GIULIA COSENZA PER START2IMPACT

ANALISI DELLA POTABILITA' DELL'ACQUA

EDA e Approcci di Machine Learning

PROGETTO FINALE DEL MASTER IN DATA SCIENCE

GIULIA COSENZA

1-Lo scopo del progetto

L'obiettivo di questo progetto è addestrare il miglior modello di machine learning per prevedere la potabilità dell'acqua, utilizzando informazioni relative a caratteristiche fisiche e chimiche. Prima dell’addestramento del modello, è stata effettuata un'analisi esplorativa dei dati per valutare come trattarli, selezionare le informazioni più significative e identificare il miglior modello, al fine di ottenere le migliori performance predittive.

GIULIA COSENZA

2-I dati

Il dataset utilizzato per l'addestramento del modello proviene da campionature di acqua e contiene informazioni relative a diverse caratteristiche fisiche e chimiche dell'acqua.

Il dataset è strutturato in:

- **3276** campioni di acqua
- **9** variabili numeriche indipendenti, che rappresentano caratteristiche chimiche e fisiche

Le variabili (o features):

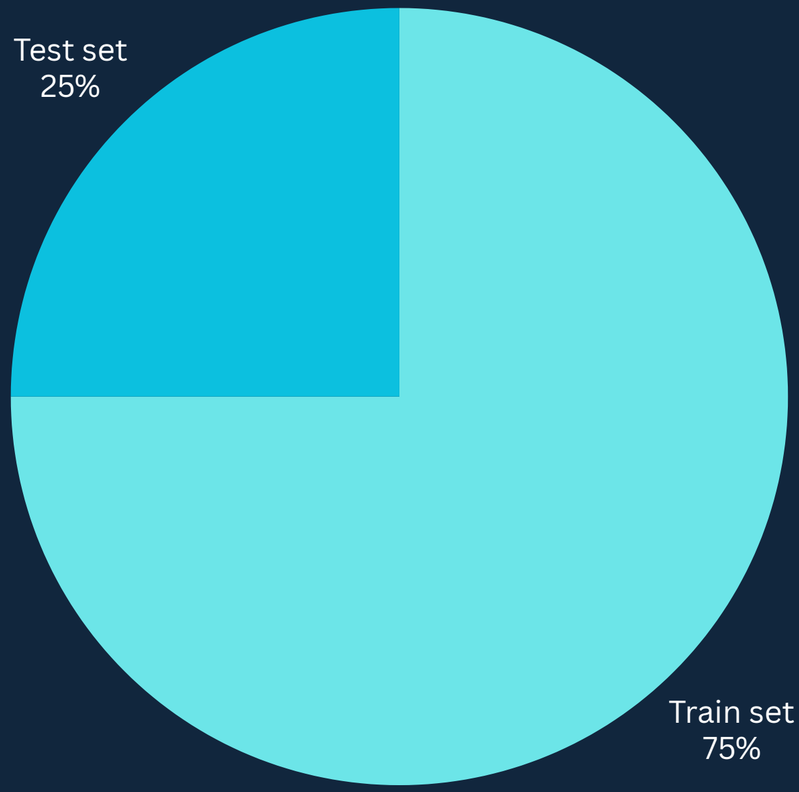
- **ph**: Misura l'acidità o l'alcalinità dell'acqua. **<float>**
- **Hardness**: Durezza dell'acqua determinata dalla concentrazione di minerali, come calcio e magnesio. **<float>**
- **Solids**: Sali disciolti nell'acqua. **<float>**
- **Chloramines**: Livello di cloroammine. **<float>**
- **Sulfate**: Livello di solfati. **<float>**
- **Conductivity**: Capacità dell'acqua di condurre elettricità. **<float>**
- **Organic_carbon**: Livello di carbone organico. **<float>**
- **Trihalomethanes**: Livello di trihalometani. **<float>**
- **Turbidity**: Misura della chiarezza dell'acqua, che è influenzata dalla presenza di particelle sospese. **<float>**

Le labels:

La variabile 'Potability' rappresenta la potabilità o la non potabilità dell'acqua, indicata rispettivamente dal valore 1 e 0.

Poiché la variabile di interesse è categorica, il problema di previsione si configura come un problema di 'CLASSIFICAZIONE'

3-Suddivisione del set



Il dataset viene suddiviso in Train set e Test set con una proporzione rispettivamente del 75% e 25%.

Prima della divisione è stato valutato se il df fosse sbilanciato.
Poichè le labels=0 rappresentavano circa il 61% di tutte le osservazioni **(df sbilanciato)** è stata fatta una suddivisione *'stratificata'* per suddividere i dati in modo che la distribuzione delle classi fosse proporzionale sia nel training set che nel test set.

