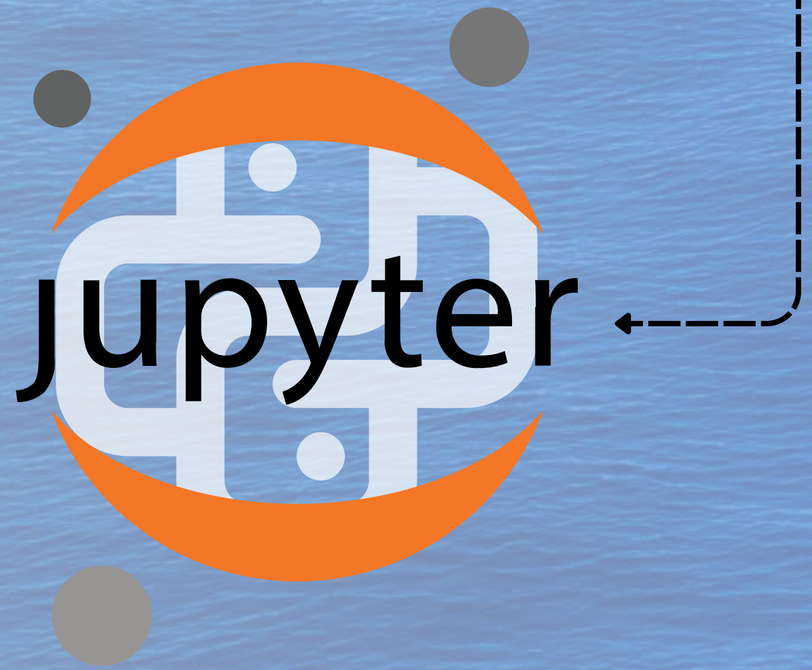
Data Science Water Quality Study

WATER POTABILITY ANALYSIS & MACHINE LEARNING PREDICTION

Federico Gori Data Scientist

Questo progetto è stato sviluppato in linguaggio Python su Jupyter Notebook, per il progetto originale cliccare sul logo



Il dataset in esame verrà utilizzato per prevedere la potabilità dell'acqua attraverso un indice binario "Potability" che classifica differenti campioni di acqua come potabili (1) o non potabili (0).

I campioni di acqua sono stati analizzati secondo 9 proprietà chimico-fisiche del liquido in relazione alla potabilità, dunque l'obiettivo di questa EDA sarà addestrare dei modelli predittivi per riconoscere l'affidabilità di un campione di acqua per il consumo.

Le 9 proprietà chimico-fisiche individuate nei campioni di acqua sono:

- **Ph**: misura che esprime l’acidità o l’alcalinità di un liquido
- **Hardness**: indica il contenuto di sali disciolti all'interno del liquido
- **Solids**: misura il totale delle sostanze organiche e non disciolte nel liquido
- **Chloramines**: disinfettanti utilizzati per il trattamento dell'acqua potabile
- **Sulfate**: sostanza presente naturalmente nell'acqua
- **Conductivity**: misura la conduttività dell'acqua in base ai sali disciolti
- **Organic Carbon**: misura il totale dei composti organici presenti nel liquido
- **Trihalomethanes**: sottoprodotti del processo di disinfezione dell'acqua
- **Turbidity**: misura la trasparenza del liquido e la capacità di assorbire o riflettere la luce

Exploratory Data Analysis

Nel dataset in esame il 43% delle righe presenta valori nulli, distribuiti tra le colonne:

- ph: 491 su 3276 casi
- Sulfate: 781 su 3276 casi
- Trihalomethanes: 162 su 3276 casi

Per evitare una perdita di dati necessari per l'analisi, date le dimensioni ridotte del dataset, si procederà a colmare le informazioni mancanti con nuovi dati, calcolati con la tecnica dell'Iterative Imputation.

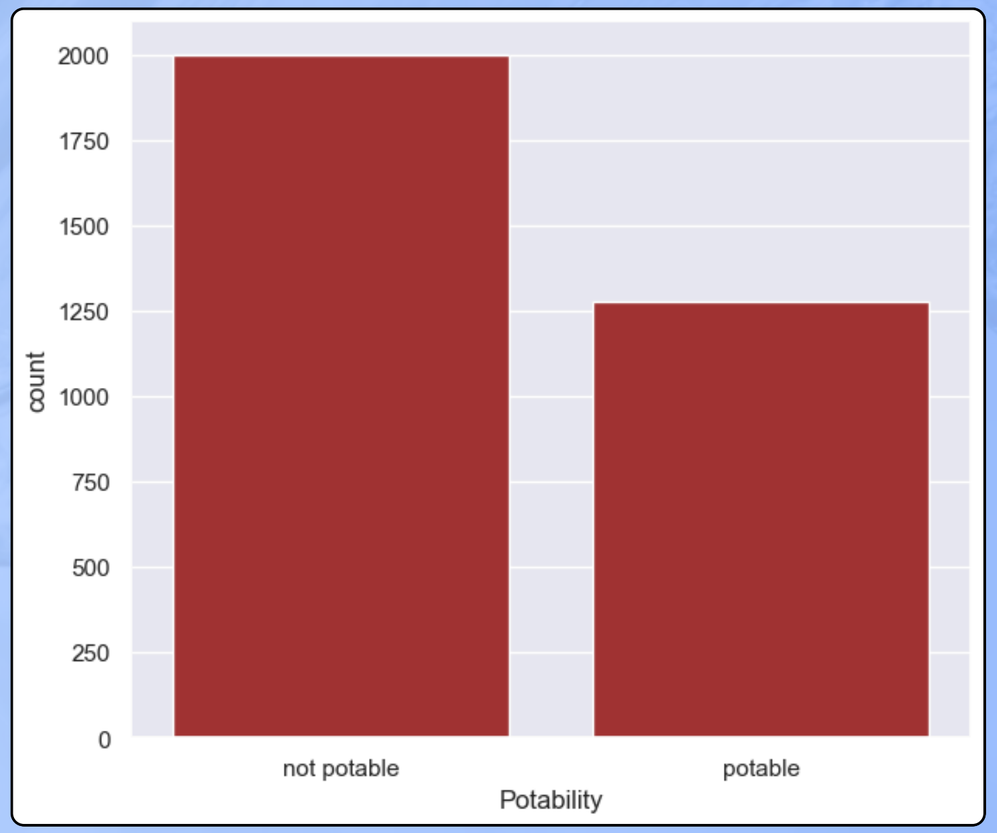
Le Features si presentano con unità di misura differenti tra loro:

- **ph : valori interi da 0 a 14**
- **Hardness, Sulfate : milligrammi per litro (mg/L)**
- **Solids, Chloramines, Organic Carbon : parti per milione (ppm)**
- **Conductivity : microsiemens per centimetro (µS/cm)**
- **Trihalomethanes : microgrammo per litro (µg/L)**
- **Turbidity : unità nefelometrica di torbidità (NTU)**

Per eguagliare l'importanza di tutte le features sulla previsione della variabile target e limitare l'influenza dei diversi valori numerici in scale differenti, le variabili saranno standardizzate mediante la funzione Standard Scaler.

