

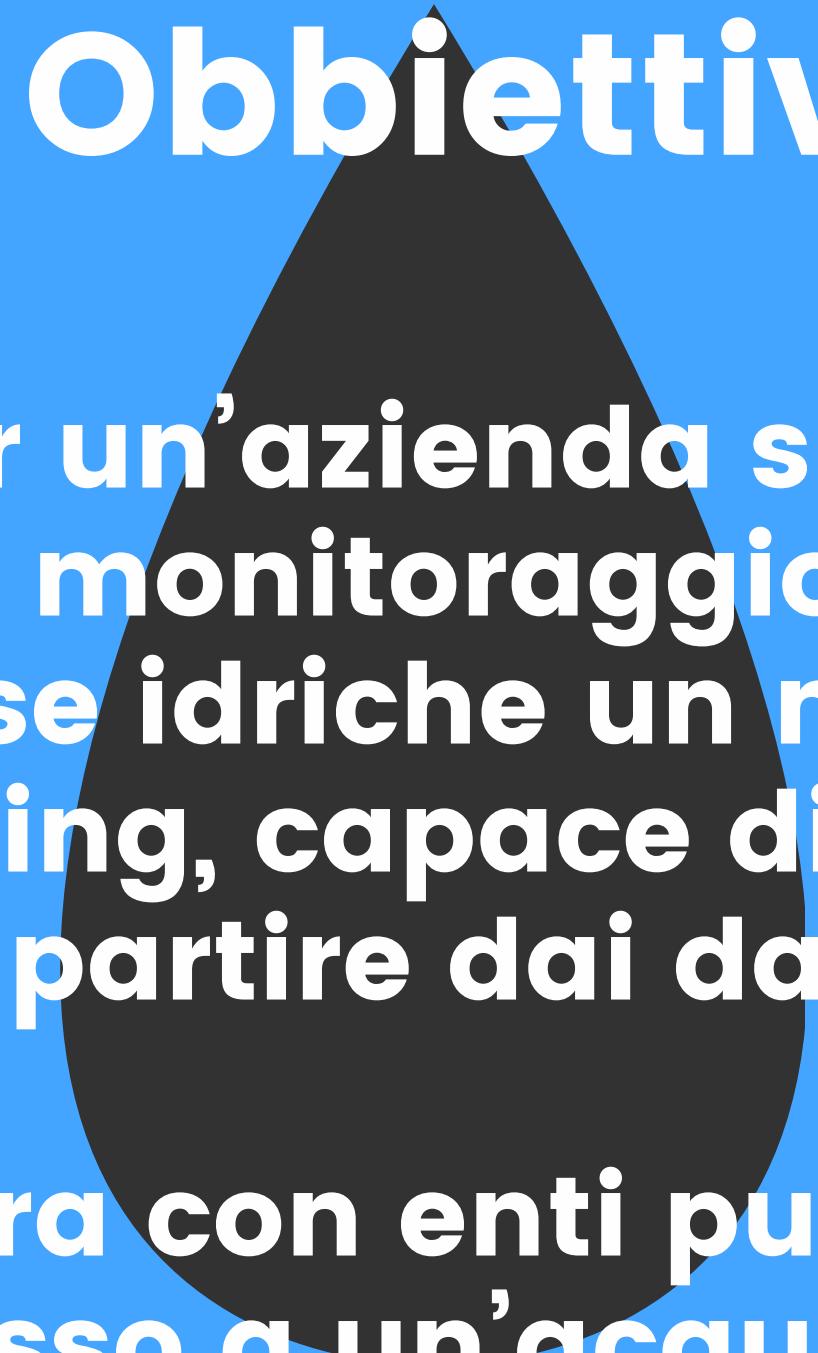
FRANCO DE GIORGIO

WATER INSIGHTS: MODELING

Previsioni sulla potabilità dell'acqua con il Machine Learning



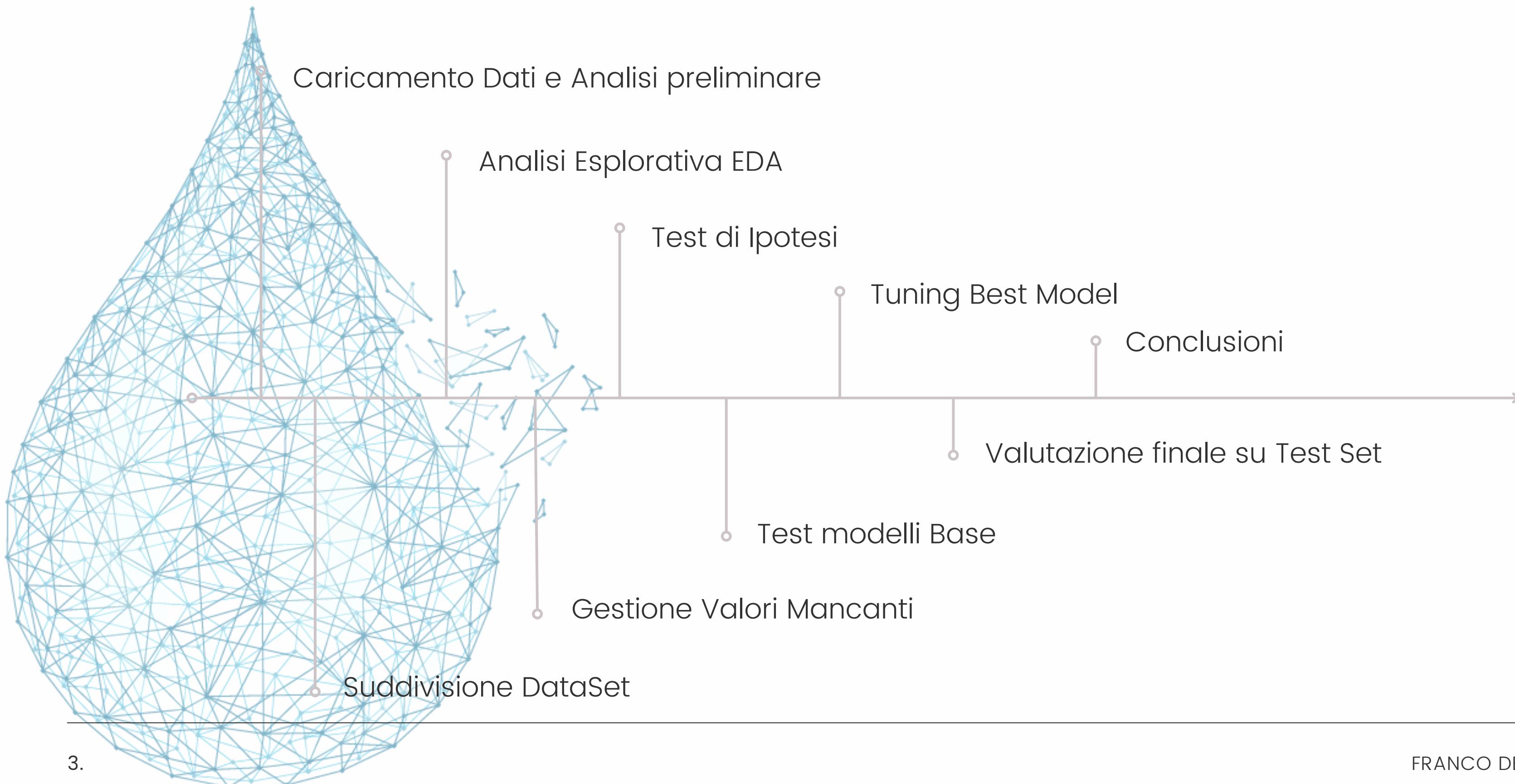
Obiettivo



Realizzare per un'azienda specializzata in tecnologie per il monitoraggio ambientale e la tutela delle risorse idriche un modello predittivo di machine learning, capace di fornire previsioni accurate a partire dai dati esplorati.

L'azienda collabora con enti pubblici e privati per garantire l'accesso a un'acqua pulita e sicura.

Fasi del Progetto



Caricamento Dati e Analisi Preliminare - Suddivisione DataSet

Nell'analisi preliminare sono emersi valori mancanti nelle colonne pH, Solfati e Trihalometani, ma non sono stati rilevati valori impossibili o fuori scala, come pH negativi.