4-Analisi esplorativa dei dati

1 .Presenza di valori nulli

- **Colonne con valori mancanti: pH, Sulfate, Trihalomethanes**
- **Distribuzione dei Dati: La distribuzione dei valori mancanti tra il train e il test set è simile. Tuttavia, i valori mancanti sono più frequenti nei campioni "non potabili", in particolare per le variabili 'ph' e 'Sulfate'.**

```
=========Train set=======
ph                  14.896215
Hardness             0.000000
Solids               0.000000
Chloramines          0.000000
Sulfate             24.053724
Conductivity         0.000000
Organic_carbon       0.000000
Trihalomethanes      5.087505
Turbidity            0.000000
dtype: float64

=========Test set=======
ph                  15.262515
Hardness             0.000000
Solids               0.000000
Chloramines          0.000000
Sulfate             23.199023
Conductivity         0.000000
Organic_carbon       0.000000
Trihalomethanes      4.517705
Turbidity            0.000000
```

```
 ph                  9.584860
Hardness             0.000000
Solids               0.000000
Chloramines          0.000000
Sulfate             14.896215
Conductivity         0.000000
Organic_carbon       0.000000
Trihalomethanes      3.266178
Turbidity            0.000000
Potability           0.000000
dtype: float64

=======Potabile========
 ph                  5.402930
Hardness             0.000000
Solids               0.000000
Chloramines          0.000000
Sulfate              8.943834
Conductivity         0.000000
Organic_carbon       0.000000
Trihalomethanes      1.678877
Turbidity            0.000000
```

percentuale di dati mancanti per ciascuna feature

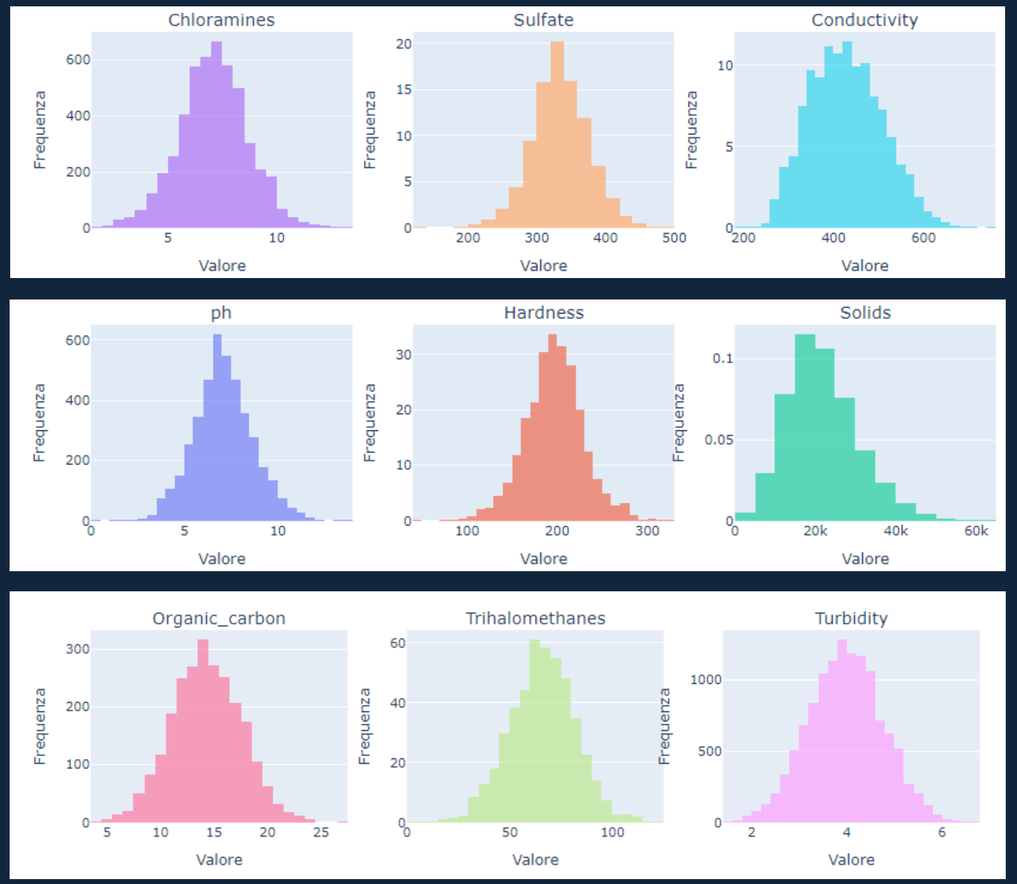
La percentuale di valori mancanti tra il train set e il test set risulta molto simile, suggerendo che i dati sono distribuiti in modo uniforme.

Se le colonne con valori mancanti, come Sulfate e pH, sono ritenute importanti per il modello, potremmo considerare l'imputazione dei valori mancanti.

4-Analisi esplorativa dei dati

2. Distribuzione delle features

- La maggior parte delle variabili ha distribuzioni simmetriche, alcune delle quali approssimano una forma normale (es. pH, Hardness, Chloramines).
- La variabile Solids ha una distribuzione asimmetrica con una coda verso destra, indicando possibili outlier o una distribuzione naturalmente sbilanciata.
- Le variabili hanno intervalli diversi: alcune, come pH e Turbidity, hanno scale limitate, mentre altre, come Solids e Conductivity, hanno scale più ampie.