La variabile target “Potability” presenta una distribuzione sbilanciata delle classi che la compongono:

- not potable (0): 1998 casi su 3276 casi (60%)
- potable (1): 1278 su 3276 casi (40%)



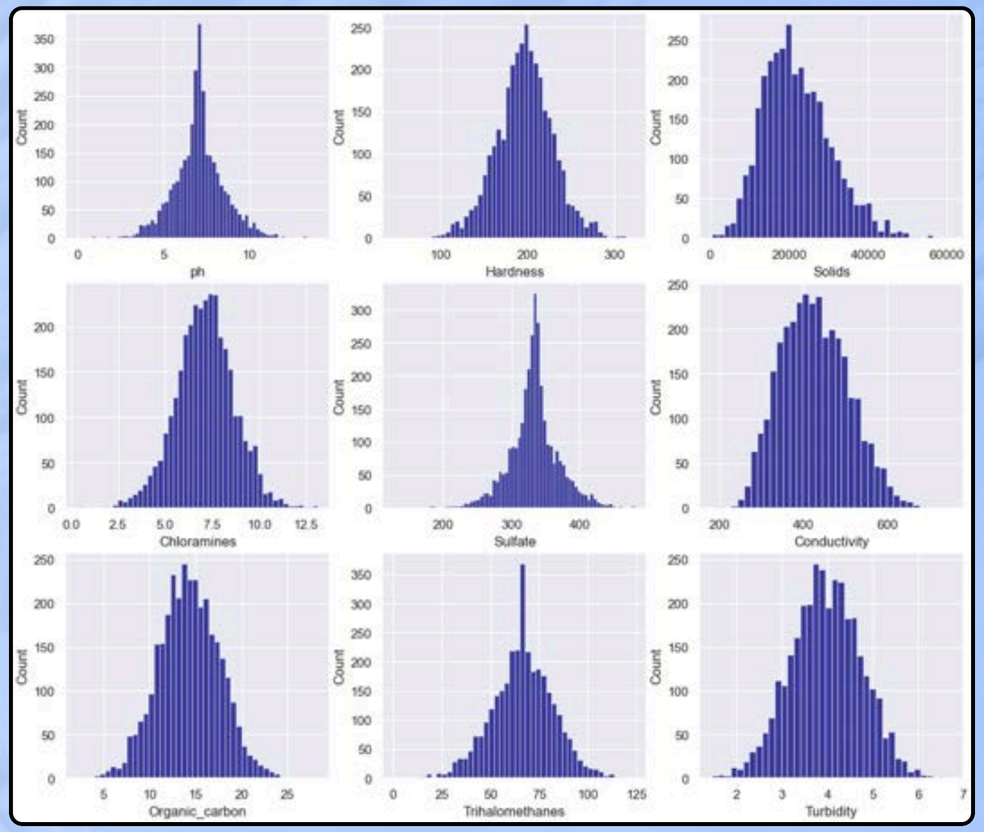
Per questo sarà necessario utilizzare metriche di valutazione dei modelli predittivi basate su Precision e Recall Score, ROC Curve e AUC, le quali danno maggiori informazioni sulle predizioni dei modelli in base alle classi predette correttamente o meno.

Inoltre sarà utile bilanciare il peso delle due classi attraverso due tecniche:

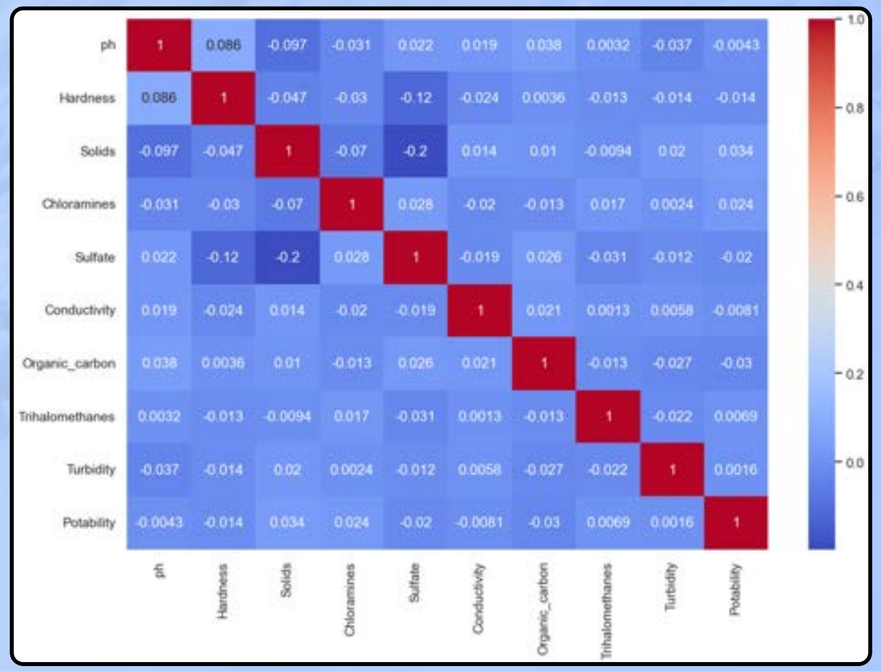
- lo Stratified Sampling nella divisione del dataset in training e test set per riequilibrare l'impatto delle due classi presenti nel target sulle predizioni dei modelli, garantendo una loro presenza omogenea nel campione utilizzato e limitando l'errore e la variabilità che potrebbe causare il Random Sampling.

- la One Sided Selection per limitare attraverso l'undersampling la grande quantità di dati appartenenti alla classe maggioritaria del target e riequilibrare le predizioni svolte dai modelli

Tutte le features presentano una distribuzione dei valori simile alla curva gaussiana, ciò suggerisce che la presenza degli outliers possa essere significativamente ridotta per la maggior parte delle stesse.

