# ANALISI STATISTICA DEL DATASET “WATER POTABILITY” E CLASSIFICAZIONE

Francesco Maria Fuligni

Matricola 0001068987

Università di Bologna

Corso di Laurea Triennale in Informatica per il Management

Esame di Statistica Numerica

Prof.ssa Elena Loli Piccolomini

18 Giugno 2024

## Descrizione del dataset

Il dataset "water_potability.csv" è stato scaricato dalla piattaforma *Kaggle* e contiene informazioni riguardanti l'analisi di diversi campioni di acqua. Per ognuno di essi, sono indicate 10 caratteristiche, ognuna costituente una colonna del dataset, di cui l'ultima è una variabile intera utilizzata come variabile target, mentre le altre sono variabili numeriche floating point:

1. **ph:** indica il livello di acidità o basicità del campione, nella scala di valori da 0 a 14. Affinché l'acqua sia potabile, la WHO indica che il valore del pH deve essere compreso tra 6.5 e 8.5;
2. **Hardness:** indica la durezza dell'acqua, data dalla concentrazione di calcio e magnesio, misurata in mg/L. Il Ministero della Salute consiglia un valore compreso tra i 150 mg/L e i 500 mg/L per questo dato;
3. **Solids:** indica i solidi totali presenti nel campione, in ppm (parti per milione). La WHO indica il range tra 0 ppm e 300 ppm per la migliore qualità dell'acqua;
4. **Chloramines:** indica la concentrazione di clorammine nel campione, in parti per milione; Tale concentrazione, secondo la WHO, deve essere compresa tra 0.5 mg/L e 1.5 mg/L affinché l'acqua sia potabile.
5. **Sulfate:** indica la concentrazione di solfati nel campione, in mg/L. La WHO consiglia per questo dato un valore massimo di 250 mg/L per garantire la potabilità.
6. **Conductivity:** indica la conducibilità elettrica dell'acqua, in μS/cm. Idealmente tale valore dovrebbe essere inferiore a 400 uS/cm;
7. **Organic carbon:** indica la quantità di carbonio organico nel campione, in ppm. Il Ministero della Salute consiglia un valore massimo di 10 ppm;

8. **Trihalomethanes:** indica la concentrazione di trialometani nel campione, in μg/L. La WHO stabilisce un valore limite per questo parametro di 100 μg/L;
9. **Turbidity:** indica la torbidità dell'acqua, misurata in NTU (Nephelometric Turbidity Units). Attraverso processi di filtrazione, questo parametro può essere ridotto fino a meno di 1 NTU, ma in generale non sono specificati valori fissi per la potabilità del campione.
10. **Potability:** indica la potabilità dell'acqua. Se il campione è potabile, ha valore 1, altrimenti ha valore 0.

Il dataset contiene in totale 3276 righe, con valori mancanti nelle colonne *ph* (491), *Sulfate* (781) e *Trihalomethanes* (162).

Si ricorda che i valori per la potabilità dell'acqua sono stabiliti dagli enti locali e possono variare a seconda dell'area geografica di riferimento. Non conoscendo la provenienza del dataset e il criterio esatto con cui è stato associato il valore di potabilità ad ogni campione, i valori limite indicati per ogni variabile sono puramente indicativi.

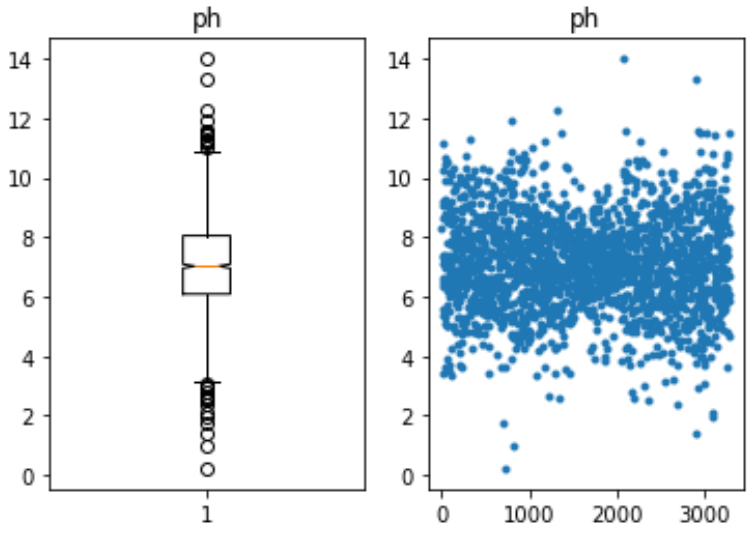
## Pre-processing

Nella prima fase di pre-processing è stato necessario eliminare dal dataset tutti i campioni che presentano valori mancanti. Ciò ha ridotto la dimensione del dataset dalle 3276 entry iniziali a 2011 entry. La rimozione dei valori mancanti (*NaN*) porta dunque alla perdita di circa un terzo del dataset, fattore che sarà discusso più avanti.