4-Analisi esplorativa dei dati

2. Distribuzione delle features

- La maggior parte delle variabili ha distribuzioni simmetriche, alcune delle quali approssimano una forma normale (es. pH, Hardness, Chloramines).
- La variabile Solids ha una distribuzione asimmetrica con una coda verso destra, indicando possibili outlier o una distribuzione naturalmente sbilanciata.
- Le variabili hanno intervalli diversi: alcune, come pH e Turbidity, hanno scale limitate, mentre altre, come Solids e Conductivity, hanno scale più ampie.

→ **Il dataset appare ben distribuito e adatto ad analisi statistiche e modellazione, ma la presenza di variabili con scale diverse e alcune asimmetrie richiedono pre-processing (standardizzazione o trasformazione)**

4-Analisi esplorativa dei dati

2. Distribuzione delle features

- Alcune feature, come pH, Hardness, Chloramines, Conductivity, Organic Carbon, Turbidity, mostrano distribuzioni abbastanza simili per acqua potabile e non potabile.
- Solids: L'acqua non potabile sembra avere una distribuzione leggermente spostata verso valori più alti.
- Sulfate: Anche qui si nota una leggera differenza, con l'acqua non potabile che tende ad avere valori più alti.

4-Analisi esplorativa dei dati

2. Distribuzione delle features

- **Alcune feature, come pH, Hardness, Chloramines, Conductivity, Organic Carbon, Turbidity, mostrano distribuzioni abbastanza simili per acqua potabile e non potabile.**
- **Solids: L'acqua non potabile sembra avere una distribuzione leggermente spostata verso valori più alti.**
- **Sulfate: Anche qui si nota una leggera differenza, con l'acqua non potabile che tende ad avere valori più alti.**

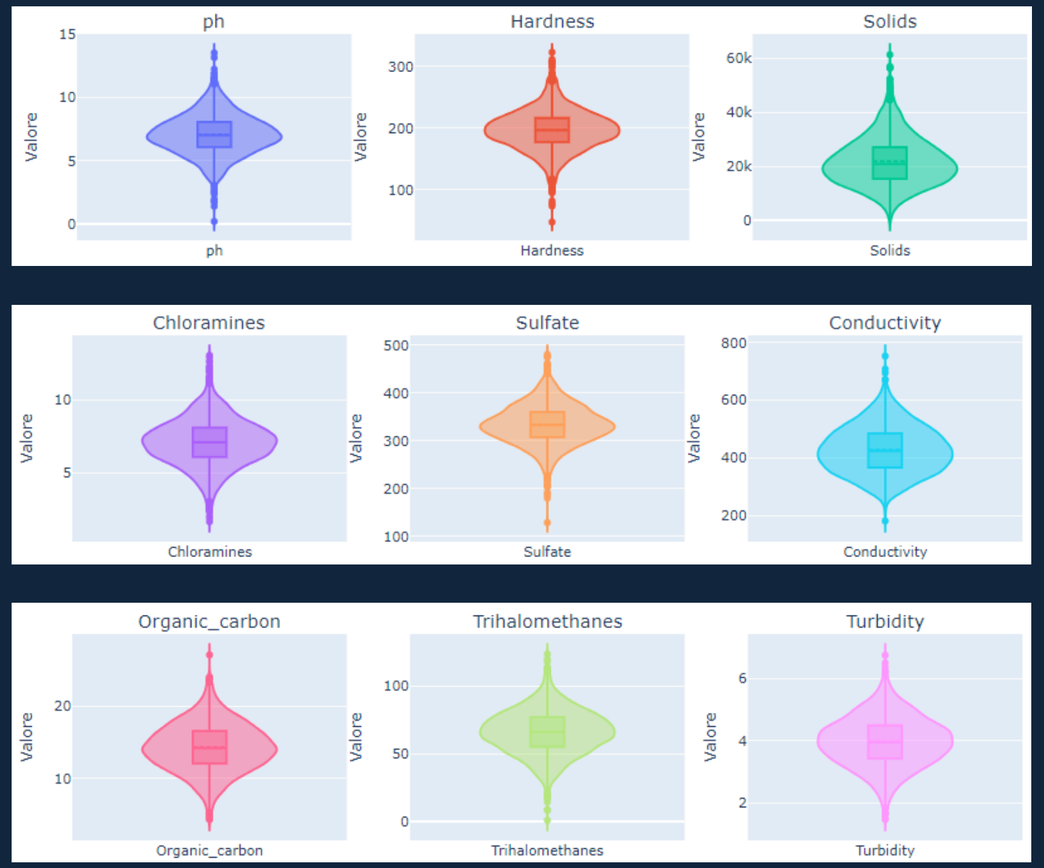
La separabilità tra le due classi non è netta per nessuna feature singolarmente, il che suggerisce che una classificazione basata su soglie semplici non sarebbe efficace. Probabilmente, una combinazione di più variabili tramite un modello statistico o di machine learning sarà necessaria per distinguere in modo efficace l'acqua potabile da quella non potabile.