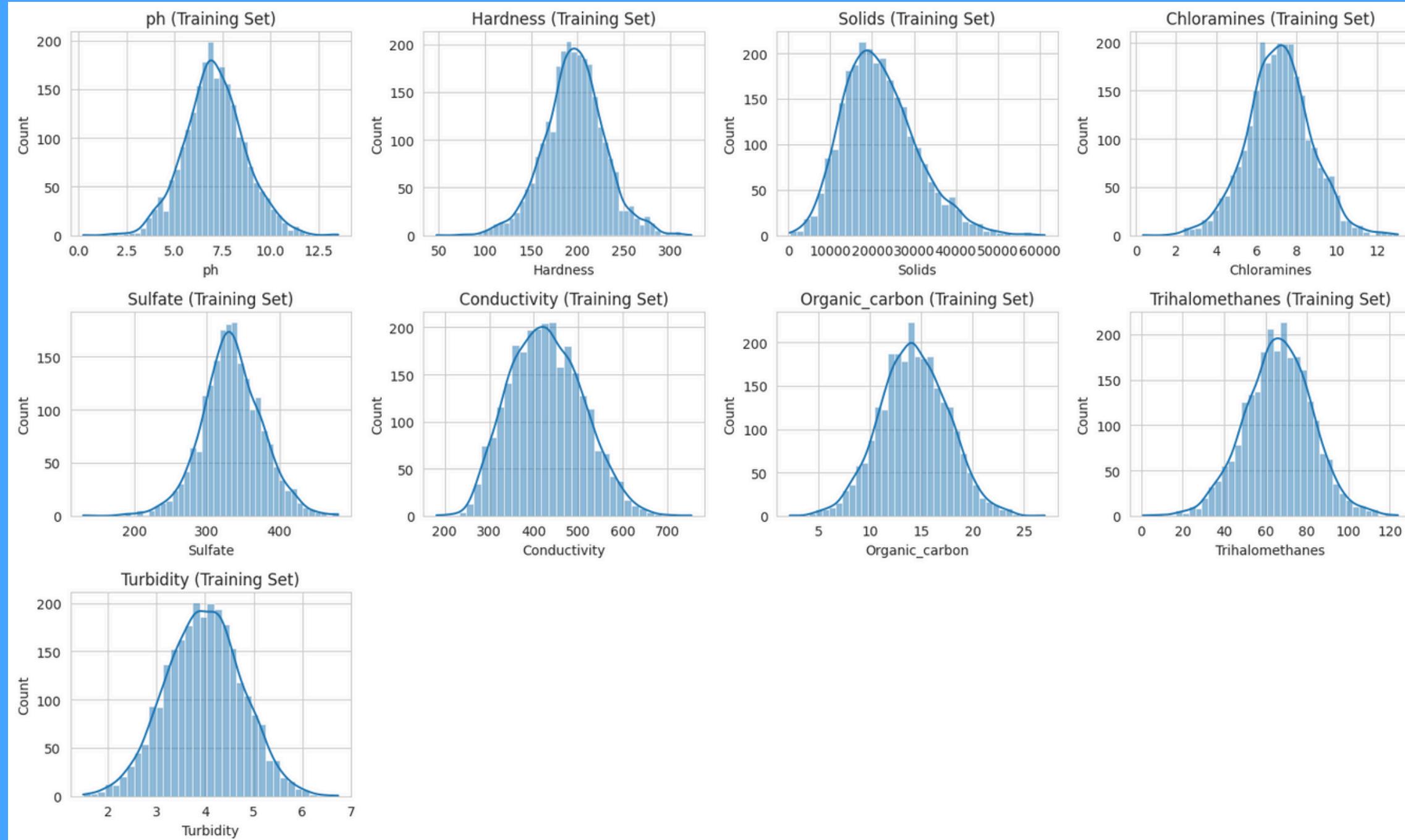
La distribuzione della variabile target ha evidenziato uno sbilanciamento: circa il 61% dei campioni riguarda acqua non potabile e il 39% acqua potabile.

Per gestire questo squilibrio, il dataset è stato suddiviso in training e test set utilizzando il parametro stratify.

Questa scelta ha garantito il mantenimento delle proporzioni tra le classi, evitando il rischio che una suddivisione casuale riducesse la rappresentanza dell'acqua potabile nei dati di test. In questo modo il modello può essere valutato in maniera più solida e realistica.

Successivamente sarà necessaria la normalizzazione delle variabili, poiché presentavano scale molto diverse tra loro. Questa fase è fondamentale per gli algoritmi che basano il calcolo su distanze o su tecniche di ottimizzazione, come K-Nearest Neighbors o la Regressione Logistica, così da rendere le caratteristiche confrontabili e migliorare la stabilità dell'addestramento.

Analisi Esplorativa EDA



L'analisi dei grafici di distribuzione (istogrammi e box plot) ha evidenziato che molte variabili seguono una forma quasi normale, come pH, durezza, sulfati, conducibilità e trialometani, pur con la presenza di alcuni outlier.

La variabile Solidi Totali Disciolti mostra invece una distribuzione asimmetrica con una lunga coda di valori molto alti, indicando casi con concentrazioni elevate che possono compromettere la potabilità.

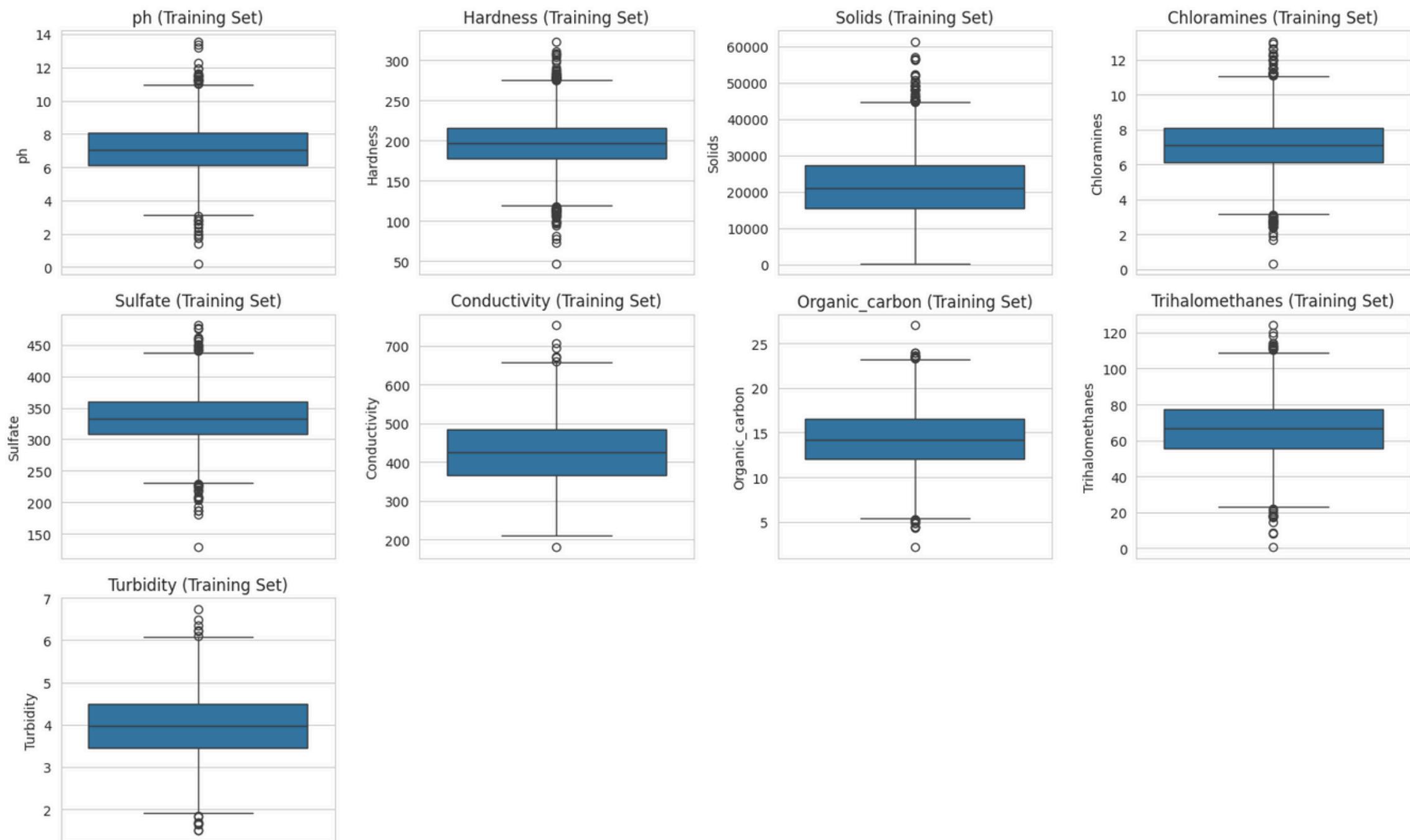
Altre variabili, come Clorammine e Carbonio Organico, risultano distribuite in modo regolare ma presentano valori anomali ai limiti estremi.