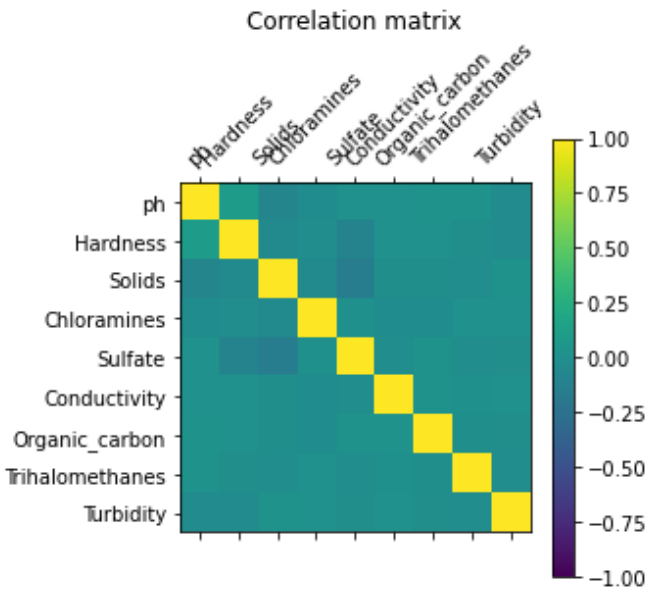
Tutte le features presentano valori numerici, ad eccezione di *Potability*, che rappresenta la variabile categorica di target per il successivo processo di classificazione.

## EDA

Nella fase di EDA sono stati esaminati i grafici boxplot e scatterplot per ogni variabile, in modo da osservare i dati a disposizione e l'eventuale presenza di outliers. Anche se sono stati evidenziati alcuni valori inferiori al range di 1,5 volte l'IQR del quantile inferiore o superiori al range di 1,5 volte l'IQR del quantile superiore, tramite il grafico boxplot, non è stato ritenuto necessario rimuovere alcun valore, per due ragioni: non si conosce la causa di tali misurazioni "estreme" e comunque tali valori sono del tutto verosimili.




La matrice di correlazione realizzata evidenzia una scarsa correlazione tra le variabili, il che porta ad ipotizzare scarsi risultati nella regressione lineare su due variabili. Le variabili più fortemente correlate, secondo tale matrice sono *Hardness* e *ph* (correlate positivamente) e *Sulfate* e *Solids* (correlate negativamente). Tali features verranno utilizzate per la fase di regressione lineare.




## Splitting

Il dataset è stato diviso in tre parti: training set, validation set e test set. I dati sono stati ripartiti in modo da destinarne il 70% per il training, il 15% per il validation e il restante 15% per il testing.

Lo splitting è stato svolto inizialmente con seed fissato a 8, per poi rimuovere questa condizione e ripetere l'esecuzione dello script con seed randomico, al fine di raccogliere i campioni casuali SRS(k) delle metriche per il modello predittore.

## Regressioni lineari

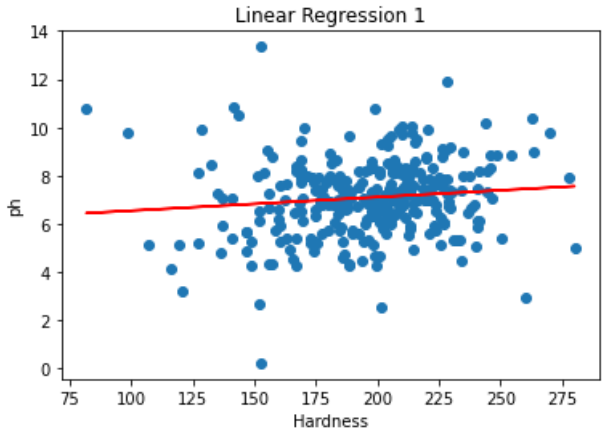
La fase di regressione lineare è stata svolta interamente con seed fissato a 8.