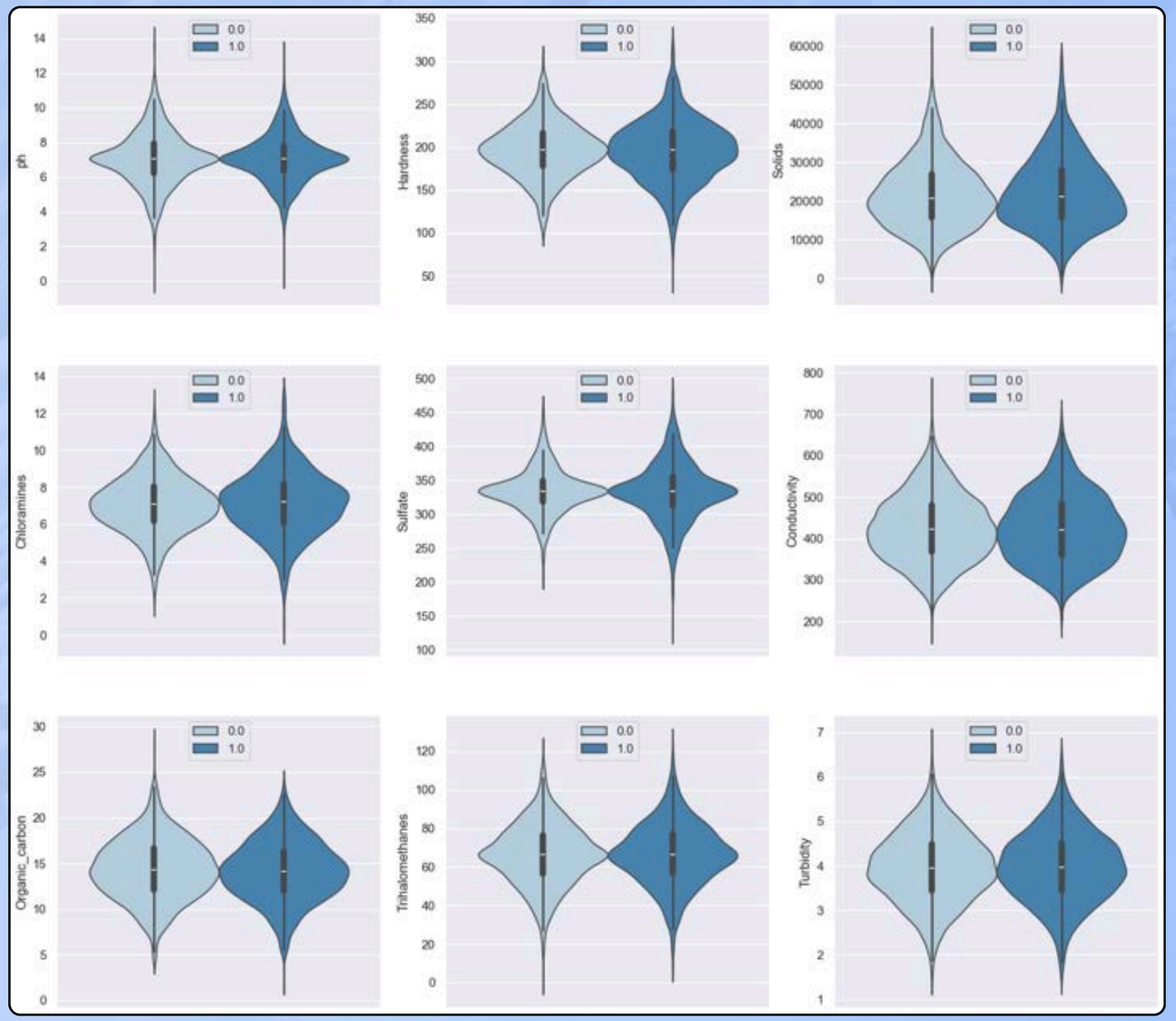


Nella verifica della correlazione tra tutte le variabili del dataset si nota la totale assenza di relazioni lineari tra tutte le features e con la variabile target; dunque, sia per la natura binaria della variabile target e sia per la mancanza di relazioni lineari tra le features, si prediligeranno modelli non lineari di Machine Learning basati sulla classificazione per la previsione del target.



La distribuzione dei valori delle features in esame suddivise rispetto all'appartenenza alle classi della variabilie target mostra un'influenza maggiore della potabilità dell'acqua in base al comportamento delle features:

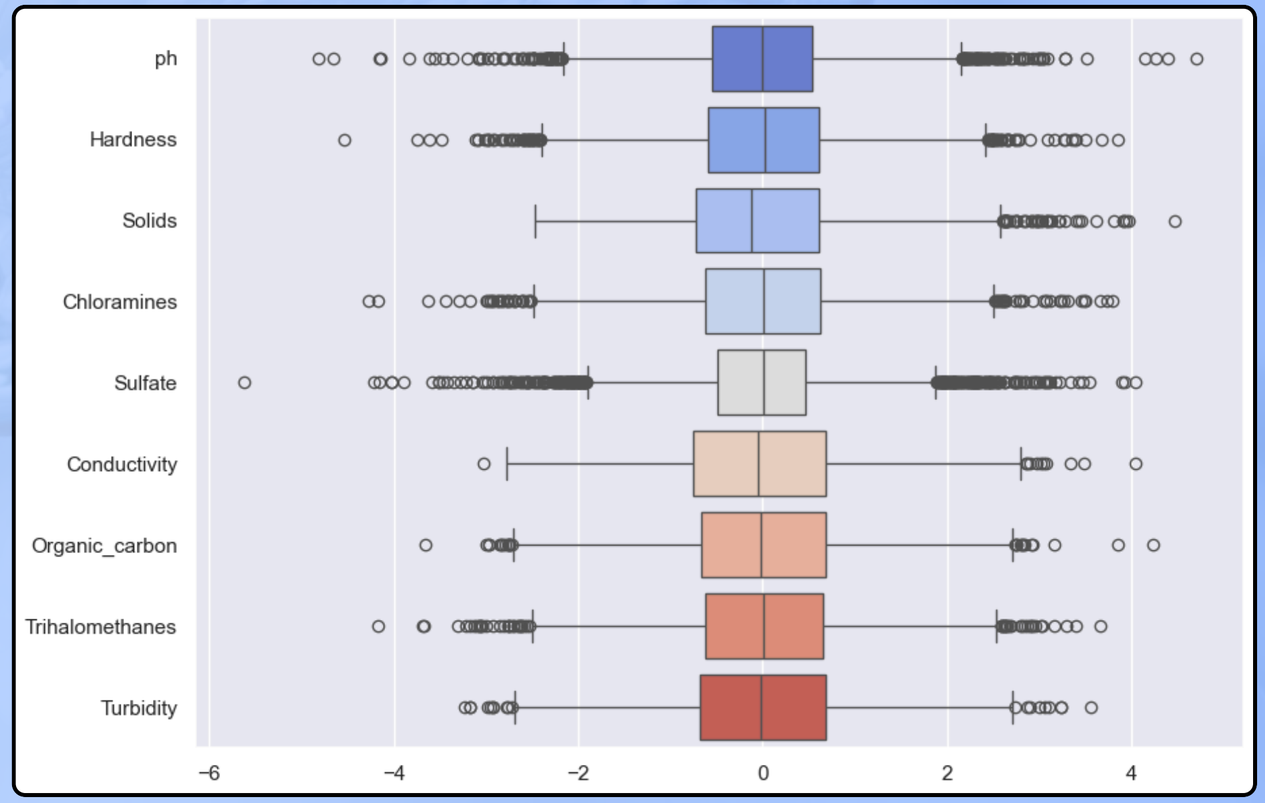
- Sulfate : differenza nelle code e nella variabilità della distribuzione
- Ph : differenza nella variabilità della distribuzione tra campioni potabili e non
- Hardness, Chloramines : lieve differenza nella variabilità della distribuzione


0.0
1.0
ph
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity

Gli outliers individuati attraverso lo Scarto Interquartile (IQR) si concentrano principalmente nelle features:

- Ph (4,33% sul totale dei casi)
- Sulfate (7,54% sul totale dei casi)

Data la natura normale delle distribuzioni e la discreta presenza di outliers, sarà utilizzato il Robust Scaler per scalare i valori estremi nelle features individuate e ridurne il rumore sugli altri valori.



Come ulteriore verifica sull'importanza delle features rispetto alle classi della variabile target, verrà svolto un Chi Square Test per constatare o rigettare l'ipotesi nulla corrispondente all'assenza di relazione tra features e target, impostando come soglia decisionale il valore P Value < 0,05.

Per questo le features del dataset sono state trasformate in variabili discrete, suddividendo i valori continui tramite degli intervalli basati sui quartili delle features, attraverso la tecnica del Binning e successivamente utilizzate per creare delle tabelle di contingenza divise per le classi del target.

I risultati del Chi Square Test sulle tabelle di contingenza prodotte dimostrano che le variabili per cui è possibile rigettare l'ipotesi nulla e quindi constatare la relazione con il target sono:

- **Hardness (P Value = 0.00 < 0.05)**
- **Chloramines (P Value = 0.00 < 0.05)**
- **Sulfate (P Value = 0.00 < 0.05)**
- **Ph (P Value = 0.02 < 0.05)**
- **Solids (P Value = 0.01 < 0.05)**

Features	Chi-Square	P-Value
Hardness binned	13.519997	0.00
Chloramines binned	15.580474	0.00
Sulfate binned	40.237992	0.00
Solids binned	11.130197	0.01
Ph binned	9.814076	0.02
Organic Carbon binned	5.194858	0.16
Conductivity binned	3.010170	0.39
Trihalomethanes binned	1.701336	0.64
Turbidity binned	0.009885	1.00

Spot-Check & Cross-Validation Pipelines

In base alle informazioni raccolte nella fase esplorativa, il dataset è stato diviso in training e test set secondo due criteri differenti di selezione delle features, formando due campioni:

- il primo comprendente tutte le features di default del dataset
- il secondo comprendente le features selezionate con il Chi Square Test.

Inoltre, per rappresentare al meglio lo squilibrio nella distribuzione delle classi della variabile target anche nei set di addestramento dei modelli predittivi, la divisione del dataset nei due set avverrà mediante lo Stratified Sampling, utilizzando l'attributo “stratify=target” della funzione train_test_split.

Per lo Spot-Check solamente i due training set creati sono stati utilizzati rispettivamente per analizzare le performance iniziali, mediante il criterio dell'Accuracy Score, di tre modelli predittivi:

- KNeighbors Classifier
- Logistic Regression
- Random Forest Classifier

Inoltre è stata attuata la Cross Validation, per verificare le performance dei modelli attraverso più partizioni differenti del dataset, tramite il metodo della Stratified K Fold, per rispettare il bilanciamento delle classi del target in ogni singola verifica.