Analisi dei Grafici di Distribuzione

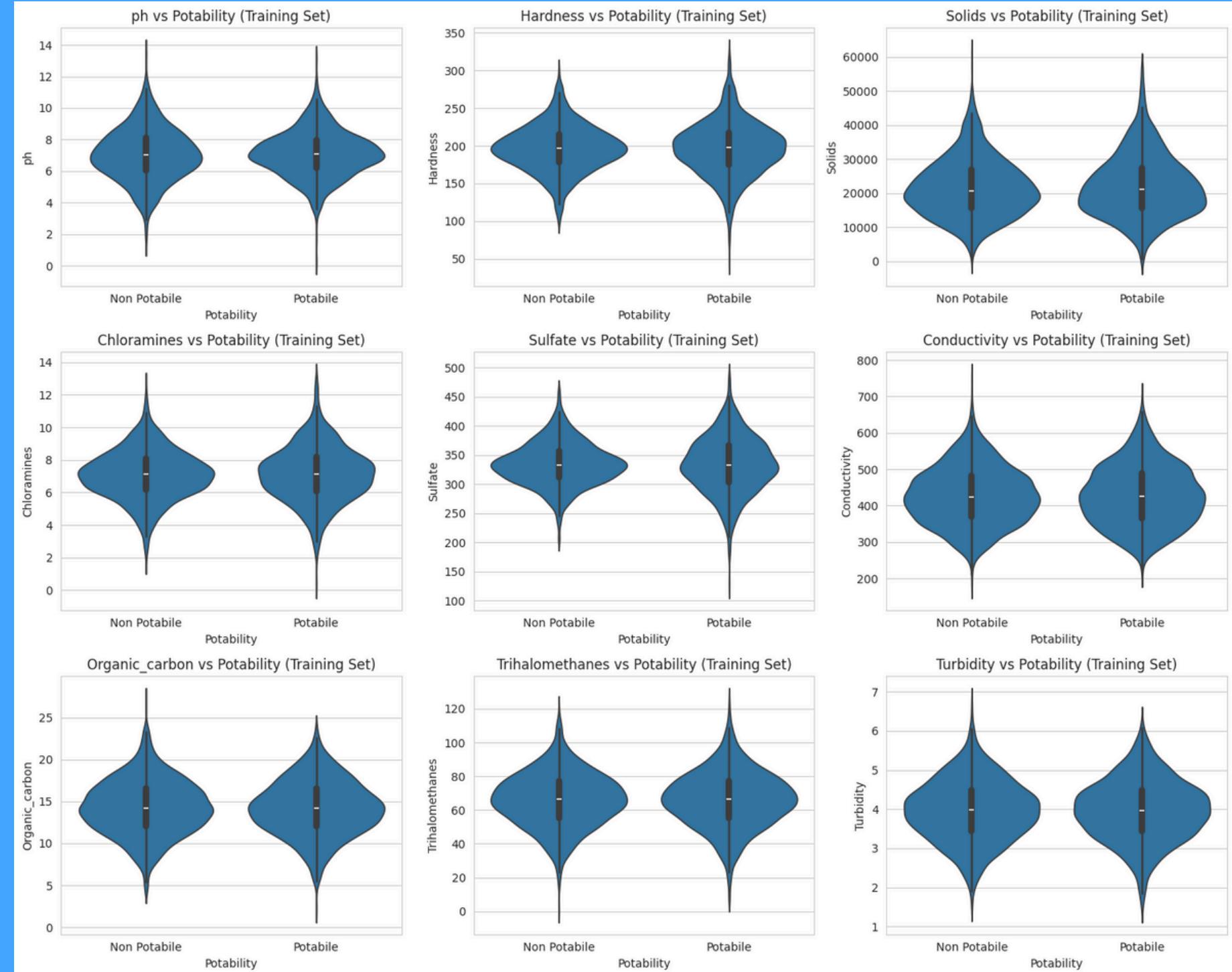
La Torbidità è mediamente normale, ma anch'essa con outlier in entrambe le direzioni.

Infine, la variabile target Potabilità, essendo binaria, conferma semplicemente lo sbilanciamento tra acqua potabile e non potabile già osservato nelle fasi iniziali.

In sintesi, la maggior parte delle caratteristiche mostra distribuzioni vicine alla normalità, ma gli outlier sono frequenti e i solidi disciolti rappresentano il caso più critico.



Analisi Violin Plot



Dai violin plot emerge che molte variabili, come pH, durezza, clorammine, solfati, carbonio organico, trialometani e torbidità, mostrano distribuzioni molto simili tra acqua potabile e non potabile, indicando che da sole non sono forti discriminanti.

Al contrario, solidi disciolti e conducibilità evidenziano differenze più marcate: l'acqua non potabile tende ad avere valori mediamente più alti e distribuzioni più estese. Questo suggerisce che possano essere variabili più informative per la classificazione.

In sintesi, solo alcune caratteristiche mostrano reali segnali distintivi, mentre per altre sarà necessario analizzare combinazioni e interazioni per individuare pattern più complessi.