### 1. *Hardness - ph*

La prima regressione lineare è stata realizzata sulle variabili *Hardness* (variabile indipendente) e *ph* (variabile dipendente). Il modello è stato allenato sui valori di *Hardness* e *ph* ottenuti dal training set, mentre le previsioni del *ph* sono state ottenute a partire dai dati di *Hardness* contenuti nel test set (dunque non noti al modello, *out of sample*). I coefficienti ottenuti sono $\beta_0 = 5.98268$ e $\beta_1 = 0,00564$.

Le metriche *coefficiente semplice di determinazione* e *Mean Squared Error* sono state ottenute confrontando i valori di *ph* predetti con quelli contenuti nel test set. Il risultato ottenuto rispecchia quanto ipotizzato in precedenza: non avendo un livello di correlazione rilevante tra le variabili, il modello di regressione lineare non *fitta* bene i dati, risultando inadeguato. Lo si nota sia graficamente, osservando come i dati non si dispongano in forma lineare, sia osservando le metriche, dati $r^2 = 0,01975$ e $MSE = 2,50884$.

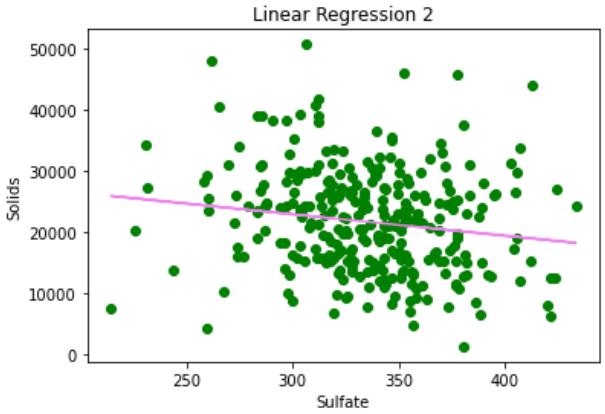
Per l'analisi di normalità dei residui è stato utilizzato il grafico QQ-Plot. Essendo le code dei residui molto distanti dalla retta che rappresenta la distribuzione normale, si conclude che i residui non presentano una distribuzione normale.

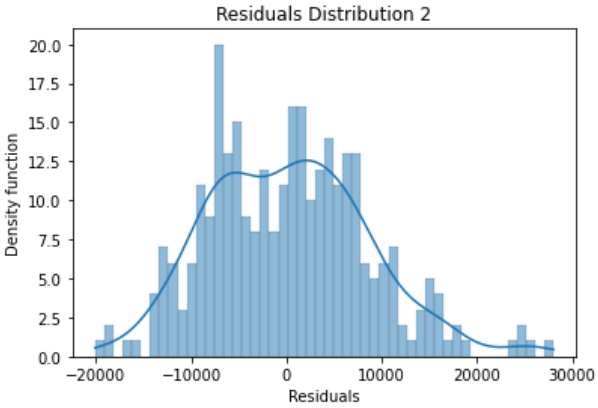



### 2. *Sulfate - Solids*

La regressione lineare sulle variabili *Sulfate* (indipendente) e *Solids* (dipendente) è stata realizzata in modo analogo e ha portato ad analoghe conclusioni. I coefficienti ottenuti sono $\beta_0 = 33{,}491$ e $\beta_1 = -35,01557$, mentre le metriche sono $r^2 = 0,03234$ e $MSE = 70{,}361{,}749$.

L'analisi di normalità dei residui è stata invece svolta utilizzando il Test di Shapiro-Wilk. Il $p_{value} = 0,0059973$ ottenuto, di molto inferiore al valore soglia 0,05, porta a rigettare l'ipotesi nulla di normalità: anche in questo caso, dunque, i residui non seguono una distribuzione normale. Ciò è riscontrabile anche visionando il grafico *Kernel Density Estimation*, con istogramma associato, non congruo con la forma caratteristica della curva Gaussiana.



## Addestramento del modello

L'addestramento del modello è stato svolto, ancora una volta, con seed 8. I modelli di regressione logistica e SVM (Support Vector Machines) sono stati allenati entrambi sulle variabili numeriche del training set, ottenuto in precedenza, e sulla variabile *Potability* come target categorico per la predizione.

