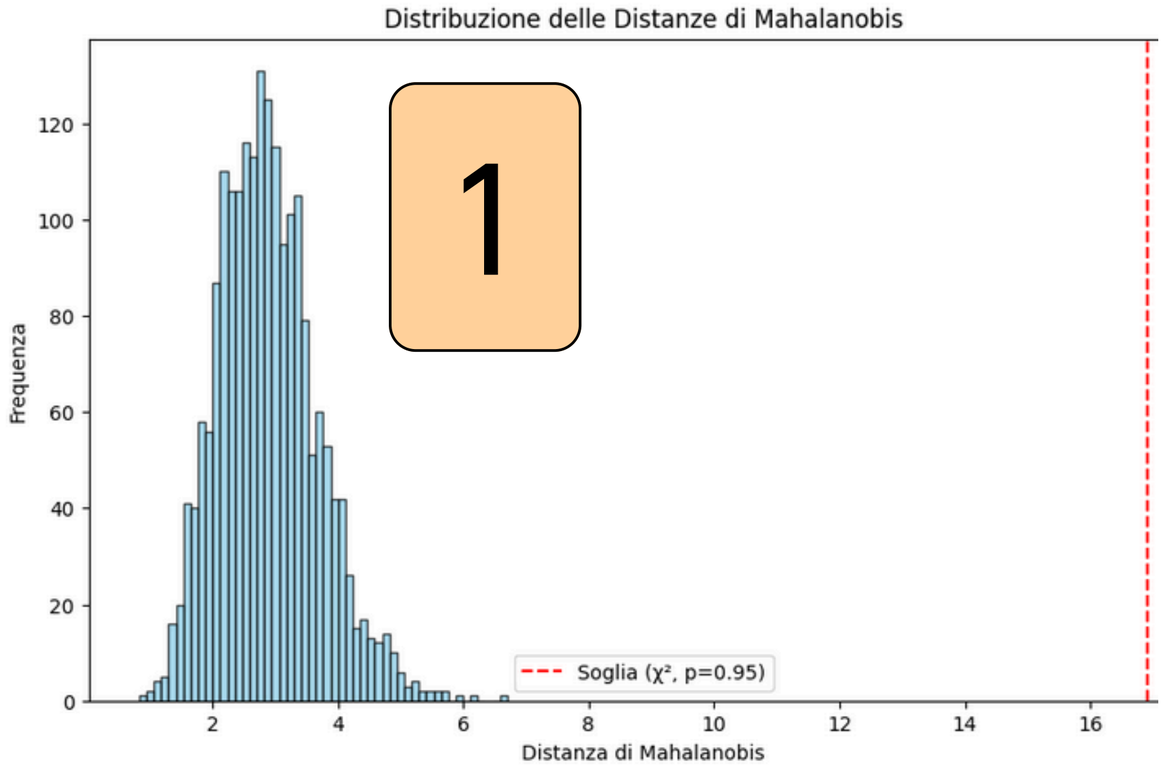
Marcello Massimi

PROGETTO FINALE DATA SCIENCE

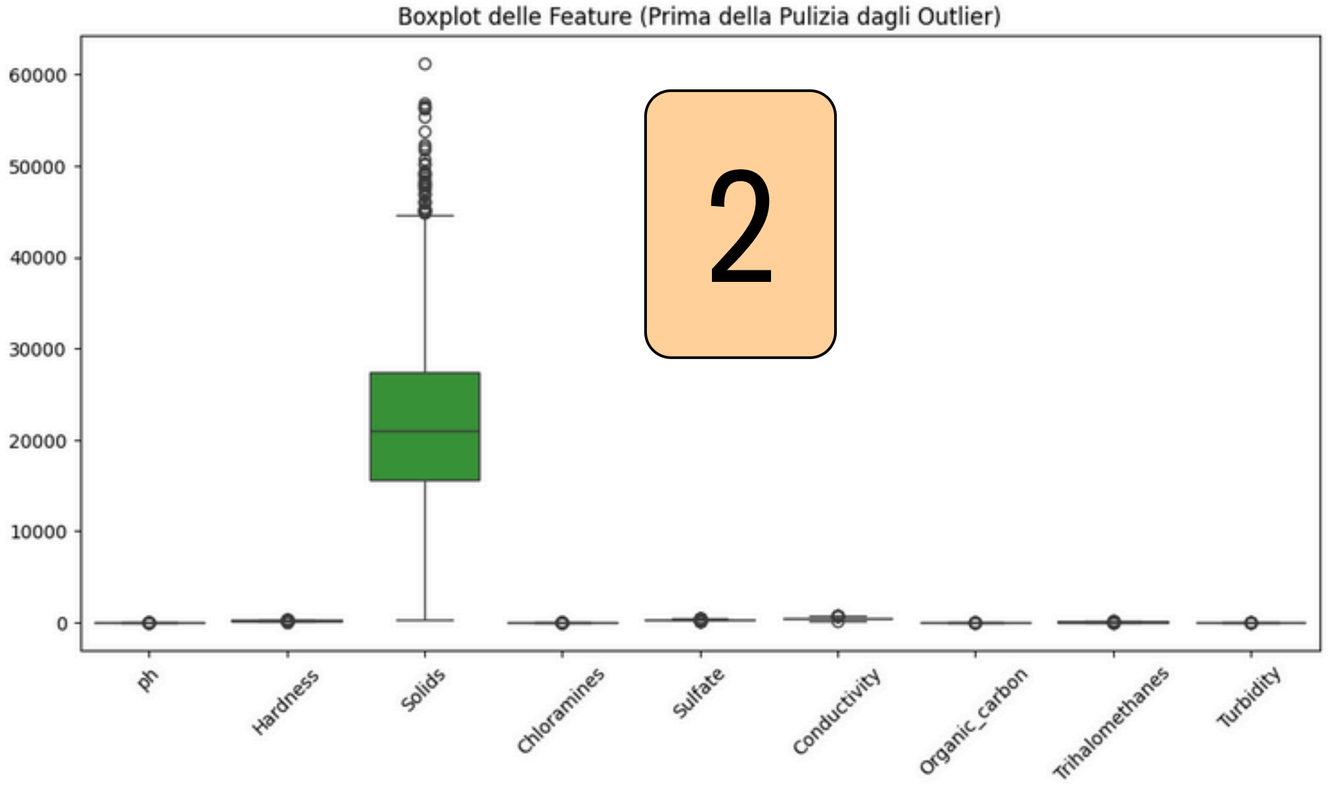
DATASET E OBIETTIVI

Utilizzeremo il dataset water_potability dataset, che è un dataset per la classificazione binaria. Il nostro obiettivo è assicurarci che il modello sia completo e comprenda tutti i controlli necessari per una valutazione corretta.