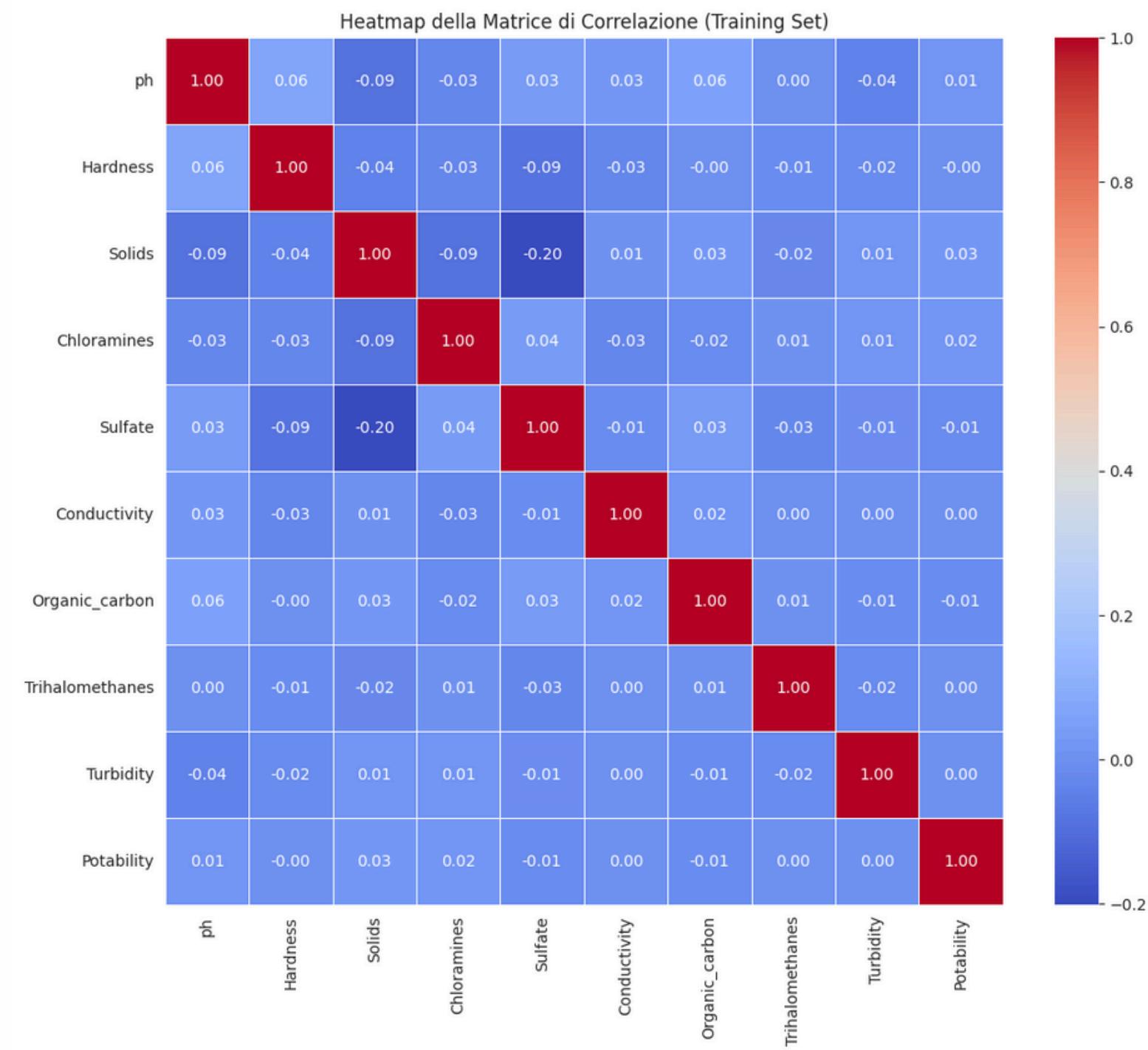
Analisi della Heatmap

La matrice di correlazione mostra che nessuna singola variabile è fortemente correlata con la potabilità: i valori sono tutti molto bassi, con un massimo di circa 0.03 per i solidi disciolti.

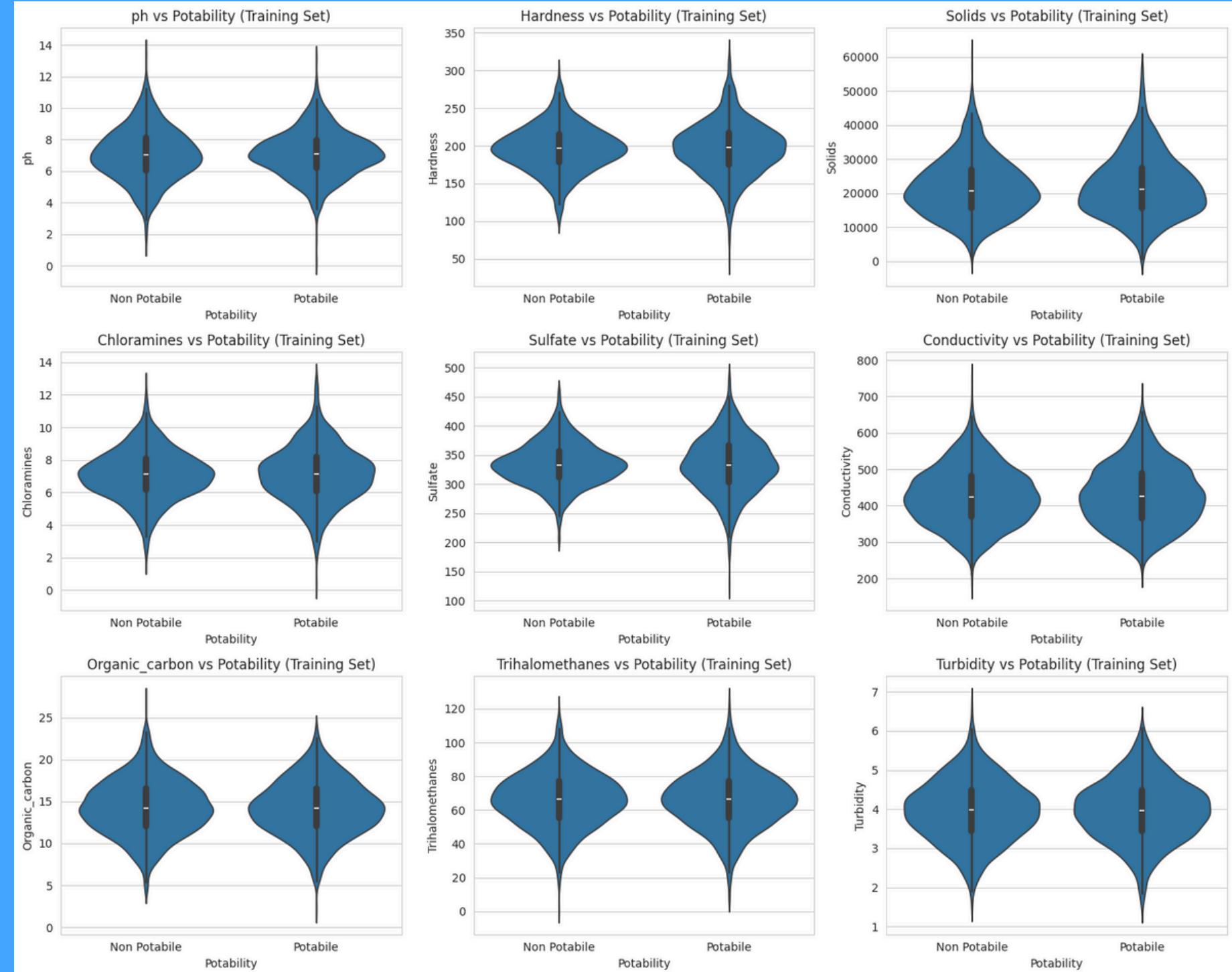
Questo conferma che la potabilità non dipende da una sola caratteristica, ma piuttosto da combinazioni di fattori.

Tra le variabili, si nota una correlazione negativa moderata tra Solidi e Solfati (-0.20), mentre le altre relazioni sono deboli. Nel complesso, non emergono problemi di multicollinearità, il che rende il dataset adatto a diversi modelli predittivi.

In sintesi, la heatmap evidenzia che sarà necessario un approccio multivariato, capace di cogliere le interazioni tra le caratteristiche, per ottenere buone prestazioni nella classificazione.



Analisi Violin Plot



Dai violin plot emerge che molte variabili, come pH, durezza, clorammine, solfati, carbonio organico, trialometani e torbidità, mostrano distribuzioni molto simili tra acqua potabile e non potabile, indicando che da sole non sono forti discriminanti.

Al contrario, solidi disciolti e conducibilità evidenziano differenze più marcate: l'acqua non potabile tende ad avere valori mediamente più alti e distribuzioni più estese. Questo suggerisce che possano essere variabili più informative per la classificazione.

In sintesi, solo alcune caratteristiche mostrano reali segnali distintivi, mentre per altre sarà necessario analizzare combinazioni e interazioni per individuare pattern più complessi.