Nello svolgere questa procedura entrambi i campioni di addestramento sono stati modificati rispetto alle statistiche riscontrate nella fase esplorativa, sfruttando una Pipeline ad hoc secondo i seguenti passaggi:

1. Imputazione dei dati mancanti con la tecnica “Iterative Imputation”
2. Scaling delle features con differenti unità di misura con “Standard Scaler”
3. Scaling delle features con una maggiore presenza di outliers con “Robust Scaler”
4. Undersampling e riequilibrazione del target con la tecnica “One Sided Selection”
5. Impostazione del Classificatore per la predizione del target

I modelli che attraverso lo Spot-Check hanno raggiunto una media di punteggi più alta per l'accuratezza predittiva con entrambi i training set sono:

- **KNeighbors Classifier** (Cross-Validation Accuracy Score medio: **0.63** / **0.64**)
- **Random Forest Classifier** (Cross-Validation Accuracy Score medio: **0.66** / **0.65**)

Models	Training Set	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean Accuracy Score
KNeighbors Classifier	All Features Set	0.64	0.61	0.58	0.66	0.62	0.63
Logistic Regression	All Features Set	0.59	0.60	0.61	0.59	0.61	0.61
Random Forest Classifier	All Features Set	0.67	0.62	0.65	0.66	0.64	0.66
KNeighbors Classifier	Chi Square Set	0.62	0.64	0.63	0.64	0.65	0.64
Logistic Regression	Chi Square Set	0.59	0.61	0.60	0.61	0.60	0.61
Random Forest Classifier	Chi Square Set	0.67	0.65	0.62	0.64	0.63	0.65

Hyperparameters Cross-Validation

Dopo aver constatato quali sono i due modelli migliori, ne è stata eseguita la Cross-Validation degli iperparametri tramite le funzioni Randomized Search CV e Grid Search CV:

- Per il modello KNeighbors Classifier è stata utilizzata la Grid Search CV, data la minore mole di iperparametri da validare

- Per il modello Random Forest Classifier è stata utilizzata la Randomized Search CV, data la maggiore complessità computazionale del modello ed il maggior numero di iperparametri da validare

Nella Cross-Validation degli iperparametri sono stati utilizzati altre due metriche di verifica delle prestazioni dei modelli, oltre all'Accuracy Score, includendo anche la valutazione della capacità dei modelli di distinguere le classi del target in modo ottimale:

Precision Score

(possibilità espressa da 0 a 1 per cui il modello è in grado di rilevare i veri negativi)

Recall Score

(possibilità espressa da 0 a 1 per cui il modello è in grado di rilevare i veri positivi)

KNeighbors Classifier

Per svolgere la Cross-Validation degli iperparametri del modello KNeighbors Classifier è stato necessario innanzitutto individuare il valore mediano dei neighbors da impostare per la validazione, calcolando la radice quadrata del numero di casi del dataset in esame e utilizzandolo come punto di partenza per determinare gli altri valori da validare.

Per fare ciò è state creata una funzione che trasformi i due training set scelti secondo le modifiche apportate dalla pipeline implementata sino ad ora, estraendo il valore ottimale di n_neighbors per entrambi, il quale verrà poi utilizzato per costruire un array di valori da testare per l'iperparametro di KNeighbors Classifier.

Model	Features Set	Best Params	Best Accuracy Score	Precision CV Mean	Recall CV Mean
KNeighbors Classifier	All Features Set	n_neighbors = 58	0.66	0.70	0.25
KNeighbors Classifier	Chi Square Set	n_neighbors = 44	0.67	0.62	0.38

Random Forest Classifier

L'impostazione della Cross Validation degli iperparametri da testare per il modello Random Forest Classifier avverrà con i seguenti criteri:

- **n_estimators: range di valori moderato, per le dimensioni ridotte del dataset e sia per il ridotto miglioramento che l'aumento degli alberi decisionali comporta**
- **min samples_split: range di valori contenuto, in quanto il rumore del dataset è stato già trattato con lo scaling e l'undersampling**
- **min_samples_leaf: range di valori contenuto, in quanto lo sbilanciamento delle classi del target è stato già trattato mediante l'undersampling**
- **max_features: test sulla casualità di scelta delle features di entrambi i metodi del modello**
- **max_depth: limitazione degli alberi in un range con dimensioni più contenute per aumentare il controllo sull'overfitting**

Model	Features Set	Best Params	Best Accuracy Score	Precision CV Mean	Recall CV Mean
Random Forest Classifier	All Features Set	• n_estimators = 200 • min_samples_split = 5 • min_samples_leaf = 1 • max_features = 'log2' • max_depth = 10	0.67	0.67	0.31
Random Forest Classifier	Chi Square Set	• n_estimators = 200 • min_samples_split = 5 • min_samples_leaf = 1 • max_features = 'log2' • max_depth = 10	0.67	0.62	0.40

Test Set Prediction Performance Evaluation

Dopo aver individuato i valori migliori degli iperparametri appartenenti ai modelli selezionati attraverso la fase dello Spot-Check, le informazioni raccolte verranno utilizzate per ottimizzare i modelli nella predizione del target del test set in associazione con entrambi i training set scelti.

In questo caso la pipeline già utilizzata per le procedure precedenti è stata modificata per analizzare l'Accuracy Score sia sul test set che sul training set, con lo scopo di rilevare il livello di Overfitting raggiunto e poter eventualmente intervenire sugli iperparametri dei modelli per la sua riduzione.

In più i modelli saranno valutati in base alle metriche individuate nella Data Exploration, aggiungendo rispetto a quelle già utilizzate fino ad ora anche:

Confusion Matrix

(per visualizzare al meglio la capacità dei modelli di distinguere tra le due classi del target)

ROC Curve

(ulteriore indicatore delle prestazioni dei modelli, specifico per i classificatori binari)