## Hyperparameter Tuning

Per il modello SVM è stata svolta la fase di *hyperparameter tuining*. In tale processo, il modello, allenato sul training set, è stato testato sul validation set variando gli iperparametri in input alla funzione SVC.

Sono stati svolti test con kernel polinomiale (*poly*), radiale (Radial Basis Function, *rbf*) e sigmoide (*sigmoid*) e per ogni kernel sono stati svolti test variando *degree* o *gamma*. Data la limitata potenza computazionale a disposizione, il massimo degree testato è stato 11. Infine, per ogni test è stato calcolato il livello di accuratezza del modello validato.

Il confronto tra i livelli di accuratezza calcolati ha portato a scegliere il kernel *poly* con *degree* 5 come iperparametri migliori, in quanto hanno prodotto l'accuratezza maggiore, pari al 60%. Su tale modello è stato poi variato anche l'iperparametro *cost*, portando alla scelta del valore 1 come ottimale.

Si è osservato anche come valori di *degree* e *cost* superiori a quelli indicati portassero al fenomeno di *overfitting*, provocando una diminuzione dell'accuratezza del modello nella predizione della variabile target.

## Valutazione della performance

I modelli di regressione logistica e SVM ottenuti sono stati testati, sempre con seed 8, sul test set e per ognuno sono state valutate le metriche *Misclassification Error* (ME), *Misclassification Rate* (MR), *Misclassification Percentage* (Mper) e *Accuracy* (Acc). Il modello che ha mostrato le prestazioni migliori è stato il SVM con kernel polinomiale, grado 5 e costo 1, avendo un'accuratezza del 57,7% a fronte dell'accuratezza di 57,3% della regressione logistica.

Si è notato che, iterando più volte lo script con seed randomico, il modello che ha presentato le prestazioni migliori è sempre stato il SVM, perciò il campione di metriche è stato misurato su questo modello.

Tuttavia, un'accuratezza del 60% rimane un valore molto basso dal punto di vista statistico. Sicuramente, i fattori che hanno inciso negativamente sulle prestazioni del predittore sono il numero ridotto di informazioni a disposizione e la bassa correlazione tra esse.

Per tentare di migliorare questo risultato, sono stati svolti diversi tentativi aumentando le dimensioni del training set (non potendo agire sulla correlazione dei dati), ma senza particolare successo. Inoltre, l'aumento eccessivo dei dati riservati alla fase di training avrebbe danneggiato le fasi di validation e testing, compromettendole eccessivamente.

Un'altra possibile iniziativa sarebbe potuta essere quella di rimuovere le tre features che presentano valori mancanti (*ph*, *Sulfate* e *Trihalomethanes*), dal momento che la rimozione dei *NaN* riduce di un terzo la dimensione del dataset. Questa strada non è stata intrapresa perché tali features sono state ritenute molto importanti per il significato del dataset: infatti, sono variabili

cruciali per la determinazione della potabilità dell'acqua e la loro rimozione avrebbe provocato una significativa perdita di informazioni.

## Studio statistico sui risultati della valutazione

Eseguendo più volte lo script con seed casuale sono stati misurati campioni aleatori SRS(k) di $k = 15$ elementi per le metriche ME, MR e Acc relative al modello predittore SVM scelto. I campioni misurati sono poi stati registrati in un file .csv e processati da un altro script, per ottenere grafici e misure di statistica descrittiva e inferenziale sui dati.

Di seguito per semplicità, verrà preso in esame solamente il campione di *Accuracy*. Per le altre metriche valgono considerazioni del tutto analoghe.

### *Statistica Descrittiva*

Sono state calcolate più misurazioni del centro e della diffusione dei dati e della loro forma.

Il centro dei dati è stato calcolato attraverso la media semplice e la mediana semplice del campione. Sul campione di *Accuracy* è stata calcolata una media di 0,59289 e una mediana di 0,59334. I due valori calcolati sono molto simili, data l'assenza di dati estremi nel campione analizzato. Entrambe le misure indicano dunque un'accuratezza del modello leggermente superiore al 59%, che rimane, come detto, un valore basso dovuto alle caratteristiche dei dati presi in esame.