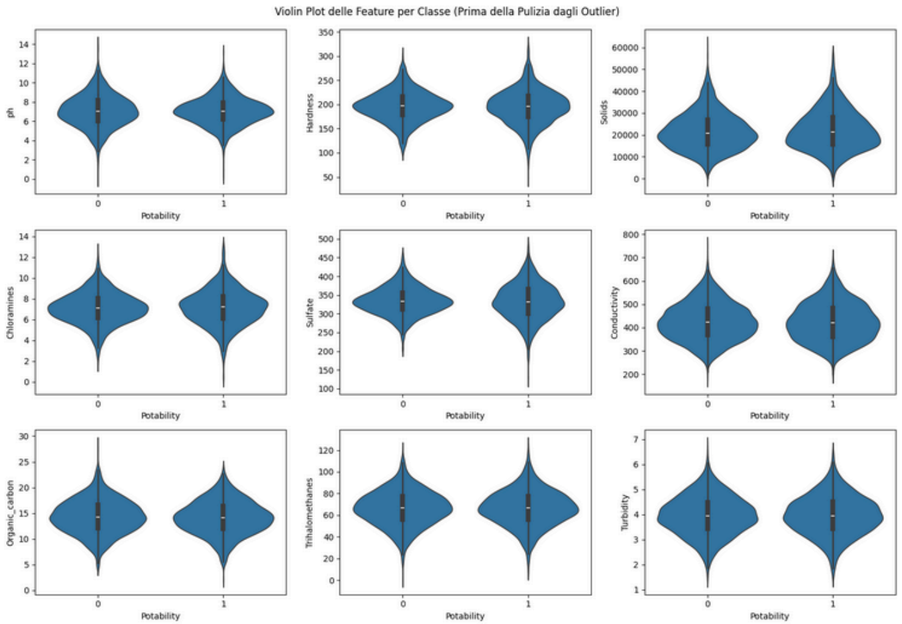
PRESENZA DI OUTLIER



1



2

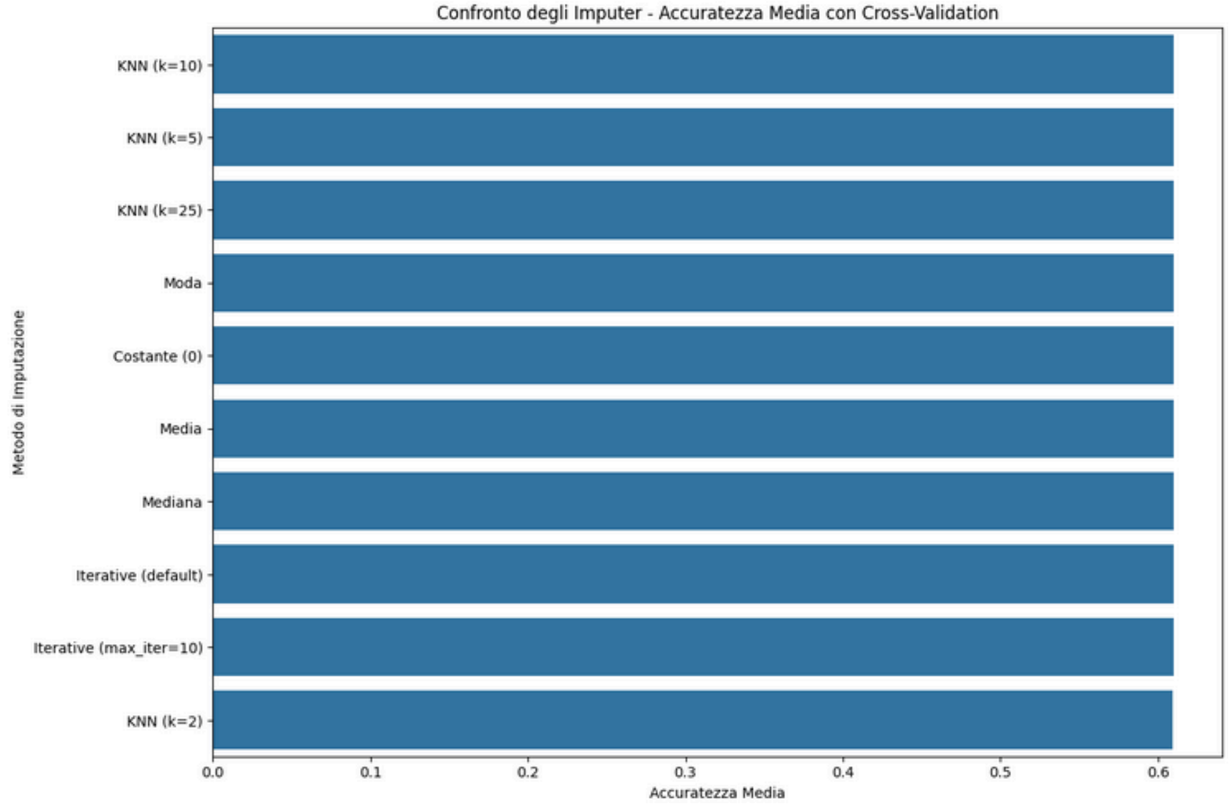


3

BEST IMPUTER

```
# Modello di riferimento
model = LogisticRegression(max_iter=5000)

# Lista imputer abbastanza completo
imputers = {
"Media": SimpleImputer(strategy="mean"),
"Mediana": SimpleImputer(strategy="median"),
"Moda": SimpleImputer(strategy="most_frequent"),
"Costante (0)": SimpleImputer(strategy="constant", fill_value=0),
"KNN (k=2)": KNNImputer(n_neighbors=2),
"KNN (k=5)": KNNImputer(n_neighbors=5),
"KNN (k=10)": KNNImputer(n_neighbors=10),
"KNN (k=25)": KNNImputer(n_neighbors=25),
"Iterative (default)": IterativeImputer(random_state=0),
"Iterative (max_iter=10)": IterativeImputer(max_iter=10, random_state=0)
}
```