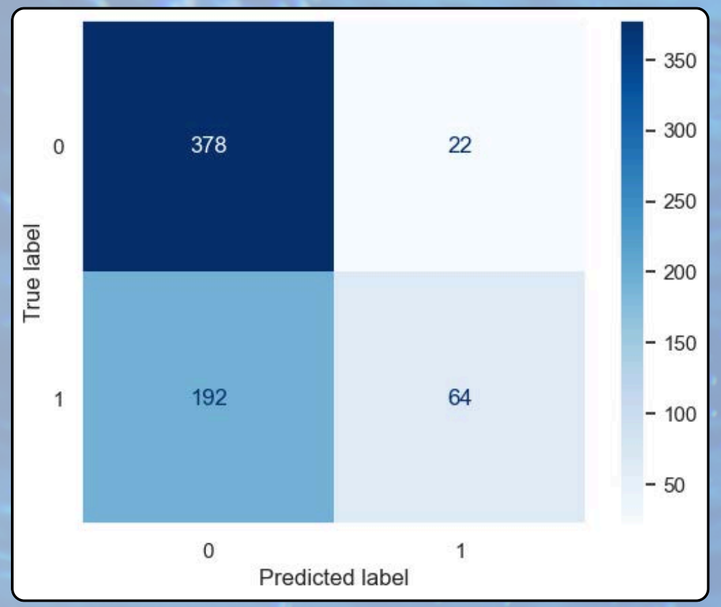
AUC

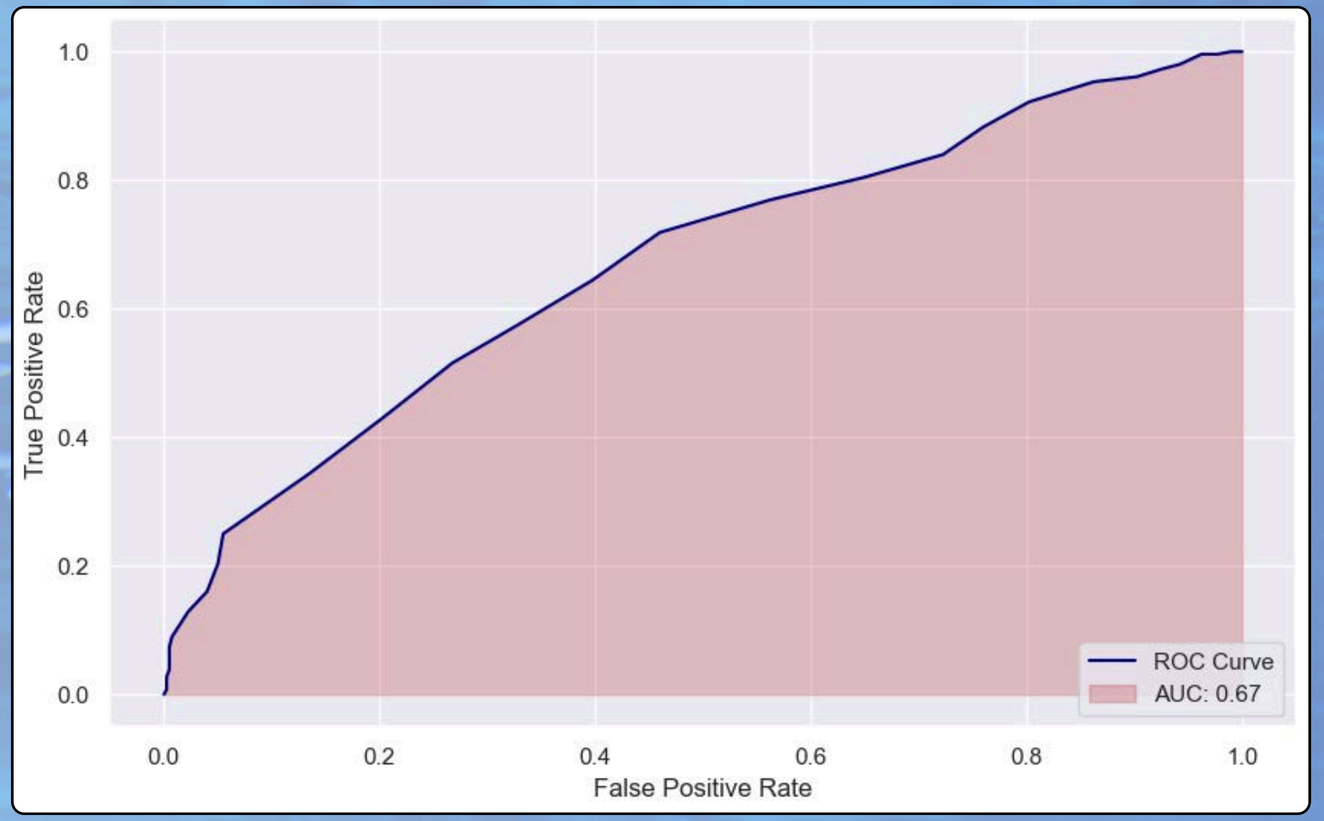
(ulteriore metrica di valutazione della capacità di distinzione tra le due classi del target)

All Features training Set - KNeighbors

Il modello KNeighbors Classifier addestrato con tutte le features originali del dataset dimostra una buona capacità predittiva attraverso il punteggio dell'Accuracy Score ed un buon equilibrio con il livello di Overfitting, ma una scarsa capacità di rappresentare le classi della variabile target, soprattutto quella minoritaria (1 = campioni di acqua potabile).

Model	Accuracy Score	Overfitting	Precision Score classe 0	Recall Score classe 0	Precision Score classe 1	Recall Score classe 1	AUC
KNeighbors Classifier	0.67	0.68	0.66	0.94	0.74	0.25	0.67