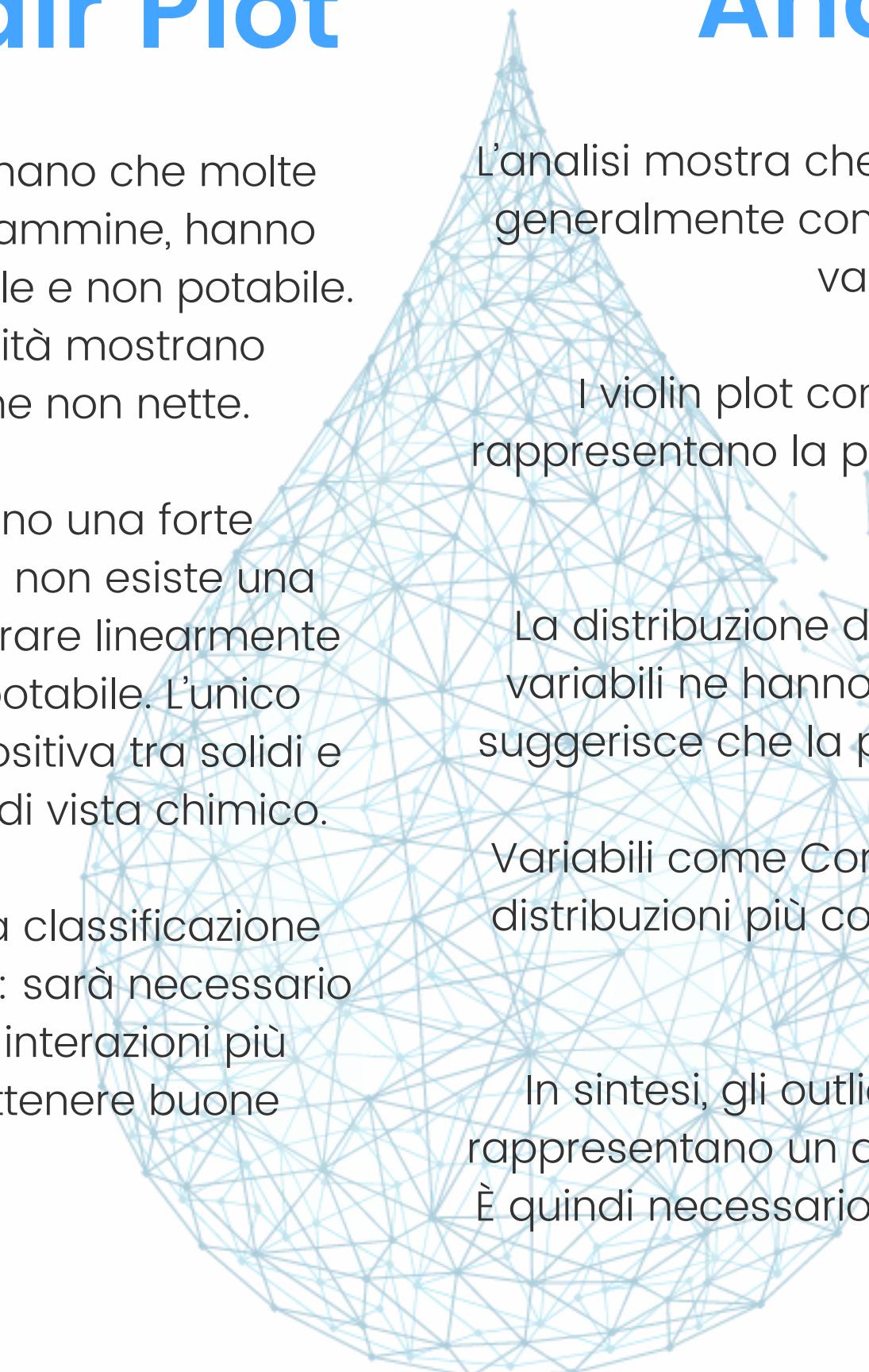
Analisi del Pair Plot

Le distribuzioni univariate confermano che molte variabili, come pH, durezza e clorammine, hanno valori molto simili tra acqua potabile e non potabile.

Solo solidi disciolti e conducibilità mostrano differenze più evidenti, sebbene non nette.

Le relazioni bivariate evidenziano una forte sovrapposizione tra le due classi: non esiste una coppia di variabili capace di separare linearmente l'acqua potabile da quella non potabile. L'unico legame chiaro è la correlazione positiva tra solidi e conducibilità, coerente dal punto di vista chimico.

In sintesi, il pair plot mostra che la classificazione non può basarsi su singole feature: sarà necessario un modello in grado di cogliere interazioni più complesse tra le variabili per ottenere buone prestazioni.



Analisi degli Outlier

L'analisi mostra che diverse feature contengono outlier, ma in percentuali generalmente contenute (sotto il 3-4%). Questo indica che non ci sono valori estremi diffusi in modo massiccio.

I violin plot confermano queste osservazioni: le code più lunghe rappresentano la presenza di outlier, come accade per Solids, soprattutto nella classe non potabile.

La distribuzione degli outlier non segue una tendenza univoca: alcune variabili ne hanno di più nell'acqua potabile, altre nel contrario. Questo suggerisce che la potabilità non dipende da singoli valori estremi, ma da combinazioni di fattori.

Variabili come Conductivity e Organic_carbon mostrano pochi outlier e distribuzioni più compatte, mentre Hardness e Chloramines evidenziano una maggiore dispersione.

In sintesi, gli outlier non compromettono la validità del dataset e non rappresentano un discriminante diretto tra acqua potabile e non potabile. È quindi necessario un modello in grado di cogliere relazioni complesse e interazioni tra le variabili.

Analisi degli Outlier con Isolation Forest



L'uso di Isolation Forest ha permesso di individuare outlier non solo in singole variabili, ma anche considerando combinazioni multivariate. Questi punti, pur non estremi in una sola dimensione, si discostano dal comportamento generale del dataset.

I risultati mostrano che gli outlier non sono numerosissimi e, soprattutto, potrebbero contenere informazioni utili: eliminarli indiscriminatamente ridurrebbe la qualità e la rappresentatività del training set. Inoltre, alcuni modelli come gli alberi decisionali sono più robusti agli outlier, mentre altri risultano più sensibili.

Per questo motivo, la scelta è stata quella di non rimuoverli in questa fase, valutando eventualmente approcci alternativi. In questo modo si preserva l'integrità del dataset, mantenendo aperta la possibilità di testare diverse strategie nei modelli successivi.



Gestione dei Valori Mancanti

Per le colonne ph, Sulfate e Trihalomethanes è stata scelta l'imputazione con la mediana, in quanto più robusta rispetto alla media di fronte agli outlier evidenziati nell'EDA.

La mediana riduce l'influenza dei valori estremi, preserva meglio la forma della distribuzione e fornisce una rappresentazione più fedele del centro dei dati.

Le mediane sono state calcolate solo sul training set ed applicate anche al test, rispettando uno scenario realistico in cui i dati futuri non sono noti durante l'addestramento.

Test di Ipotesi

Per confrontare le distribuzioni delle feature continue tra acqua potabile e non potabile, è stato scelto il Test U di Mann-Whitney, poiché:

- è adatto a confrontare due gruppi indipendenti,
- non richiede l'assunzione di normalità,
- è più robusto agli outlier rispetto al t-test.

I risultati mostrano che, per tutte le feature, il p-value > 0.05, quindi non è stato possibile rifiutare l'ipotesi nulla: non emergono differenze significative tra le distribuzioni delle feature considerate singolarmente.

Questo conferma quanto osservato in EDA: forte sovrapposizione tra le classi e correlazioni lineari molto basse con la variabile target. La potabilità sembra dunque dipendere da combinazioni complesse di feature, piuttosto che da singole proprietà, evidenziando l'importanza di modelli di machine learning capaci di catturare tali interazioni.

Analisi dello Spot Check dei Modelli Base

Lo spot check è stato condotto su Regressione Logistica, Random Forest Classifier e K-Nearest Neighbors utilizzando Stratified K-Fold Cross-Validation sul set di training. Le prestazioni sono state valutate non solo tramite accuratezza, ma anche con metriche più significative in presenza di dati sbilanciati: Precisione, Recall, F1-Score e AUC-ROC.

I risultati mostrano come la Regressione Logistica raggiunga un'accuratezza media di circa 0.61, sostanzialmente in linea con la proporzione della classe maggioritaria (acqua non potabile). Tuttavia, tutte le altre metriche si attestano a zero, segnalando che il modello non riesce a riconoscere correttamente la classe minoritaria. Questo conferma che le relazioni lineari, su cui si basa, non sono sufficienti per affrontare la complessità del dataset.