• Feature come Solids e Sulfate potrebbero essere più influenti nella determinazione della potabilità rispetto ad altre variabili.

4-Analisi esplorativa dei dati

3. Presenza di valori outliers

- **L'analisi dei ViolinPlot evidenzia una presenza non trascurabile di outlier che possono influenzare la qualità dell'analisi e della modellazione.**
- **Le variabili Hardness, Chloramines, Solids, Sulfates e ph in particolare mostra un'elevata presenza di outlier superiori, spesso molto distanti dai valori principali.**



4-Analisi esplorativa dei dati

3. Presenza di valori outliers

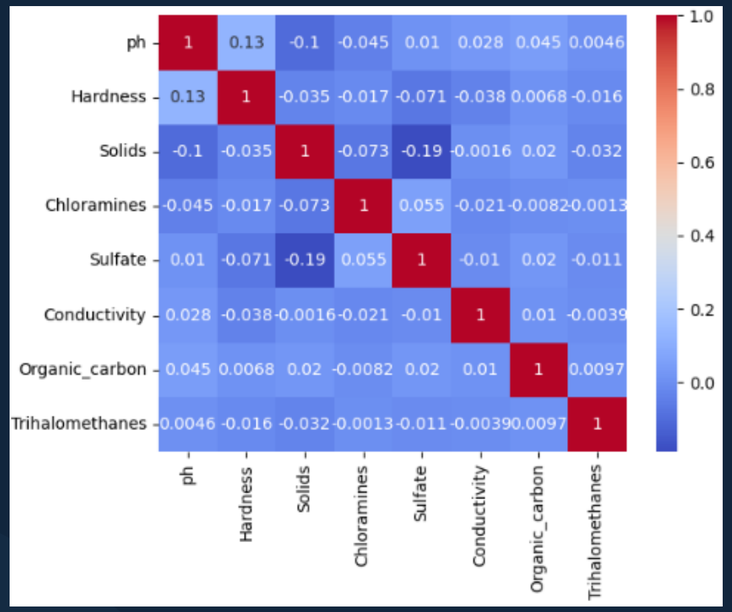
- L'analisi dei ViolinPlot evidenzia una presenza non trascurabile di outlier che possono influenzare la qualità dell'analisi e della modellazione.
- Le variabili Hardness, Chloramines, Solids, Sulfates e ph in particolare mostra un'elevata presenza di outlier superiori, spesso molto distanti dai valori principali.

```
Percentuale di outliers in ph: 1.26%
Percentuale di outliers in Hardness: 2.85%
Percentuale di outliers in Solids: 1.51%
Percentuale di outliers in Chloramines: 1.91%
Percentuale di outliers in Sulfate: 1.26%
Percentuale di outliers in Conductivity: 0.24%
Percentuale di outliers in Organic_carbon: 0.77%
Percentuale di outliers in Trihalomethanes: 0.98%
```

**La forte presenza di outlier in alcune variabili richiede una gestione attenta per evitare distorsioni analitiche.
La pre-elaborazione dei dati, inclusa la rimozione o trasformazione degli outlier, è fondamentale per ottenere risultati più accurati e rappresentativi.**

4-Analisi esplorativa dei dati

4. Correlazione tra feature e labels	**Le variabili hanno correlazioni generalmente basse tra loro, indicando che nessuna singola feature è fortemente predittiva delle altre. Le correlazioni più alte sono tra Solids e Sulfate, ma restano deboli.**



Questo suggerisce che tutte le variabili contribuiscono in modo indipendente alla classificazione della potabilità e che sarà necessario un approccio di machine learning, preferibilmente con modelli non lineari (es. Random Forest). Inoltre, un'analisi di feature importance potrebbe aiutare a individuare le variabili più influenti.

4-Analisi esplorativa dei dati

5. Rilevanza delle features

Prima dei test statistici sono state rimosse le righe con valori mancanti che avrebbero potuto interferire

- ***T-TEST:*** **nessuna feature ha mostrato p-value < 0.05 significa che, secondo il test statistico, non ci sono evidenze sufficienti per rifiutare l'ipotesi nulla (H_0). In altre parole, non esiste una differenza statisticamente significativa tra le medie dei due gruppi.**

```
ph
T-statistic: -1.701208729587031, p-value: 0.08913390664301572

Hardness
T-statistic: -1.5275140239988816, p-value: 0.12686731805985885

Solids
T-statistic: -1.6312918301840482, p-value: 0.10306182403280839

Chloramines
T-statistic: -0.6328590905616739, p-value: 0.5269327653707445
```

```
Sulfate
T-statistic: -1.0559935002401897, p-value: 0.2911599803356202

Conductivity
T-statistic: 0.40609963747063277, p-value: 0.6847337045576771

Organic_carbon
T-statistic: -0.3160080191746569, p-value: 0.7520452340575812

Trihalomethanes
T-statistic: -0.277264884531486, p-value: 0.7816191389046245

Turbidity
T-statistic: -0.9879842836590009, p-value: 0.32333725293914384
```