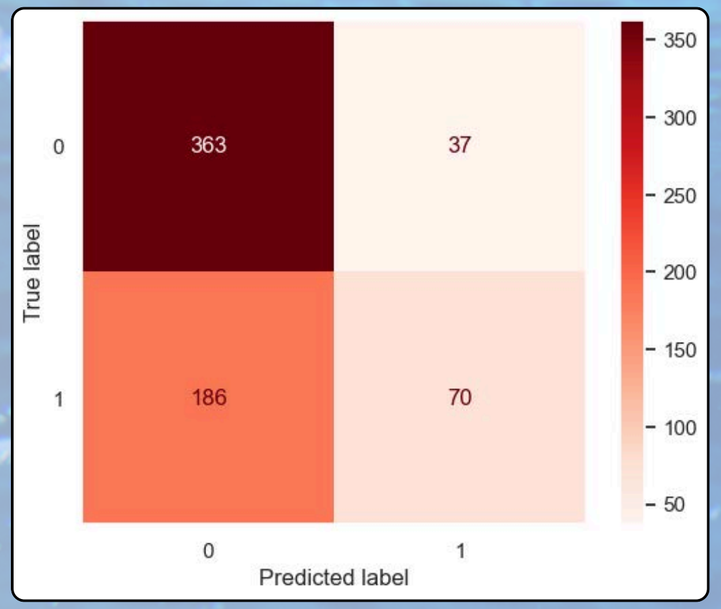

378
22
192
64
0
1
True label
Predicted label
350
300
250
200
150
100
50

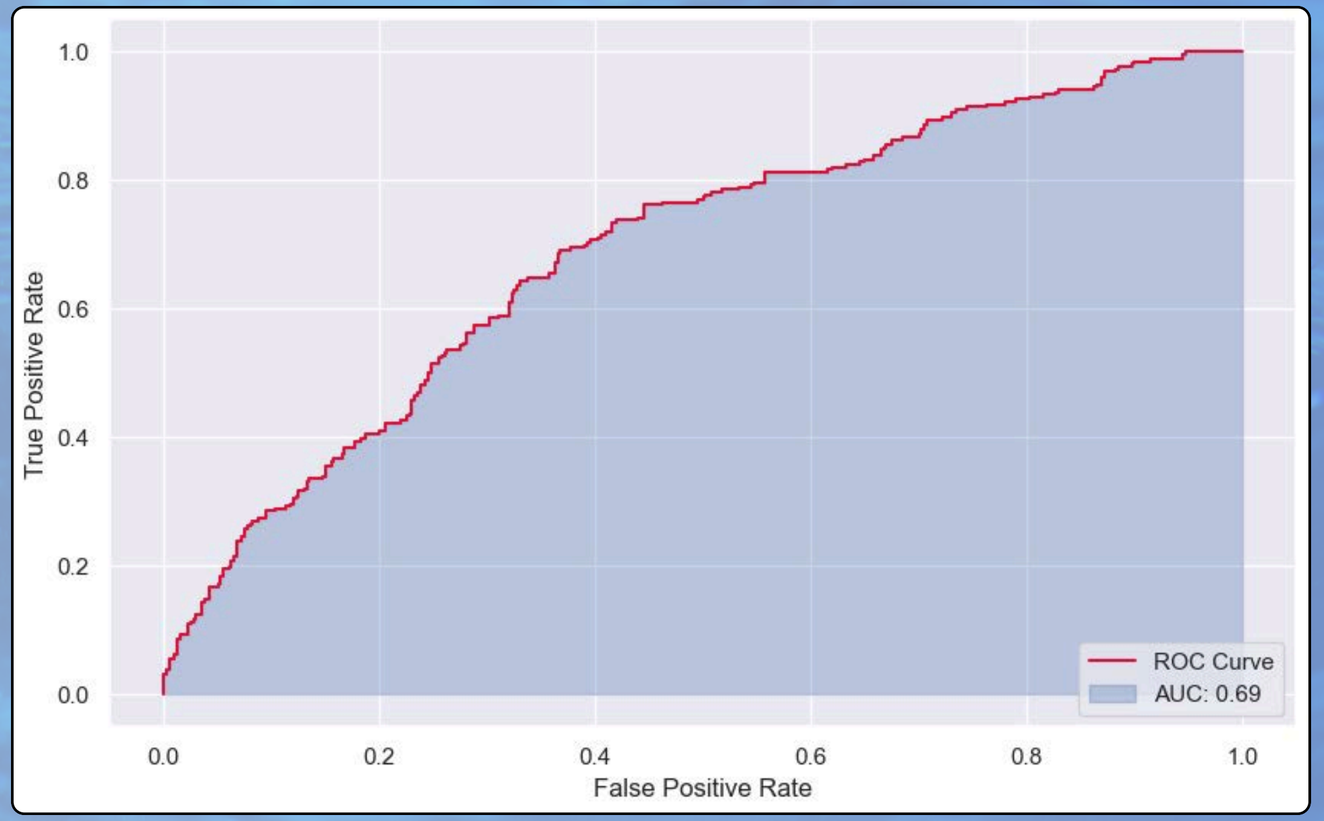

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve
AUC: 0.67

All Features training Set - Random Forest

Il modello Random Forest Classifier addestrato con tutte le features originali del dataset dimostra un elevato livello di Overfitting, proprio dei modelli Ensemble ed una capacità di distinguere le due classi del target molto sbilanciata, allo stesso modo del modello KNeighbors Classifier.

Model	Accuracy Score	Overfitting	Precision Score classe 0	Recall Score classe 0	Precision Score classe 1	Recall Score classe 1	AUC
Random Forest Classifier	0.66	0.83	0.66	0.91	0.65	0.27	0.69

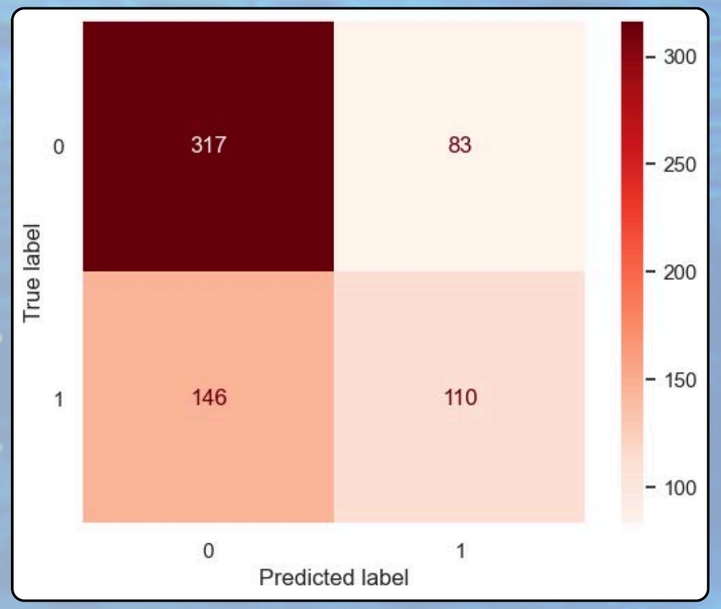

363
37
186
70
0
1
True label
0
1
Predicted label
350
300
250
200
150
100
50

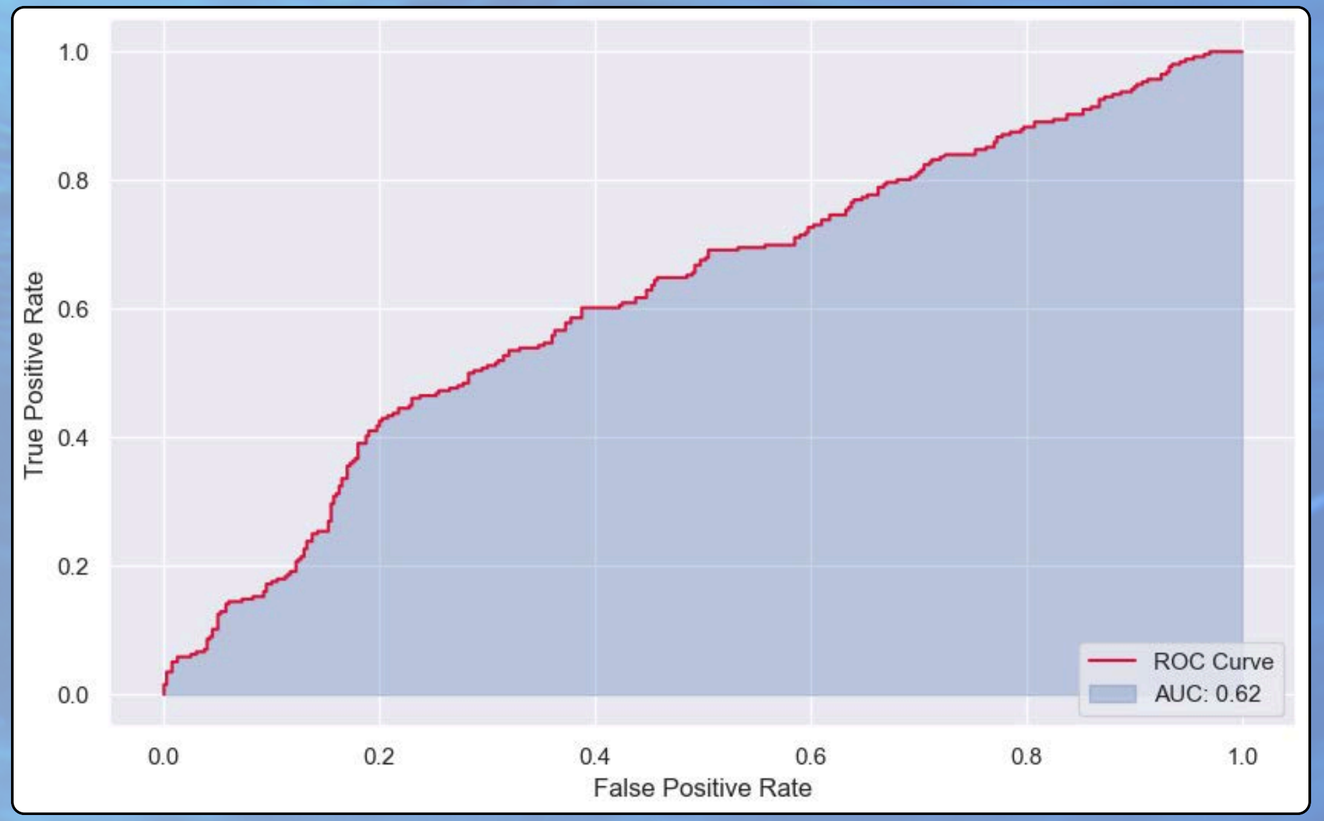

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve
AUC: 0.69

Data la possibilità del modello in questione di adattarsi meglio alla differenza numerica dei campioni di entrambe le classi, attraverso la clausola interna “class_weight = 'balanced'“ e la possibilità di ridurre l'overfitting diminuendo la complessità degli alberi con la riduzione della loro profondità (max_depth), è stato prodotto un ulteriore test per ottimizzare i risultati ottenuti.

Attuando questi cambiamenti nel modello si nota un netto miglioramento sia nel livello di Overfitting e sia nella rappresentazione di entrambe le classi del target.

