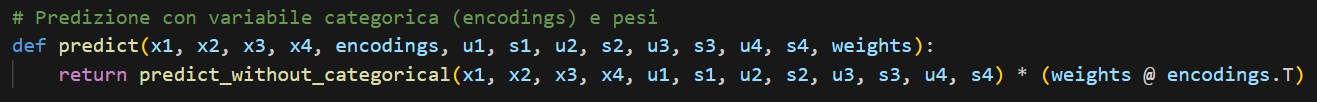
Data la precedente analisi dei dati, si è notato che le feature hanno una distribuzione che ricorda quella di una gaussiana.

Si procede calcolando 4 densità di probabilità della distribuzione normale con media e deviazione standard per ogni feature

Si calcola il prodotto delle 4 gaussiane per ottenere una nuova distribuzione delle feature.



Si sfrutta One-Hot Encoder per trasformare i dati categorici. La trasformazione converte ogni categoria in un vettore binario (one-hot encoded).

Si effettua la predizione con la variabile categorica (encodings), per ciascun valore di essa si avrà un peso associato .  


Si calcola il gradiente della funzione di errore rispetto a tutte le feature, come somma di derivate parziali.

Immagine che contiene testo, schermata, software

Descrizione generata automaticamente

Di seguito il codice in cui applichiamo la discesa del gradiente.

Si aggiorna il valore di ciascun parametro con l'opposto del gradiente moltiplicato per lo step, in quanto il gradiente è la direzione dove il loss aumenta maggiormente; quindi, si va nella direzione opposta per minimizzare la funzione di errore (loss).

Immagine che contiene testo, schermata, software, schermo

Descrizione generata automaticamente

I valori ottenuti per la loss dopo 10.000 iterazioni:

Immagine che contiene testo, Carattere, tipografia

Descrizione generata automaticamente

Calcoliamo l’errore:

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

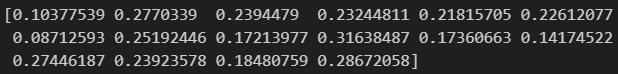
Infine, si calcola l’accuratezza del modello, andando a vedere quanti tra i valori predetti rientrano in un range di al massimo un ordine di grandezza come errore:

Immagine che contiene testo, Carattere, schermata

Descrizione generata automaticamenteImmagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamente

Osservando i pesi assegnati a ciascuna variabile categorica si nota che alcune influenzano maggiormente la previsione, seppur in modo poco marcato.



Essendoci pochi valori alti rispetto a quelli bassi, durante la discesa del gradiente lo step tende ad approssimare di più i valori bassi.

Immagine che contiene schermata, testo, diagramma, Diagramma

Descrizione generata automaticamente

Andiamo infine a verificare i risultati del modello senza variabile categorica.

La precisione ottenuta peggiora anche se in modo non troppo drastico, ne traiamo quindi che l’informazione riguardo alla stazione è rilevante ai fini della previsione.



Immagine che contiene schermata, testo, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, Carattere, schermata, tipografia

Descrizione generata automaticamente

**2.3 Boostrap resampling**

Essendo che il modello funziona ma continuava a riscontrare difficoltà nel prevedere concentrazioni elevate (in particolare quelle sopra i 10.000, valore di nostro interesse), abbiamo provato ad implementare una tecnica di oversampling, al fine di bilanciare la classe minoritaria (quella delle concentrazioni > 10.000).

Partendo dal dataset originale si selezionano elementi dalla classe target per generare un campione di bootstrap, si calcola la statistica di interesse e si ripete il processo diverse volte, generando diverse stime della statistica di interesse.

La funzione “resample” utilizzata sfrutta tale meccanismo al suo interno, esegue la selezione casuale con sostituzione dal set di dati originale, consentendo la selezione multipla delle stesse osservazioni.

Nel nostro caso abbiamo valutato essere sufficienti 140 valori nuovi per ottenere una previsione più accurata.

Immagine che contiene schermata, testo, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, Carattere, bianco, schermata

Descrizione generata automaticamente