4-Analisi esplorativa dei dati

5. Rilevanza delle features

Prima dei test statistici sono state rimosse le righe con valori mancanti che avrebbero potuto interferire

- ***MANOVA TEST:*** **le combinazioni ['ph', 'Hardness', 'Solids', 'Sulfate'] , ['ph', 'Solids', 'Sulfate'], ['ph', 'Hardness', 'Solids', 'Sulfate', 'Turbidity'], ['ph', 'Solids', 'Sulfate'], ['ph', 'Hardness', 'Solids'] e ['ph', 'Solids'] mostrano p-values < 0.05**

```
Combinazione: ['ph', 'Hardness', 'Solids', 'Sulfate']
- P-value: 0.0354
- Wilks' Lambda: 0.9924
- Pillai's Trace: 0.0076
- Hotelling-Lawley Trace: 0.0077
- Roy's Largest Root: 0.0077
Combinazione: ['ph', 'Hardness', 'Solids', 'Sulfate', 'Turbidity']
- P-value: 0.0418
- Wilks' Lambda: 0.9915
- Pillai's Trace: 0.0085
- Hotelling-Lawley Trace: 0.0086
- Roy's Largest Root: 0.0086
```

```
Combinazione: ['ph', 'Solids', 'Sulfate']
- P-value: 0.0446
- Wilks' Lambda: 0.9940
- Pillai's Trace: 0.0060
- Hotelling-Lawley Trace: 0.0060
- Roy's Largest Root: 0.0060
Combinazione: ['ph', 'Hardness', 'Solids']
- P-value: 0.0466
- Wilks' Lambda: 0.9941
- Pillai's Trace: 0.0059
- Hotelling-Lawley Trace: 0.0059
- Roy's Largest Root: 0.0059
```

```
Combinazione: ['ph', 'Solids']
- P-value: 0.0469
- Wilks' Lambda: 0.9955
- Pillai's Trace: 0.0045
- Hotelling-Lawley Trace: 0.0045
- Roy's Largest Root: 0.0045
```

4-Analisi esplorativa dei dati

5. Rilevanza delle features

Prima dei test statistici sono state rimosse le righe con valori mancanti che avrebbero potuto interferire

- ***T-TEST:*** **nessuna feature ha mostrato p-value < 0.05 significa che, secondo il test statistico, non ci sono evidenze sufficienti per rifiutare l'ipotesi nulla (H_0). In altre parole, non esiste una differenza statisticamente significativa tra le medie dei due gruppi.**

- ***MANOVA TEST:*** **le combinazioni ['ph', 'Hardness', 'Solids', 'Sulfate'] , ['ph', 'Solids', 'Sulfate'], ['ph', 'Hardness', 'Solids', 'Sulfate', 'Turbidity'], ['ph', 'Solids', 'Sulfate'], ['ph', 'Hardness', 'Solids'] e ['ph', 'Solids'] mostrano p-values < 0.05**

I risultati mostrano che, sebbene il T-test per le singole caratteristiche abbia restituito p-value superiori a 0.05 (indicando nessuna differenza significativa tra i gruppi per ciascuna variabile singolarmente), l'analisi delle combinazioni di variabili ha prodotto p-value inferiori a 0.05, suggerendo che la combinazione di queste caratteristiche potrebbero ben differenziare i due gruppi (Acque potabili e non).

5-Pre-processing del dataset

1.Rimozione degli outliers

E' stata implementata una funzione per la rimozione degli Outliers con il metodo IQR.
Gli outlier sono stati rimossi solo dal train set.
Rimuovere gli outliers dal training set aiuta il modello a imparare meglio, mentre mantenerli nel test set assicura che la valutazione del modello rifletta accuratamente le sue capacità in situazioni reali.

2.Scalatura dei dati

3.Imputazione dei dati mancanti

Gli outliers sono stati rimossi prima di fare qualsiasi analisi statistica (come la correlazione, il T-test, il test MANOVA e l'imputazione) per evitare che influenzassero i risultati, distorcendo la valutazione delle relazioni tra le variabili e compromettendo l'accuratezza dei modelli statistici.

5-Pre-processing del dataset

1.Rimozione degli outliers

E' stata implementata una funzione per la rimozione degli Outliers con il metodo IQR. Gli outlier sono stati rimossi **solo dal train set**.
Rimuovere gli outliers dal training set aiuta il modello a imparare meglio, mentre mantenerli nel test set assicura che la valutazione del modello rifletta accuratamente le sue capacità in situazioni reali.

2.Scalatura dei dati

La scalatura è stata applicata **sia sul train che sul test set** per garantire che i dati fossero scalati correttamente e in modo consistente, rispettando la stessa trasformazione su entrambi i set. Questo aiuta a evitare distorsioni nei modelli e assicura una valutazione realistica delle performance.