Model	Accuracy Score	Overfitting	Precision Score classe 0	Recall Score classe 0	Precision Score classe 1	Recall Score classe 1	AUC
Random Forest Classifier	0.65	0.63	0.68	0.79	0.57	0.43	0.62


317
83
146
110
0
1
True label
0
1
Predicted label
300
250
200
150
100

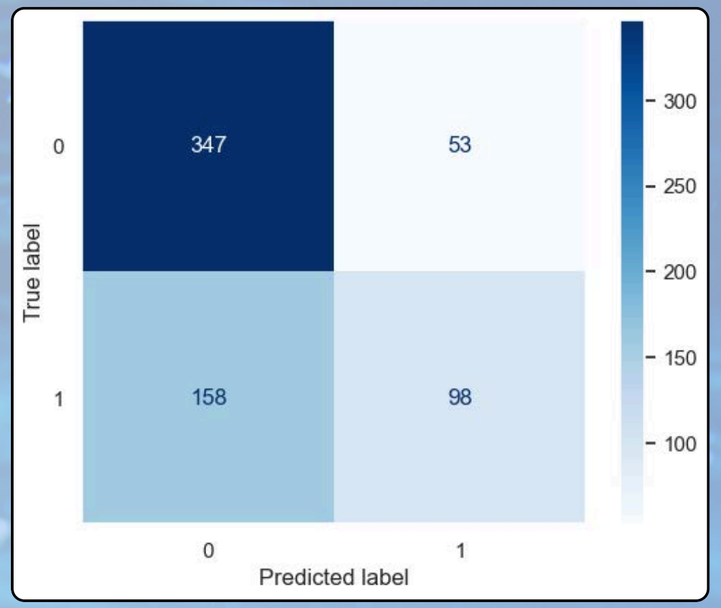

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve
AUC: 0.62

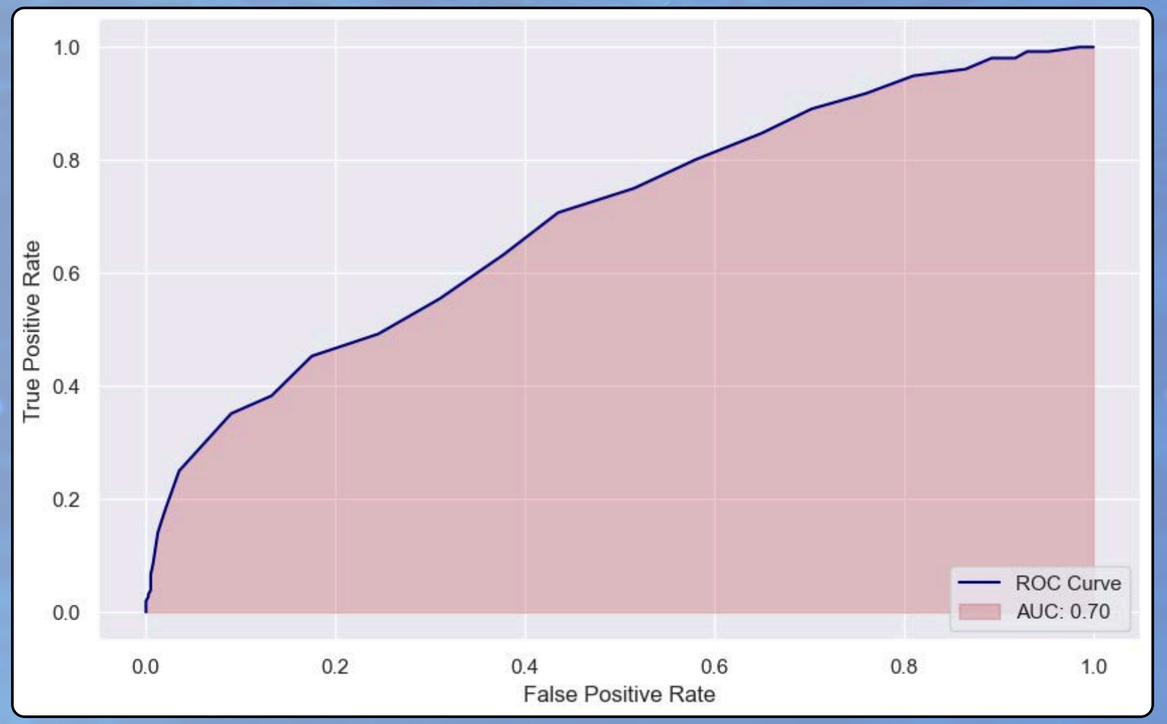
Chi Square Test Features Training Set- KNeighbors

Il modello KNeighbors Classifier addestrato con le features selezionate attraverso Chi Square Test dimostra una buona capacità predittiva attraverso il punteggio dell'Accuracy Score ed un buon equilibrio con il livello di Overfitting.

Model	Accuracy Score	Overfitting	Precision Score classe 0	Recall Score classe 0	Precision Score classe 1	Recall Score classe 1	AUC
KNeighbors Classifier	0.68	0.69	0.69	0.87	0.65	0.38	0.70

Rispetto alla sua versione addestrata con tutte le features originali del dataset si nota un miglioramento sia nella capacità predittiva con un Accuracy Score più alto, sia una migliore rappresentazione della distinzione tra le classi del dataset, questa volta più equilibrata rispetto alla classe minoritaria dei campioni potabili, dimostrata attraverso il valore dell'AUC e i punteggi del Precision e del Recall Score.



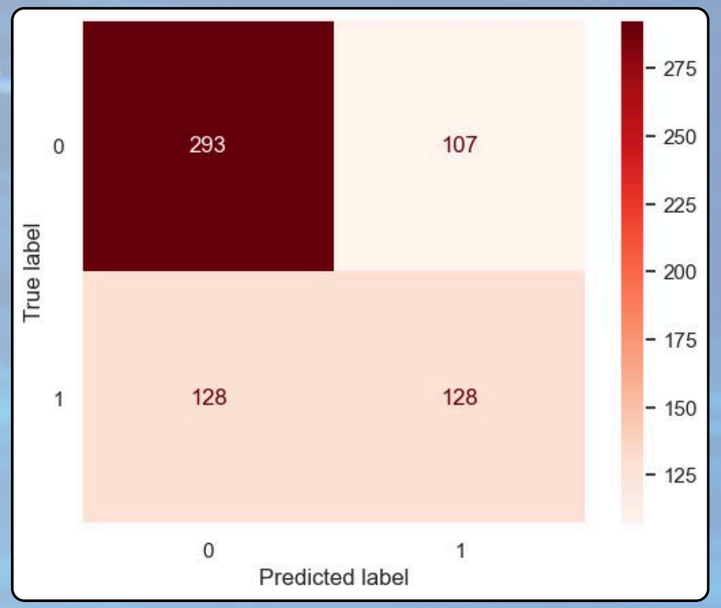


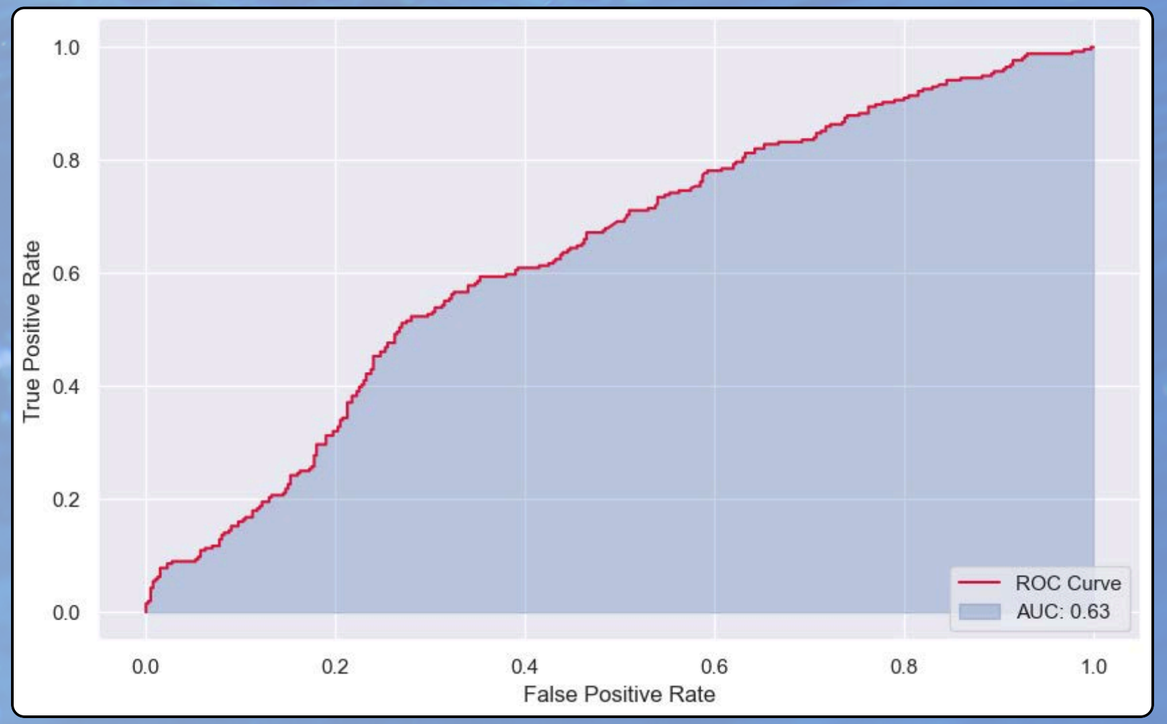
Chi Square Test Features Training Set- Random Forest