Modello	Accuratezza Media	Precisione Media	Recall Media	F1-Score Media	AUC-ROC Media
LogisticRegression_Pipeline	0.6099	0.0000	0.0000	0.0000	0.4775
RandomForestClassifier_Pipeline	0.6721	0.6484	0.3493	0.4535	0.6772
KNeighborsClassifier_Pipeline	0.6351	0.5436	0.4041	0.4635	0.6299

Analisi dello Spot Check dei Modelli Base

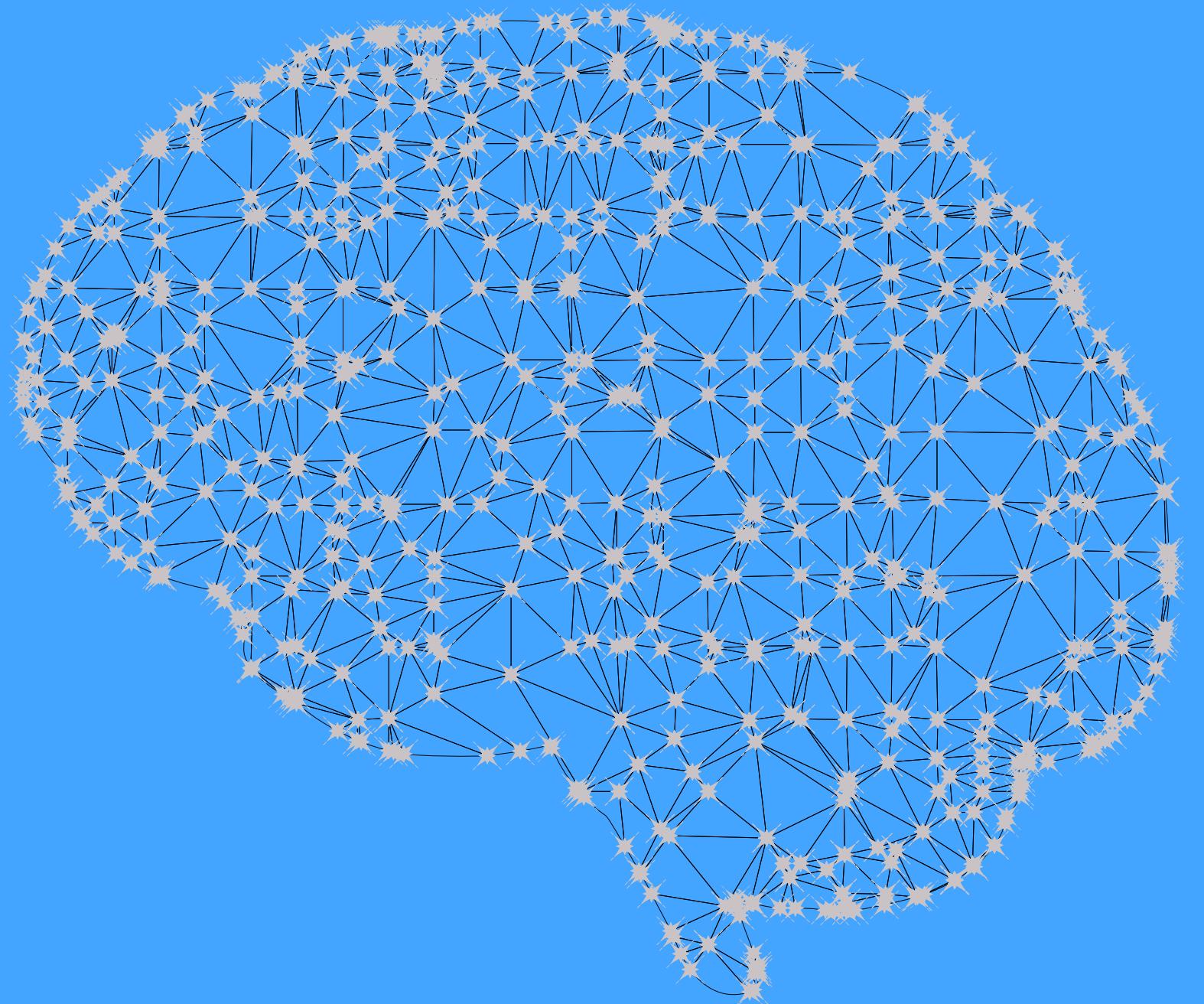
Random Forest e K-Nearest Neighbors hanno evidenziato prestazioni nettamente superiori. Random Forest si distingue per la sua capacità discriminatoria complessiva, con un AUC-ROC medio pari a 0.6772 e una precisione particolarmente elevata (0.6484), a indicare che quando classifica un campione come potabile tende a farlo con buona affidabilità. Il suo recall rimane però moderato, intorno a 0.35, e quindi il modello non intercetta una quota significativa di acque realmente potabili.

K-Nearest Neighbors ha mostrato invece un equilibrio leggermente diverso. Pur avendo una precisione inferiore rispetto a Random Forest (0.5436), ottiene un recall più alto (0.4041) e un F1-score medio marginalmente superiore (0.4635 contro 0.4535). Questo significa che KNN è più capace di riconoscere i campioni della classe minoritaria, anche se al costo di una maggiore frequenza di falsi positivi. Il suo AUC-ROC, pari a 0.6299, resta comunque inferiore rispetto a quello della Random Forest, segnalando una minore capacità globale di separare le due classi.

Il confronto conferma che la Regressione Logistica non è adatta a questo problema, mentre i due modelli non lineari riescono a catturare meglio la struttura dei dati. Random Forest appare più solido nella distinzione complessiva delle classi, mentre KNN mostra un vantaggio nel recall e nel bilanciamento tra precisione e sensibilità.

Alla luce di questi risultati, i modelli più promettenti su cui concentrare gli sforzi successivi sono Random Forest e K-Nearest Neighbors. Le prossime fasi saranno dedicate all'ottimizzazione dei loro iperparametri, con particolare attenzione al miglioramento di metriche come F1-score e Recall, che sono cruciali in un contesto applicativo in cui i falsi negativi (classificare acqua potabile come non potabile) potrebbero avere conseguenze più rilevanti rispetto ai falsi positivi.

Tuning dei Modelli



Dopo aver identificato Random Forest e K-Nearest Neighbors come i modelli più promettenti, procedo con l'ottimizzazione dei loro iperparametri tramite GridSearchCV, che valuta sistematicamente diverse combinazioni usando la cross-validation per individuare la configurazione migliore.

Per Random Forest vengono considerati parametri chiave come il numero di alberi, la profondità massima, i campioni minimi per split e foglia e il criterio di splitting, in modo da bilanciare prestazioni e rischio di overfitting. Per K-Nearest Neighbors, il focus è sul numero di vicini, sulla funzione di peso e sulla metrica di distanza, variando tra Manhattan ed Euclidea.

Poiché il dataset è sbilanciato, l'accuratezza non è indicativa: la selezione avviene quindi tramite l'F1-score, che bilancia precisione e recall e riflette meglio la capacità del modello di riconoscere la classe minoritaria. GridSearchCV viene impostato con scoring='f1' e refit='f1', mentre il ROC-AUC viene monitorato come indicatore aggiuntivo della capacità discriminatoria dei modelli.