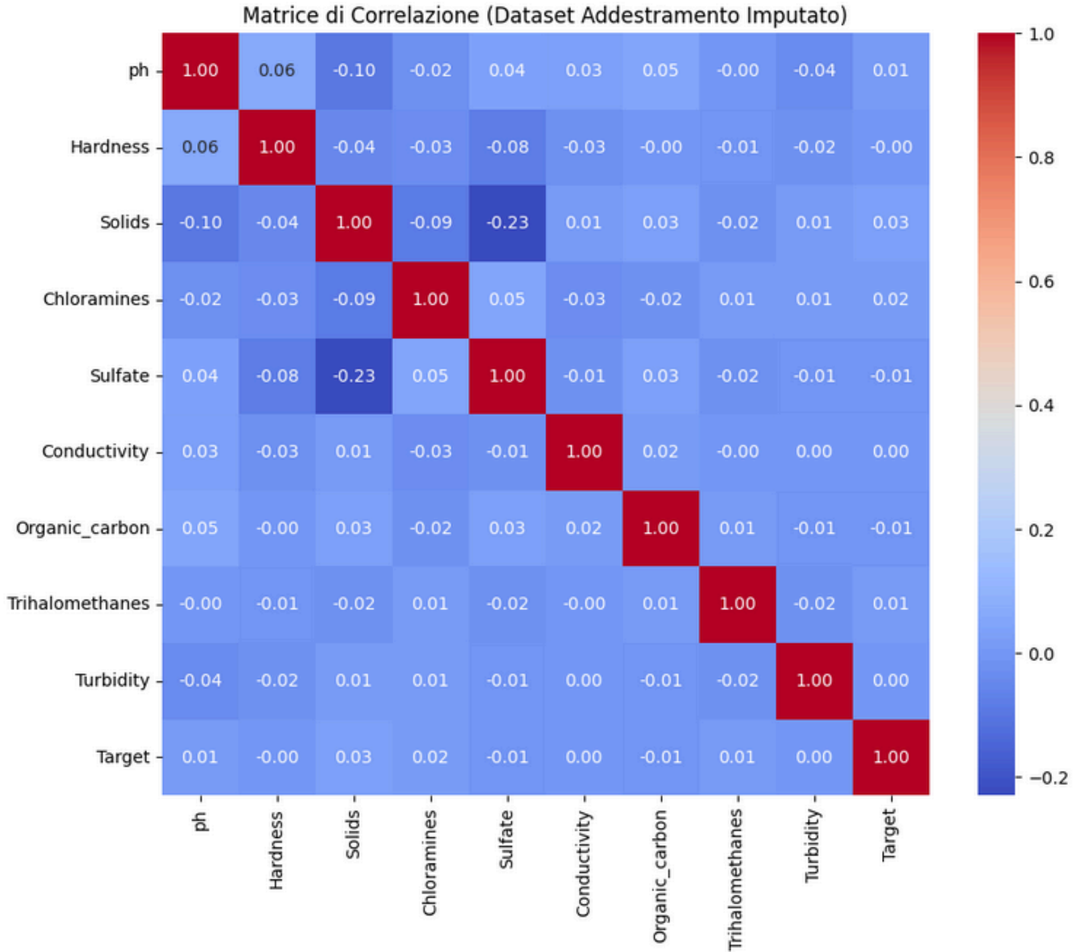
Pulizia e gestione valori nulli

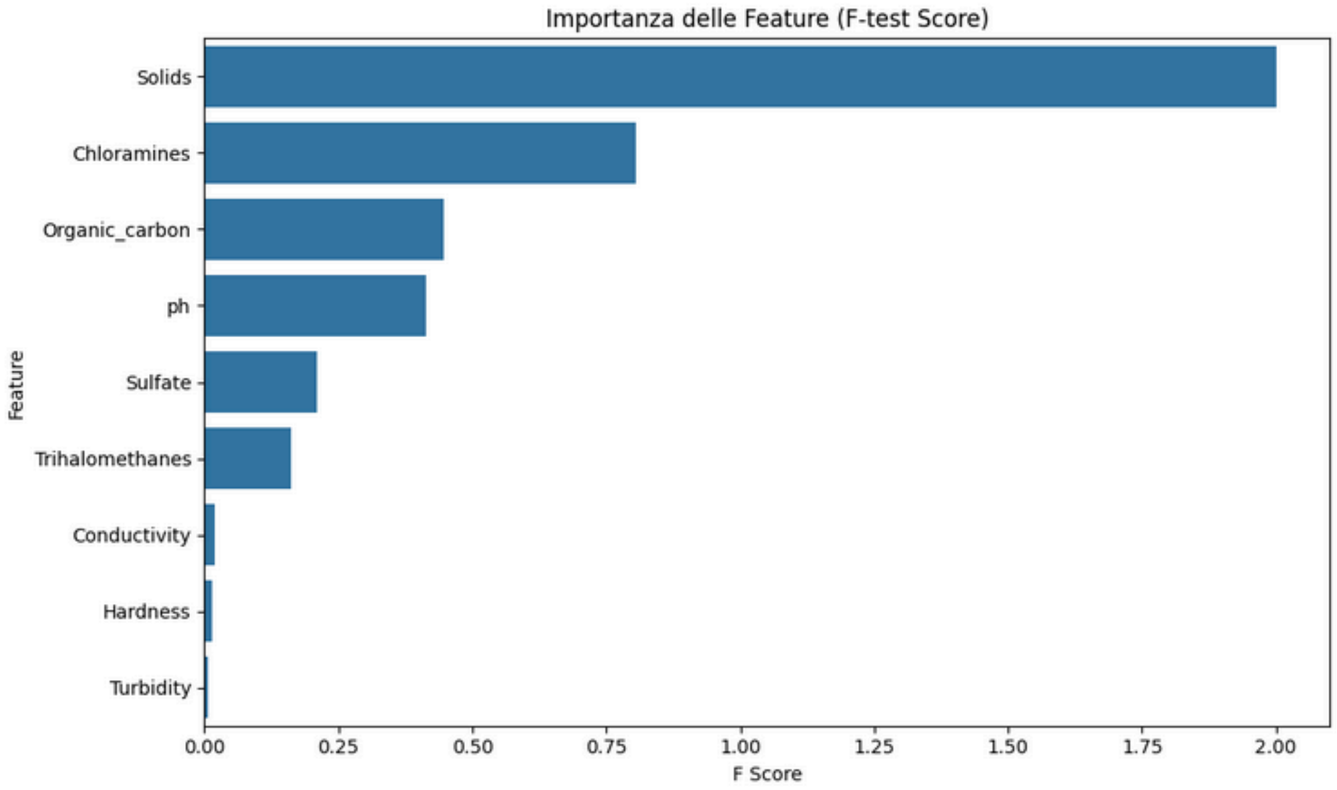
DData la moderata quantità di dati nulli definiamo una lista di imputer per trovare il migliore.