La diffusione dei dati è stata calcolata attraverso la varianza e la deviazione standard campionarie, il range interquartile (*IQR*) e la deviazione assoluta dalla media (*MAD*). Per l'*Accuracy* sono stati calcolati i seguenti valori:

- $S^2 = 0,00054793$
- $S = 0,023408$
- $IQR = 0,030$
- $MAD = 0,02470$

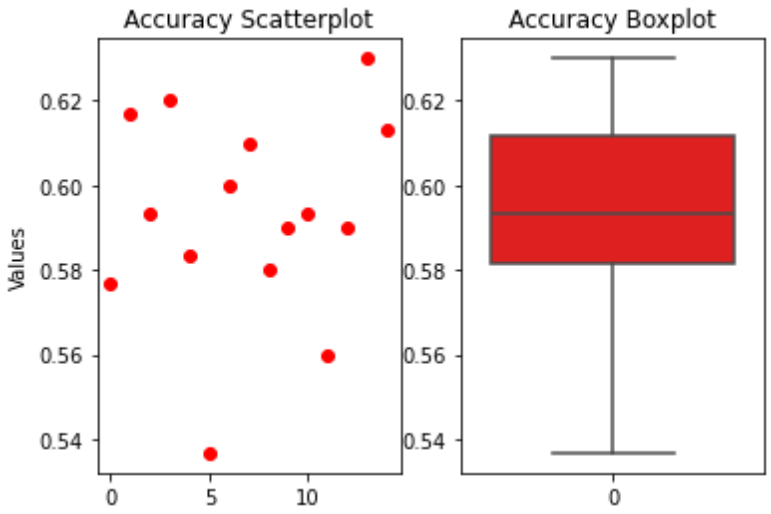
La bassa deviazione standard (e analogamente la varianza) indica una bassa dispersione dei dati, che tendono a concentrarsi intorno al valore medio. Quindi, si può dedurre che le prestazioni dell'algoritmo non sono particolarmente soggette a come viene effettuato lo splitting dei dati con seed randomico, anche se in alcuni casi rimane la possibilità di avere valori più dispersi (come il minimo di 0,54 e il massimo di 0,63). Anche la deviazione assoluta dalla media (MAD), piuttosto bassa, indica che i valori di *Accuracy* sono vicini alla media.

Il IQR indica invece la concentrazione della parte centrale dei dati, compresa tra i quantili a 25% e 75%. Anche in questo caso, avendo ottenuto un valore basso, i dati centrali mostrano una concentrazione elevata, visualizzabile anche graficamente.

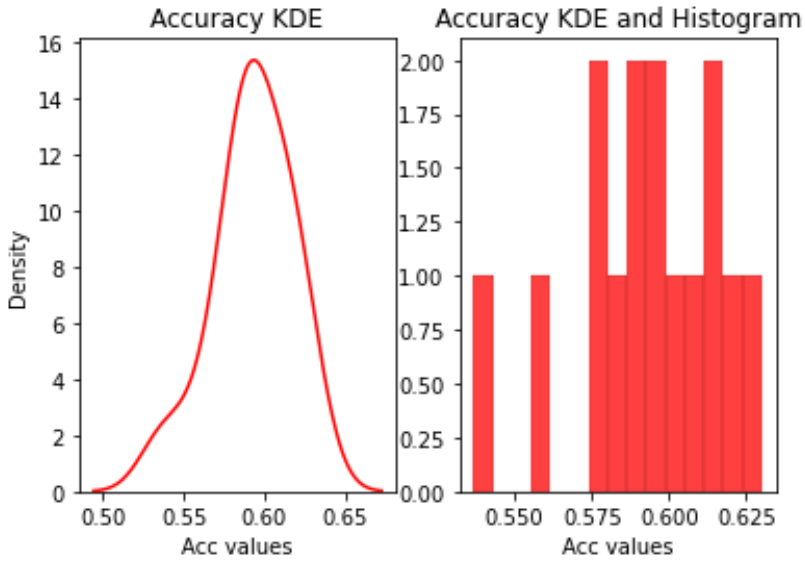
Inoltre, è stata valutata anche la forma dei campioni mediante misure della simmetria e della curtosi. Per l'*Accuracy* è stato ottenuto un valore di simmetria $g_1 = -0,63533$ e di curtosi $g_2 = 0,17208$. Il primo, essendo inferiore a 0, ma minore in valore assoluto di $\epsilon = 1,265$, evidenzia una lieve asimmetria a sinistra, come riscontrato anche nel grafico KDE. Il secondo, essendo maggiore di 0, ma comunque inferiore a $\epsilon = 2,530$, classifica la curva come mesocurtica.