3.Imputazione dei dati mancanti

E' stato impiegato KNNImputer per l'imputazione dei valori mancanti **solo nel train set**. KNNImputer sfrutta la somiglianza tra le osservazioni per imputare i valori mancanti, senza introdurre distorsioni significative conservando la variabilità nelle variabili e permettendo di catturare relazioni non lineari tra le features.

6-Addestramento e predizioni

Modelli da testare: Logistic regression, Random Forest, K-neares Neighbors

STRADA 1

Valutare i modelli utilizzando tutte le features

STRADA 2

Valutare i modelli utilizzando solo alcune feature selezionate

La combinazione scelta è stata: **['ph', 'Hardness', 'Solids', 'Sulfate']**.

Mostrava infatti il p-value più basso. Inoltre l'aggiunta di Turbidity non migliorava le statistiche di separazione e tra le combinazioni ['ph', 'Solids', 'Sulfate'] e ['ph', 'Hardness', 'Solids'] le metriche erano troppo simi tanto da rendere la scelta di Sulfate piuttosto che Hardness non è ovvia.

6-Addestramento e predizioni

STRADA 1 **SPOT CHECK**

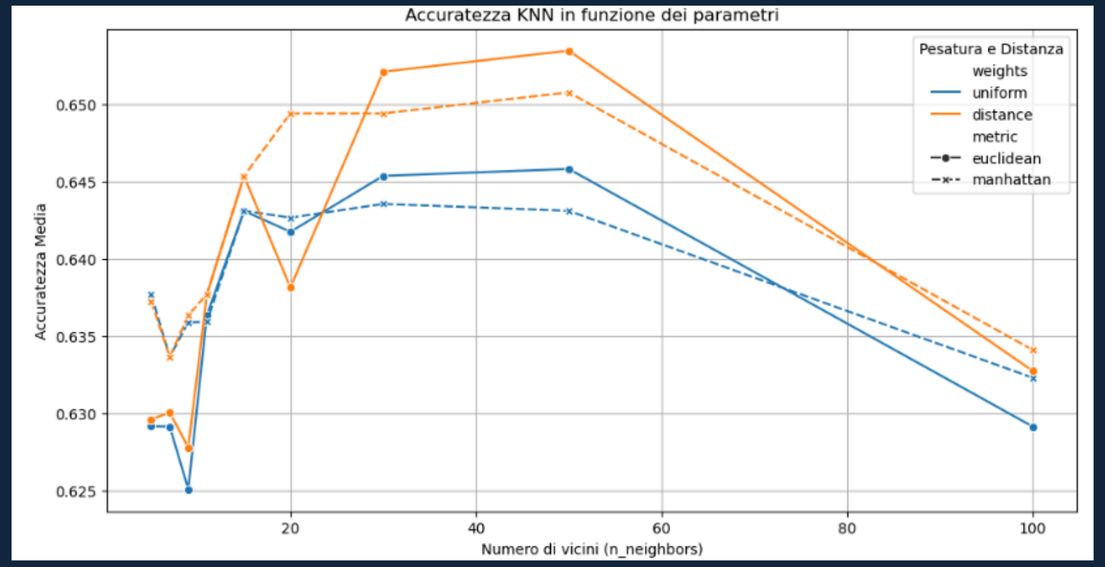
Lo Spot Check ha individuato come migliori modelli il Random Forest seguito dal KNN.

```
Spot check con tutte le feature:
{'Logistic Regression': 0.6165636488506443, 'Random Forest': 0.657516260655274l, 'K-nearest Neighbors': 0.6291803013776108}
```

6-Addestramento e predizioni

STRADA 1 → SPOT CHECK → **TUNING DEGLI IPERPARAMETRI**

n_neighbors	[5, 7, 9, 11, 15, 20, 30, 50, 100]
weights	[uniform, distance]
metric	[euclidean, manhattan]