CORRELAZIONE E RELAZIONE DELLE FEATURES

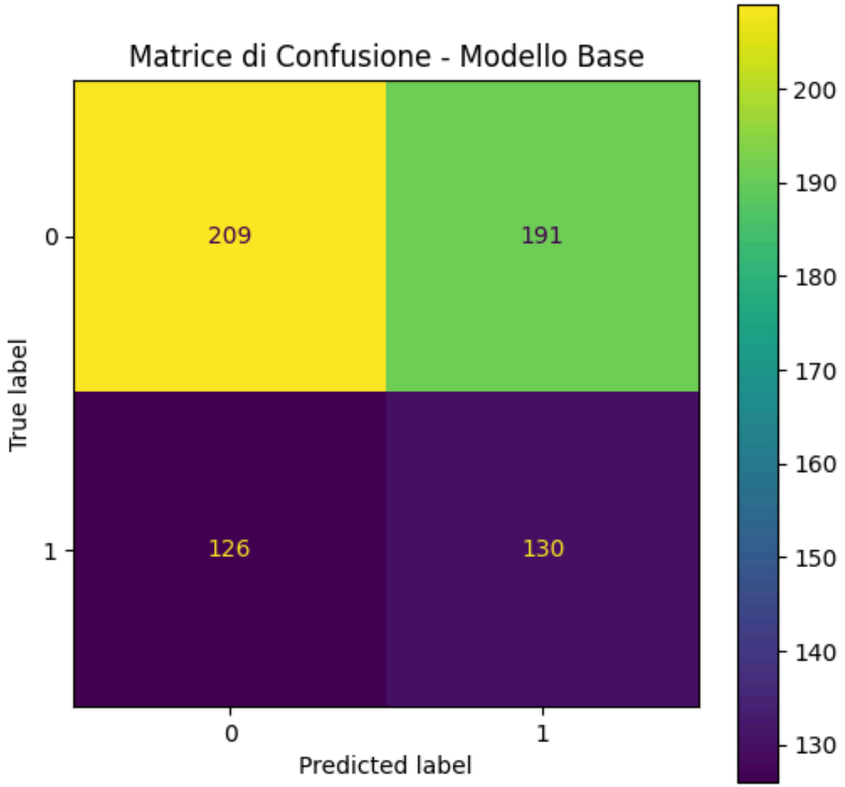
Nessuna relazione lineare presente

Relazione statistiche presente, ma il p-value ne smentisce l'importanza poiché nessuno arriva alla soglia significativa.