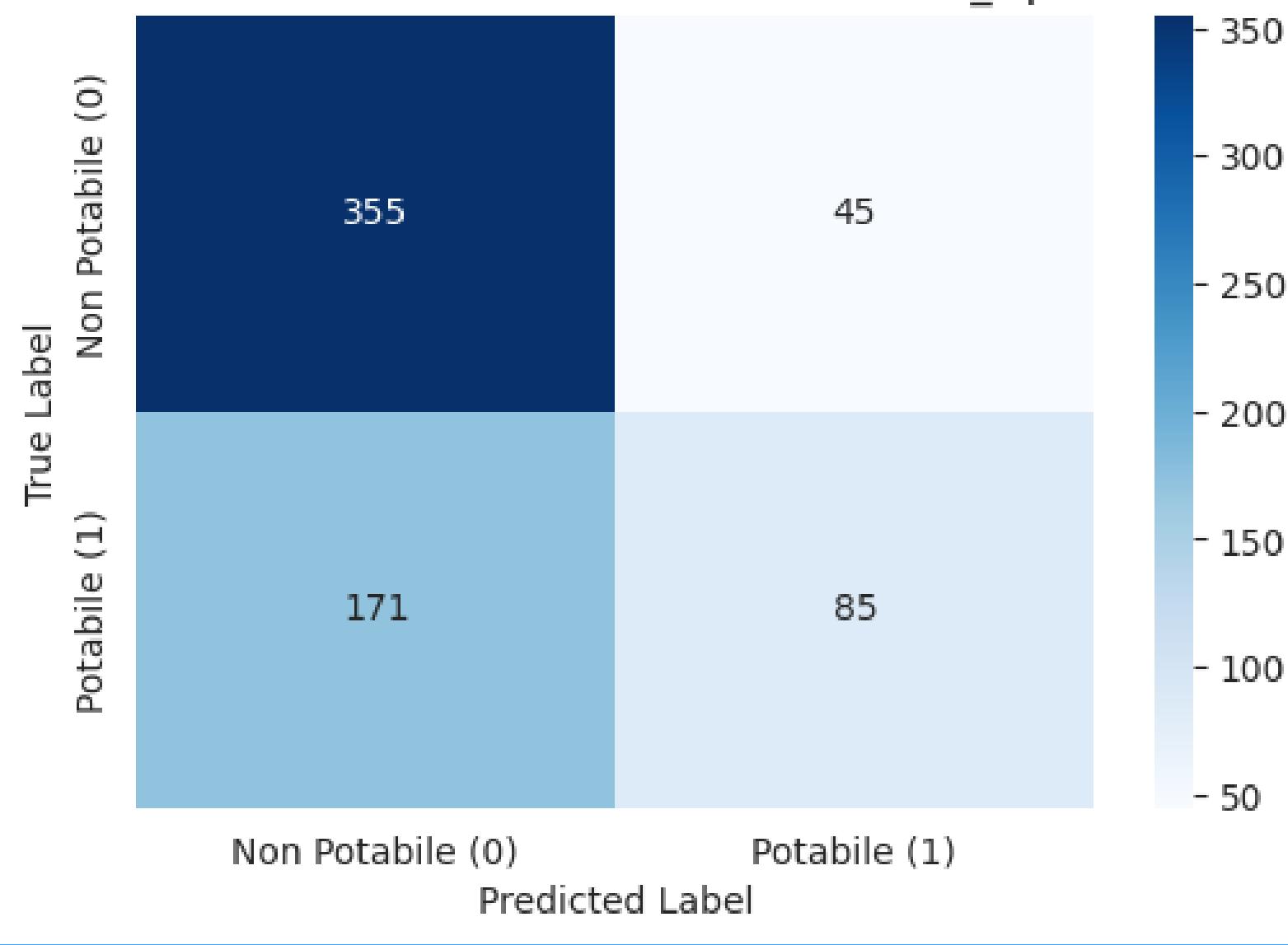
Analisi Finale e Conclusioni

Ho seguito un approccio standard partendo dall'esplorazione dei dati fino alla valutazione dei modelli finali. L'EDA sul set di training ha evidenziato valori mancanti in 'ph', 'Sulfate' e 'Trihalomethanes', uno sbilanciamento significativo tra la classe maggioritaria "non potabile" e la minoritaria "potabile", e una forte sovrapposizione delle distribuzioni delle feature tra le due classi. I test univariati (Mann-Whitney e t-test) hanno confermato che nessuna feature presa singolarmente mostra differenze significative tra le classi, suggerendo l'uso di modelli in grado di catturare interazioni e relazioni non lineari.

Per i valori mancanti ho scelto l'imputazione con la mediana, calcolata sul solo training set per evitare data leakage, e applicata anche al test set. Lo spot check su Regressione Logistica, Random Forest e K-Nearest Neighbors ha mostrato che i modelli lineari come la regressione logistica sono poco efficaci sul dataset sbilanciato, mentre Random Forest e KNN hanno ottenuto metriche significativamente migliori in Precisione, Recall, F1-score e AUC-ROC.

L'ottimizzazione degli iperparametri tramite GridSearchCV ha confermato che KNN con 3 vicini, distanza Manhattan e pesi basati sulla distanza raggiunge un F1-score leggermente superiore in cross-validation, ma sul test set Random Forest ottimizzato ha ottenuto risultati complessivamente migliori, con Accuratezza 0.6707, Precisione 0.6538, F1-score 0.4404 e AUC-ROC 0.6539. Entrambi i modelli mostrano un recall basso per la classe positiva (circa 33-34%), indicando che solo un terzo delle acque realmente potabili viene identificato correttamente. La matrice di confusione conferma che Random Forest commette meno falsi positivi rispetto a KNN.