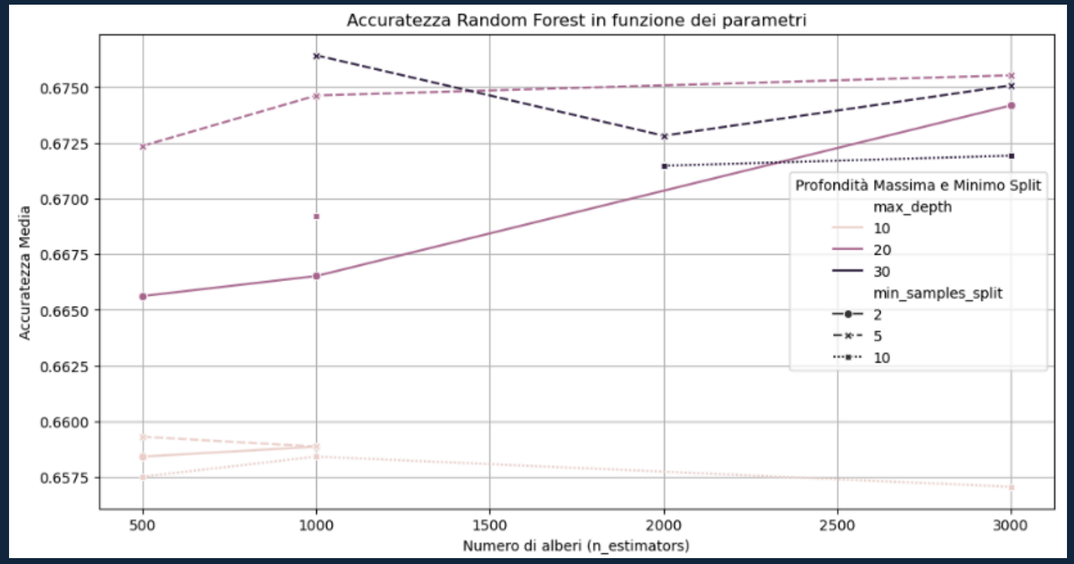


Migliori parametri:
{'metric': 'euclidean', 'n_neighbors': 50, 'weights': 'distance'}

6-Addestramento e predizioni

STRADA 1 → SPOT CHECK → **TUNING DEGLI IPERPARAMETRI**

n_estimators	[500, 1000, 2000, 3000]
max_depth	[10, 20, 30]
min_samples_split	[2, 5, 10]



Migliori parametri:
{'n_estimators': 1000, 'min_samples_split': 5, 'max_depth': 30}

6-Addestramento e predizioni

STRADA 1 → SPOT CHECK → TUNING DEGLI IPERPARAMETRI → **PREDIZIONE**

Modello	Accuracy	Precision	Recall	F1-score
KNN	0.63	0.62	0.12	0.2
Random Forest	0.65	0.61	0.30	0.40

6-Addestramento e predizioni

STRADA 2 **SPOT CHECK**

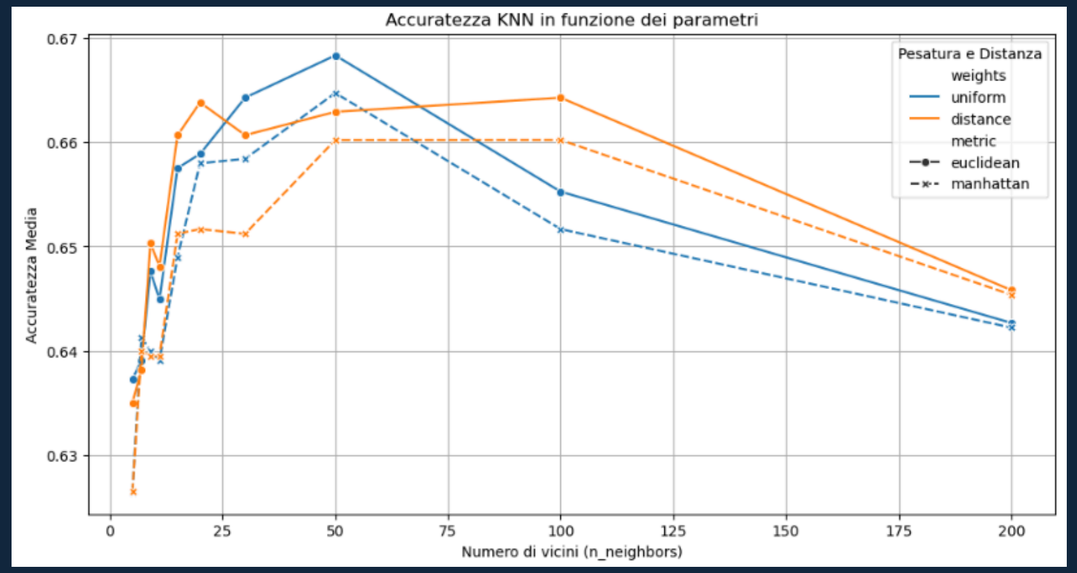
Lo Spot Check ha individuato come migliori modelli il Random Forest seguito dal KNN.

```
Spot check con feature selezionate:
{'Logistic Regression': 0.6165636488506443, 'Random Forest': 0.6620328849028401, 'K-nearest Neighbors': 0.6372702298711268}
```

6-Addestramento e predizioni

STRADA 2 → SPOT CHECK → **TUNING DEGLI IPERPARAMETRI**

n_neighbors	[5, 7, 9, 11, 15, 20, 30, 50, 100]
weights	[uniform, distance]
metric	[euclidean, manhattan]