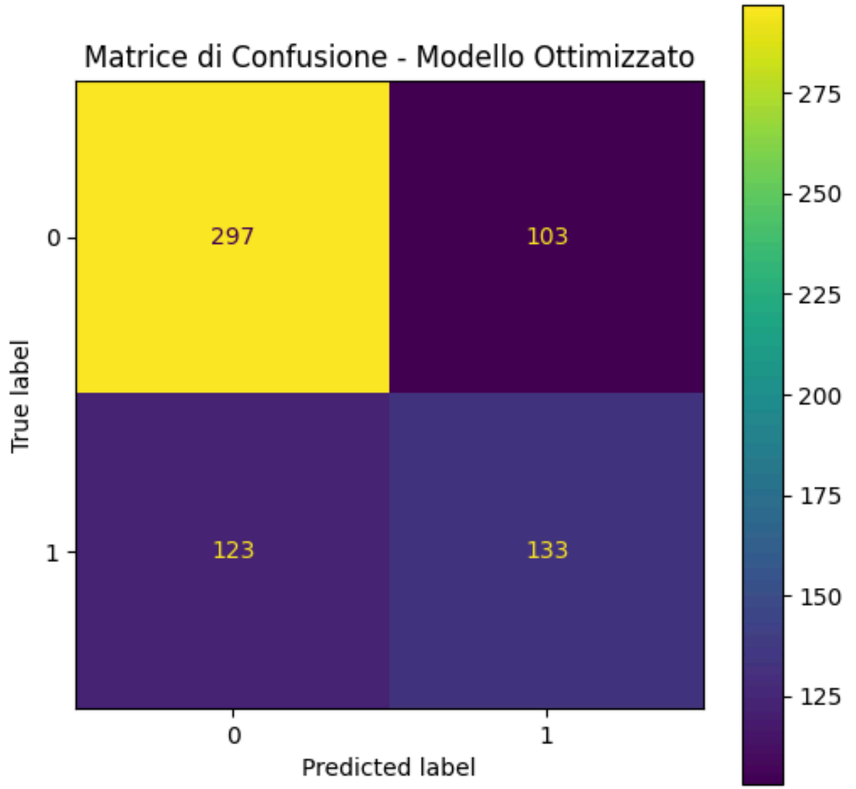




MODELLO BASE E MODELLO OTTIMIZZATO

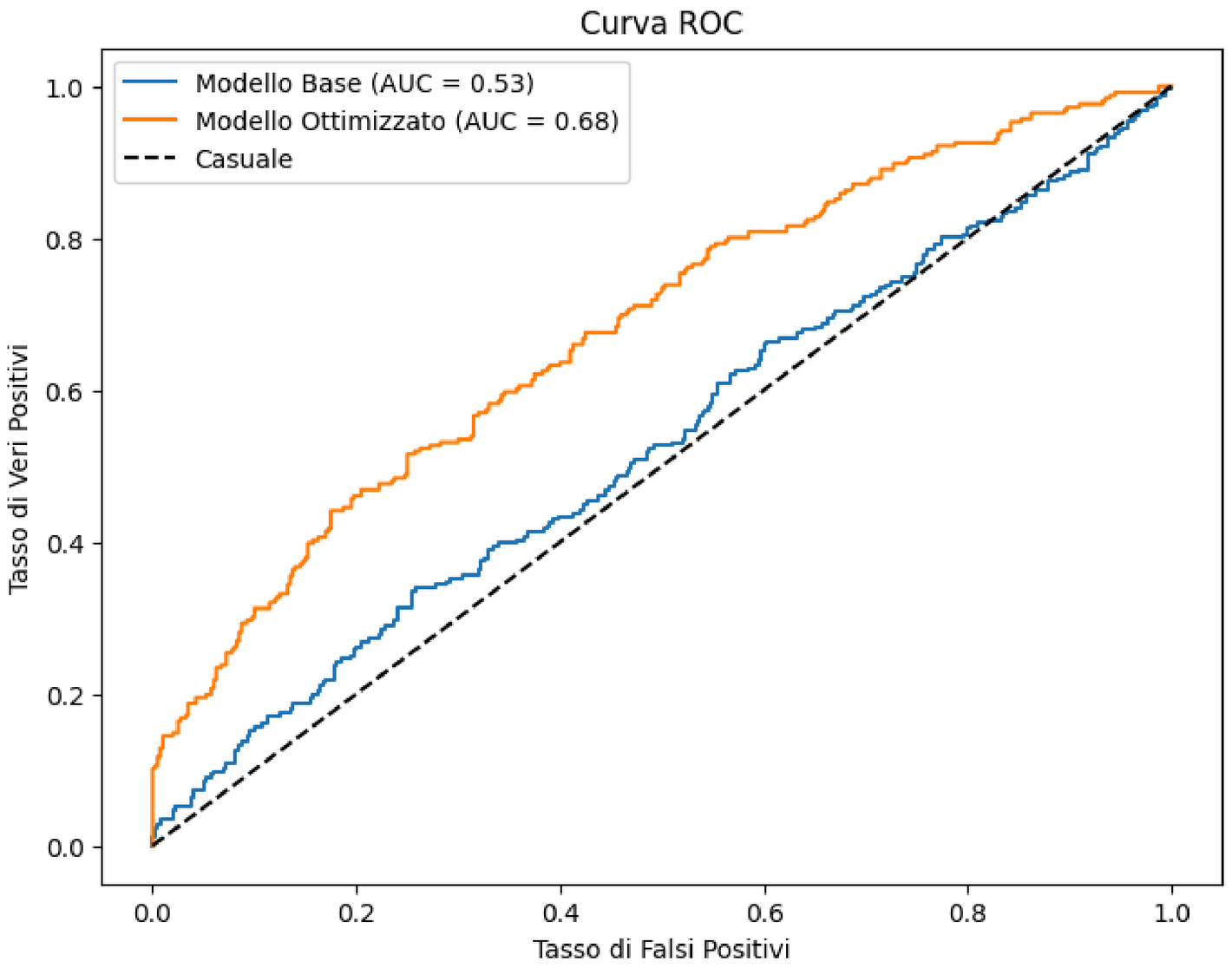


Modello base:
Risultati molto deludenti, ma lui ci fare solo da baseline.



Modello ottimizzato:
Risultati di certo migliori ma un 70% per il tipo di modello che dovremo ambire è ancora basso, quando si deve predire la potabilità dell'acqua dei valori maggiori sono necessari.

CONCLUSIONE



Qui abbiamo la conferma di quello che abbiamo discusso prima però i risultati andrebbero migliorati ulteriormente, purtroppo il grande peso del dataset non ci permette di spremere troppe informazioni.