***Grafici***

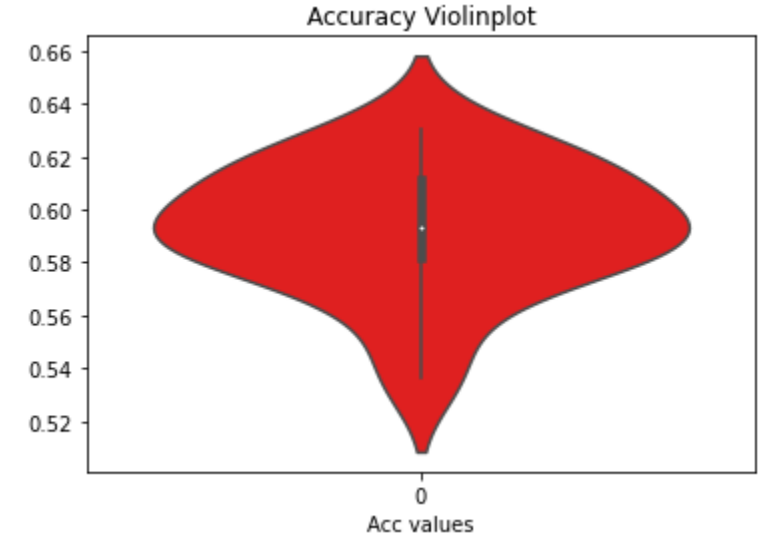
Sono stati realizzati diversi grafici per i dati di *Accuracy* campionati. Lo scatterplot e il boxplot mostrano come non vi siano valori estremi nei dati campionati.




Dal grafico KDE (realizzato a partire dall'istogramma) si possono osservare invece le caratteristiche della forma della distribuzione dei dati dell'*Accuracy*. In particolare, si osserva la lieve asimmetria a sinistra evidenziata in precedenza e anche un'elevata somiglianza con la distribuzione normale.



Infine è stato realizzato un violinplot come grafico riassuntivo, in quanto comprende in un'unica visione il boxplot e il KDE visti in precedenza.




.

***Statistica inferenziale***

È stato svolto il test di Shapiro-Wilk sui campioni, che non ha rigettato l'ipotesi nulla di normalità nel caso dell'*Accuracy*, avendo ottenuto un $p_{value} = 0,67411$ superiore al valore soglia 0,05.

È stato infine calcolato l'intervallo di confidenza per la media campionaria misurata precedentemente sui campioni esaminati. L'intervallo di confidenza è stato calcolato con $\alpha = 0,05$, cioè con un livello di confidenza del 95%. In particolare, essendo il numero di osservazioni campionate (15) di molto inferiore a 50 e la deviazione standard della popolazione non nota, è stato utilizzato il quantile della distribuzione t di Student, con 14 gradi di libertà ( $df = n - 1$) e la deviazione standard campionaria.

L'intervallo di confidenza calcolato per la media campionaria calcolata sul campione di *Accuracy*, pari a 0,59289, è $(0.57993,\ 0.60585)$. L'ampiezza piuttosto contenuta dell'intervallo

calcolato (meno di 3%) indica una buona precisione nella misura della media campionaria e rispecchia quanto già notato nell'analisi statistica, ovvero la forte concentrazione dei dati intorno al valore medio e quindi la bassa dispersione degli stessi.

Si osserva anche che in tale intervallo rientra l'*Accuracy* misurata sul modello SVM in fase di *hyperparameter tuning*, ma non i valori misurati in fase di valutazione dei modelli. Da questo risultato si può intuire la capacità, seppur limitata, del seed (fissato a 8 in quel caso) di influenzare (positivamente o negativamente) le prestazioni del modello, a seconda di come avviene lo *splitting* dei dati.

**Bibliografia**

E. Loli Piccolomini, A. Messina, *Statistica e Calcolo con R*, McGraw Hill, 2015

World Health Organization, *Guidelines for drinking-water quality, Fourth edition incorporating the first and second addenda*, 2022. Link alla fonte

Ministero della Salute, *Acque potabili*, 2016. Link alla fonte

Nayana CK, *Water Potability*, Kaggle, 2024. Link alla fonte