Il modello Random Forest Classifier addestrato solamente con le features selezionate attraverso Chi Square Test dimostra una minore capacità predittiva rispetto alla sua versione addestrata con tutte le features originali del dataset, con un punteggio più basso ottenuto nell'Accuracy Score.

Model	Accuracy Score	Overfitting	Precision Score classe 0	Recall Score classe 0	Precision Score classe 1	Recall Score classe 1	AUC
Random Forest Classifier	0.64	0.62	070	0.73	0.54	0.50	0.63

Rispetto alla predizione della distinzione tra le due classi il modello dimostra di essere più performante, sia dal punto di vista dell'equilibrio presente tra i risultati di precision e recall score, sia attraverso il risultato ottenuto per l'AUC.





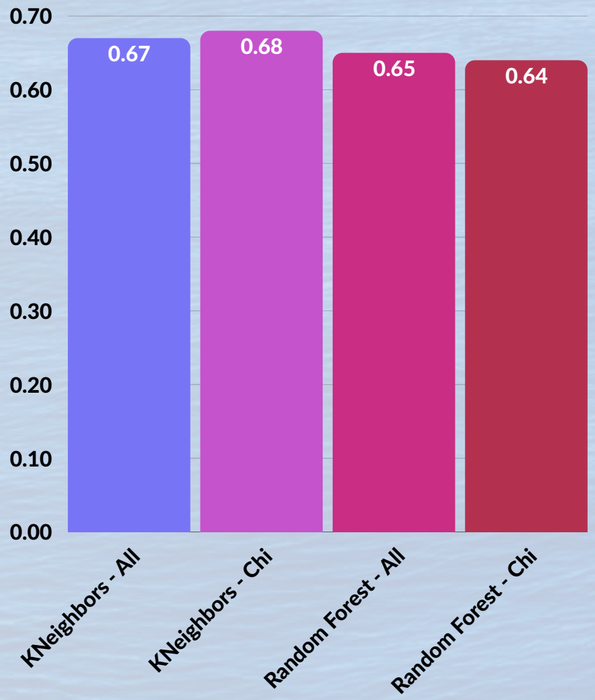
Model	Features Set	Accuracy Score	Overfitting	Precision Score classe 0	Recall Score classe 0	Precision Score classe 1	Recall Score classe 1	AUC
KNeighbors Classifier	All Features	0.67	0.68	0.66	0.94	0.74	0.25	0.67
KNeighbors Classifier	Chi Square Test	0.68	0.69	0.69	0.87	0.65	0.38	0.70
Random Forest Classifier	All Features	0.65	0.63	0.68	0.79	0.57	0.43	0.62
Random Forest Classifier	Chi Square Test	0.64	0.62	0.70	0.73	0.54	0.50	0.63

Risultati e Conclusioni

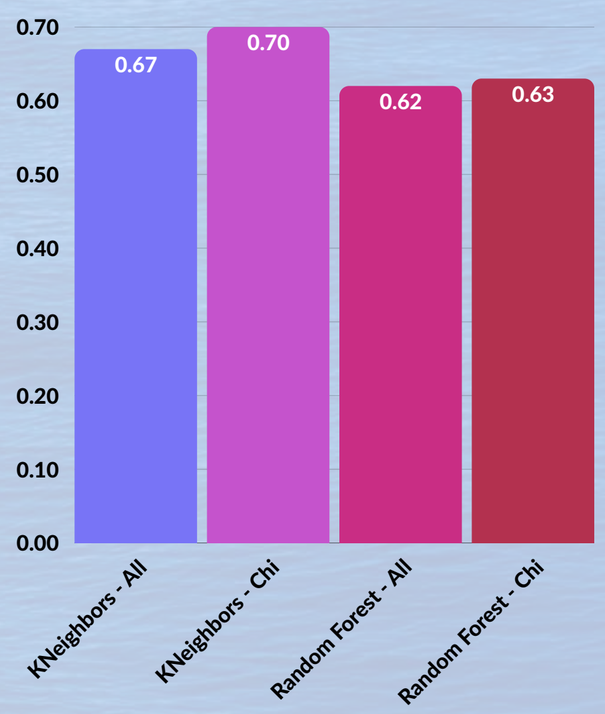
I risultati ottenuti attraverso l'Hyperparameters Cross-Validation e l'utilizzo di due set di features differenti per allenare i modelli scelti hanno dimostrato come l'indicatore di potabilità dell'acqua possa essere analizzato e predetto in maniera più efficace dal KNeighbors Classifier, nello specifico quando viene allenato con il set di features selezionate tramite il Chi Square Test.

In primo luogo per la sua maggior capacità di riportare previsioni più attendibili, dimostrata attraverso il punteggio più elevato ottenuto per l'Accuracy Score tra tutti i modelli utilizzati; infine per l'equilibrio che riesce a mantenere nell'identificazione dei campioni potabili e non potabili e nella loro distinzione tra veri e falsi positivi, soprattutto nel caso della classe minoritaria nel dataset, rappresentata dai campioni di acqua potabile.

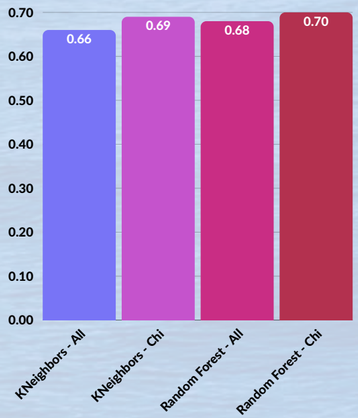
Accuracy Scores



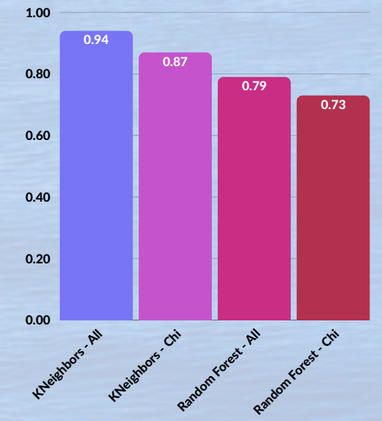
AUC Scores



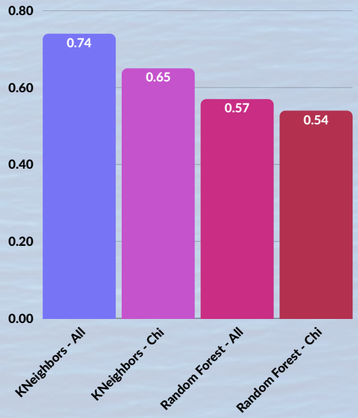
Precision Scores (Classe 0)



Recall Scores (Classe 0)



Precision Scores (Classe 1)



Recall Scores (Classe 1)