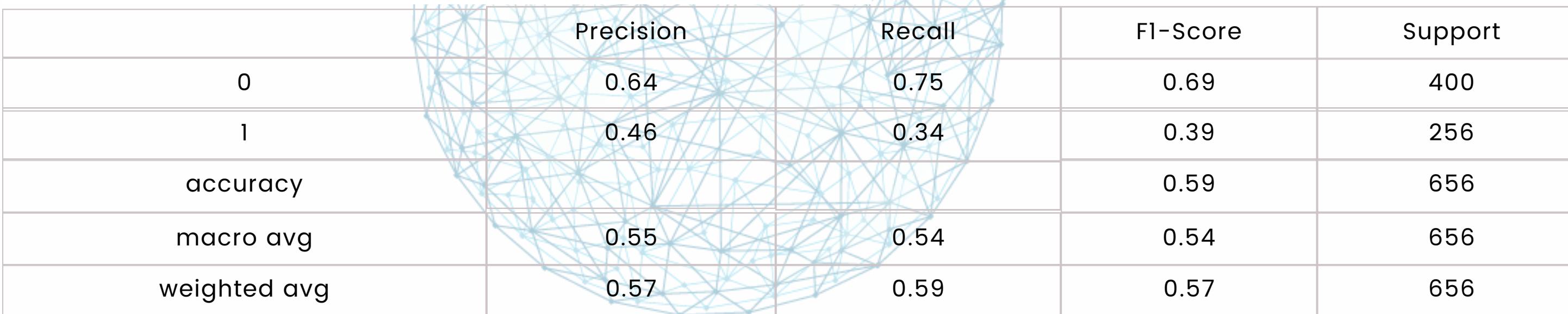
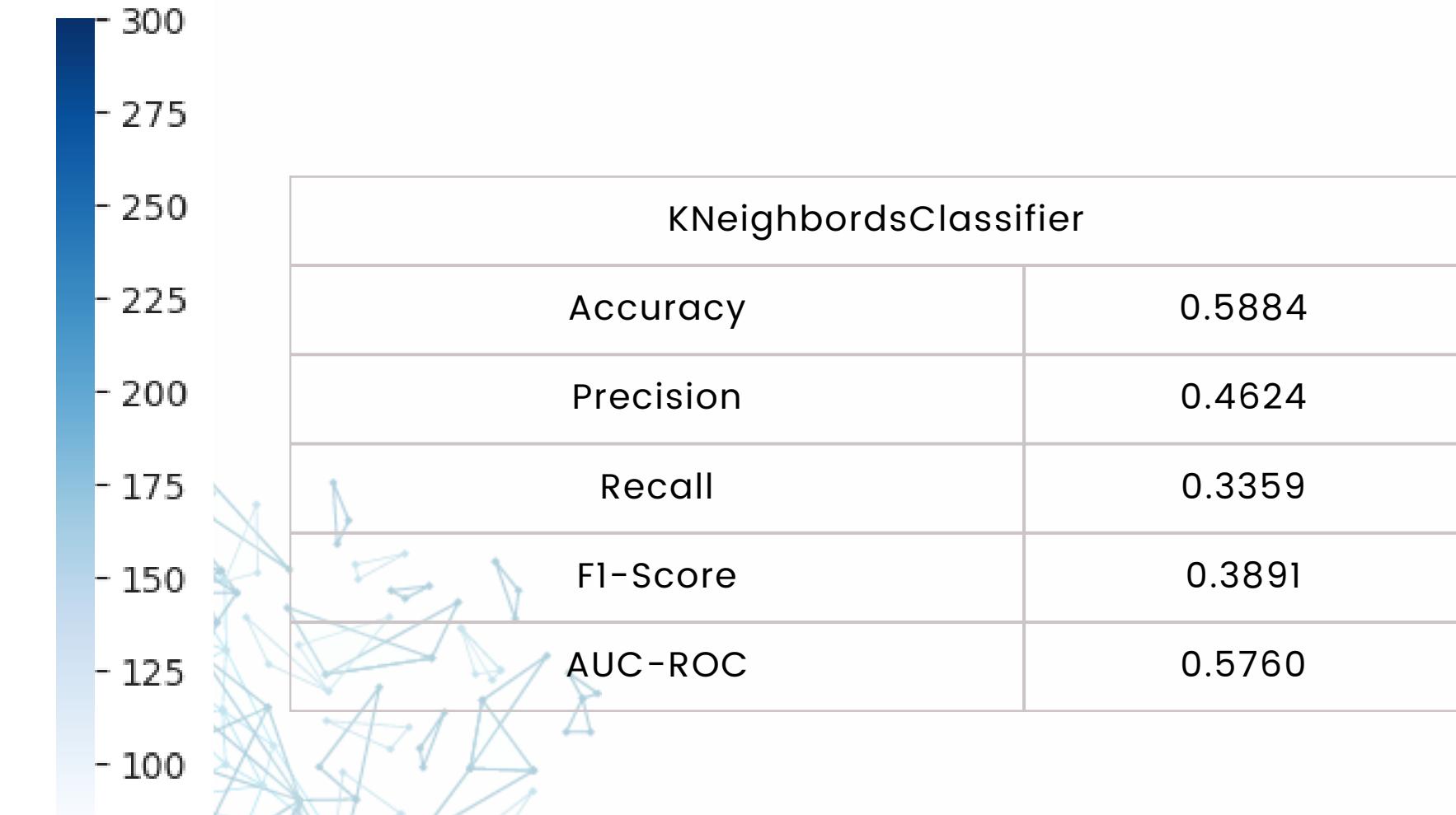
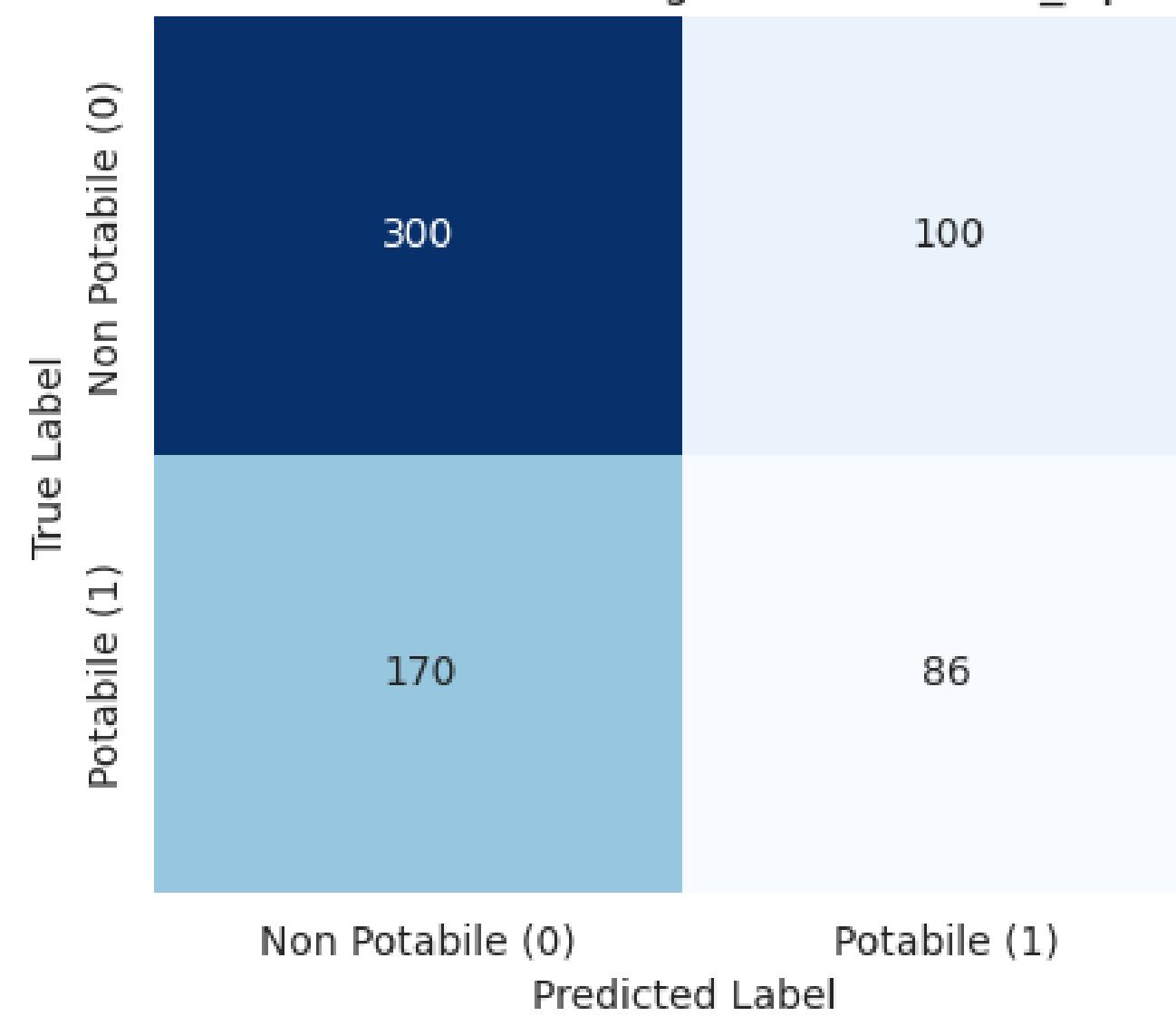
Confusion Matrix for RandomForestClassifier_Optimized



RandomForestClassifier	
Accuracy	0.6707
Precision	0.6538
Recall	0.3320
F1-Score	0.4404
AUC-ROC	0.6539

	Precision	Recall	F1-Score	Support
0	0.67	0.38	0.77	400
1	0.65	0.33	0.44	256
accuracy			0.67	656
macro avg	0.66	0.61	0.60	656
weighted avg	0.67	0.67	0.64	656

Confusion Matrix for KNeighborsClassifier_Optimized

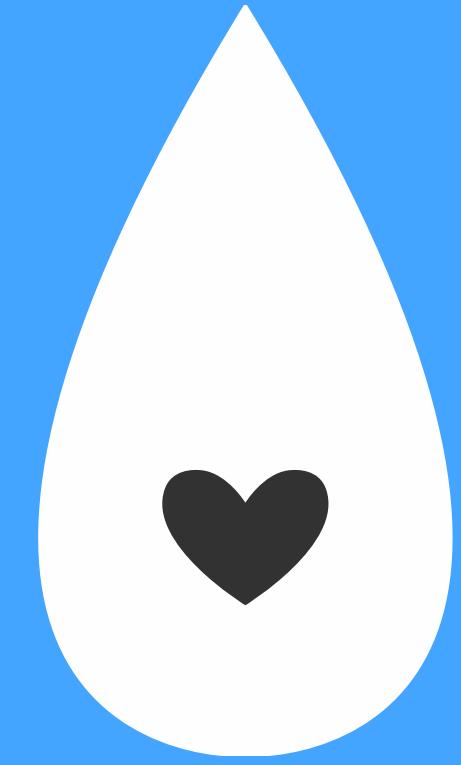


Analisi Finale e Conclusioni

Il Random Forest ottimizzato risulta quindi il modello più efficace in termini di F1-score e capacità discriminatoria generale. Tuttavia, la classificazione della potabilità dell'acqua rimane complessa a causa dell'elevata sovrapposizione tra le classi e della mancanza di un segnale forte nelle feature.

Dal punto di vista aziendale, la scelta tra precisione e recall dipende dai costi associati ai falsi positivi (rischi sanitari e legali) e ai falsi negativi (sprechi o opportunità mancate). Possibili passi futuri includono la regolazione della soglia di classificazione, l'uso di pesi di classe, tecniche di bilanciamento come SMOTE, feature engineering, sperimentazione di algoritmi come Gradient Boosting o SVM non lineari e ottimizzazione avanzata delle metriche.

Per questo progetto, il Random Forest Classifier ottimizzato rappresenta la soluzione migliore trovata, pur riconoscendo la necessità di ulteriori affinamenti per allineare le prestazioni del modello alle esigenze e ai costi aziendali.



Thankyou

FRANCO DE GIORGIO