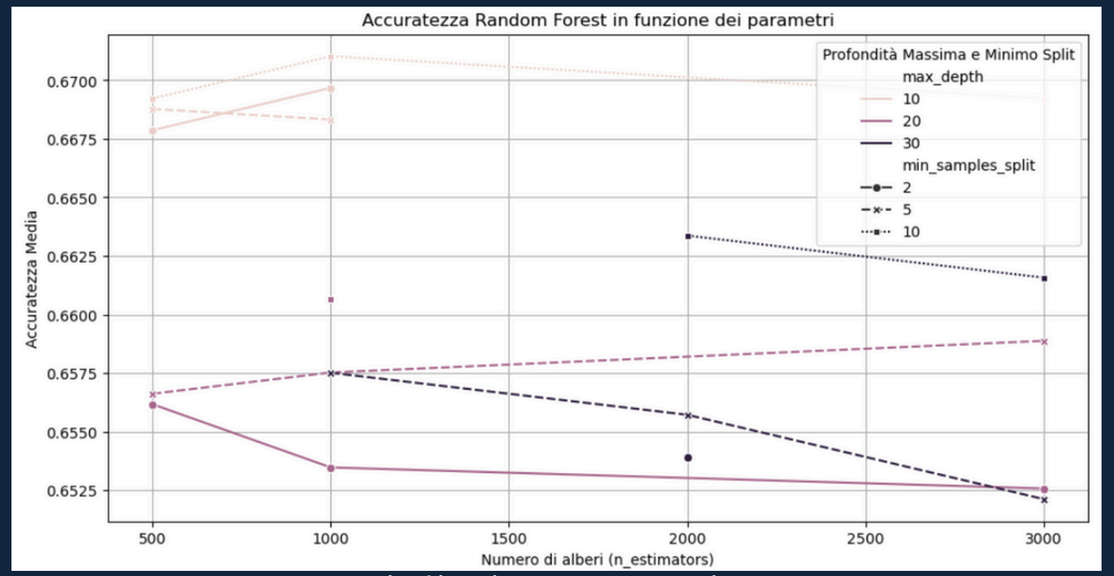


Migliori parametri:
{'metric': 'euclidean', 'n_neighbors': 50, 'weights': 'uniform'}

6-Addestramento e predizioni

STRADA 2 → SPOT CHECK → **TUNING DEGLI IPERPARAMETRI**

n_estimators	[100, 500, 1000, 2000, 3000]
max_depth	[2, 5, 10, 20, 30]
min_samples_split	[2, 5, 10, 20, 50]



Migliori parametri:
{'n_estimators': 1000, 'min_samples_split': 10, 'max_depth': 10}

6-Addestramento e predizioni

STRADA 2 → SPOT CHECK → TUNING DEGLI IPERPARAMETRI → **PREDIZIONE**

Modello	Accuracy	Precision	Recall	F1-score
KNN	0.67	0.72	0.25	0.37
Random Forest	0.66	0.64	0.28	0.39

7-Risultati

STRADA 1: TUTTE LE FEATURES

	KNN	Random Forest
Accuracy	0.63	0.65
Precision	0.62	0.61
Recall	0.12	0.30
F1-score	0.2	0.40

STRADA 2: FEATURES SELEZIONATE

	KNN	Random Forest
Accuracy	0.67	0.66
Precision	0.72	0.64
Recall	0.25	0.28
F1-score	0.37	0.39

8-Discussione e Conclusioni

- **Problema:** L'obiettivo del progetto era prevedere la potabilità dell'acqua utilizzando un dataset con 9 variabili. Poiché la variabile target ("Potability") è categorica, si trattava di un *problema di classificazione*. Le principali sfide includevano valori mancanti, outlier e variabili con distribuzioni differenti.

- **Approccio:**
  - **Pre-elaborazione dei dati:** Imputazione dei valori mancanti, rimozione degli outlier nel training set, e scalatura dei dati per garantire coerenza.
  - **Approccio di modellazione a due strade:**
    - Testare i modelli con tutte le features.
    - Testare i modelli con selezione delle features per ottimizzare le performance. **->** T-test e MANOVA Test
  - **Modelli testati:** Logistic Regression, Random Forest, KNN.
  - **Tuning dei parametri:** Sono stati ottimizzati i parametri dei modelli per migliorare la performance.

8-Discussione e Conclusioni

'Strada' migliore:

- KNN con la selezione delle features (STRADA 2) migliora tutte le metriche di performance rispetto al modello allenato con tutte le features (STRADA 1).
- Random Forest con la selezione delle features (STRADA 2) migliora l'accuracy e la precision rispetto al modello allenato con tutte le features (STRADA 1). La recall, invece, diminuisce leggermente riflettendosi in una leggera diminuzione del F1-score.

Modello migliore:

- Random Forest è il modello più performante, grazie al suo miglior recall e F1-score rispetto a KNN.
- Nonostante KNN migliori nella precisione con la selezione delle features (72%), Random Forest bilancia meglio precisione e recall, risultando più adatto al task di predizione.

Interpretazione dell'accuratezza:

- Se consideriamo solo l'accuratezza, KNN con la selezione delle features(STRADA 2) (67%) supera Random Forest (66%).
- Tuttavia, l'accuratezza non sempre riflette la capacità di predizione.

Link al codice

https://drive.google.com/file/d/13Ug4qHGdXT496MfkI290UKgYl0Vy076X/view?usp=sharing