Acqua Potabile e non :
classificazione con Machine Learning

EDA

Preprocessing

Machine Learning

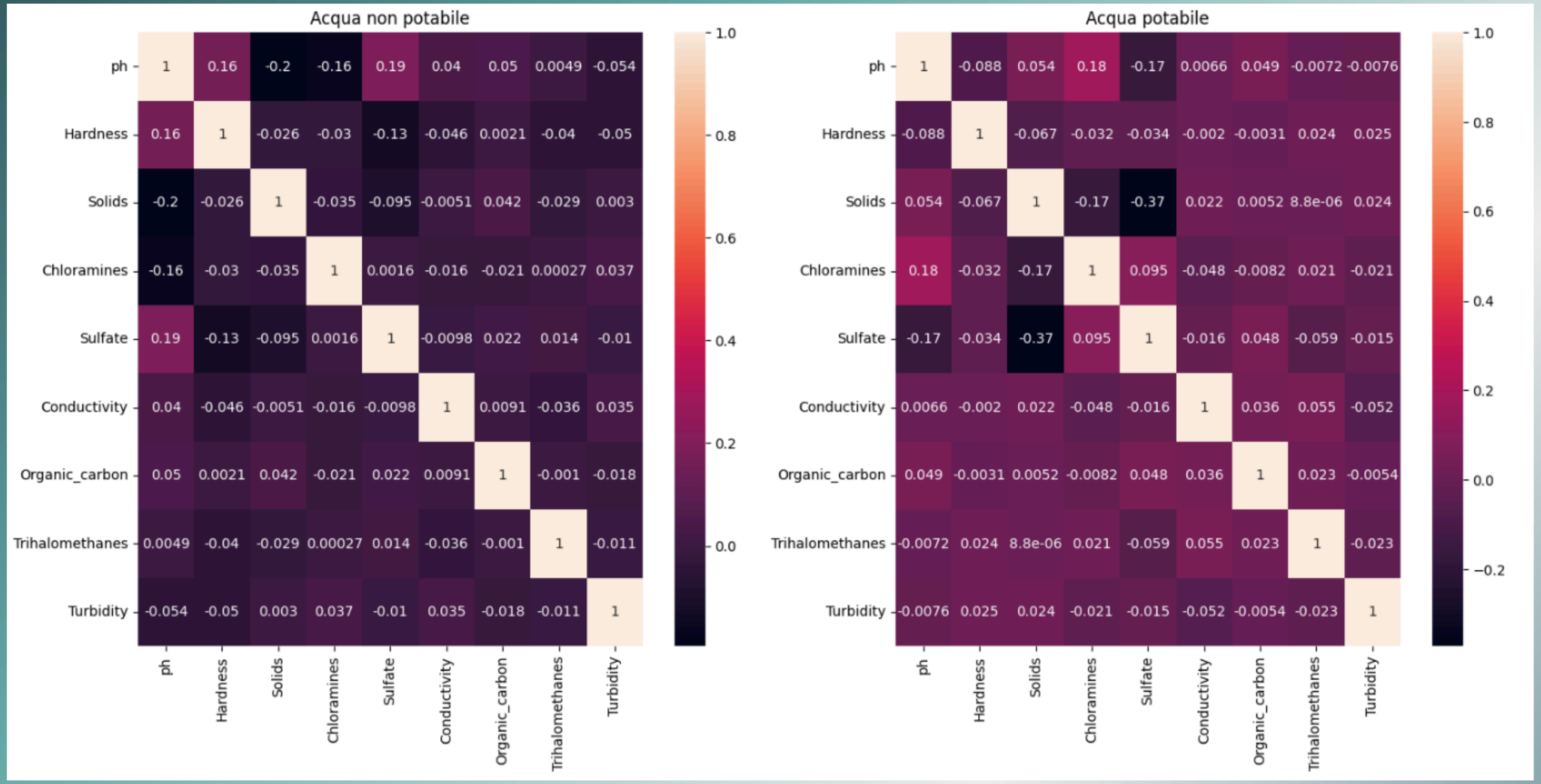
Nicola Carcangiu

EDA

- sulla parte di dataset usata come training set, proporzione tra classi originale;
- analisi differenze nei dati tra acqua potabile e non, impiego di Test Statistici:
  - Levene
  - Shapiro
  - T-Test
  - Mann-Whitney U

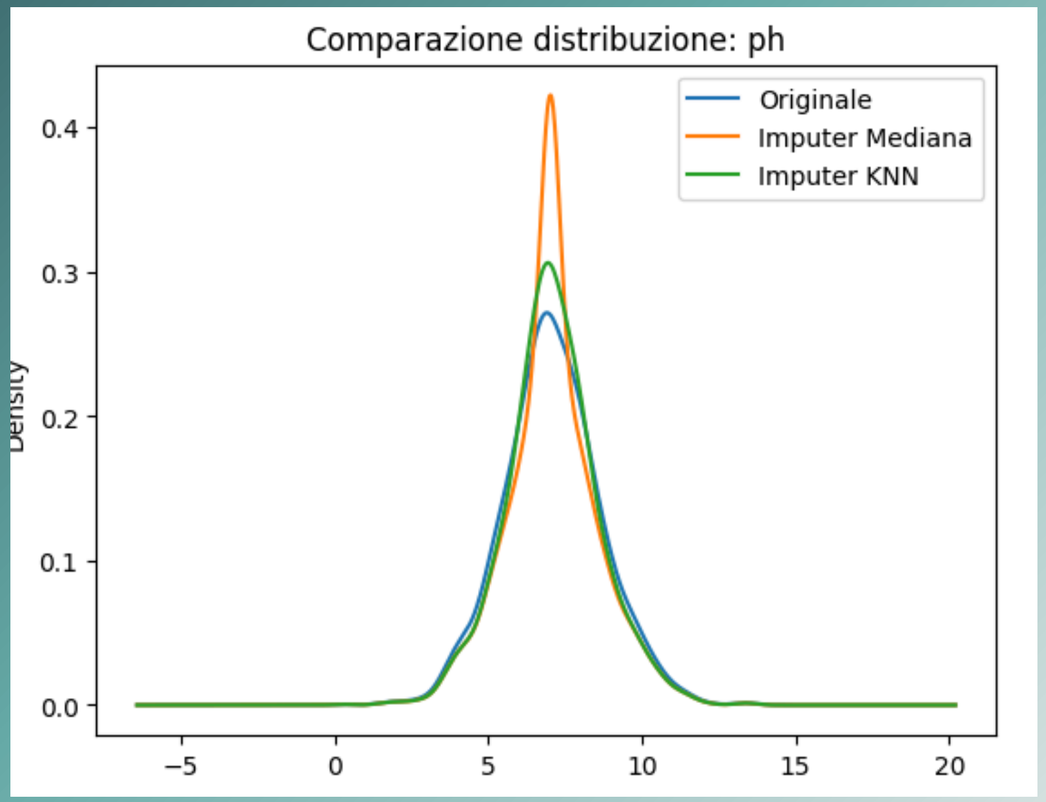
Nessuna differenza significativa trovata

- analisi correlazioni tra features:



<u>Valori molto bassi: il dataset non ha features ridondanti</u>

Preprocessing

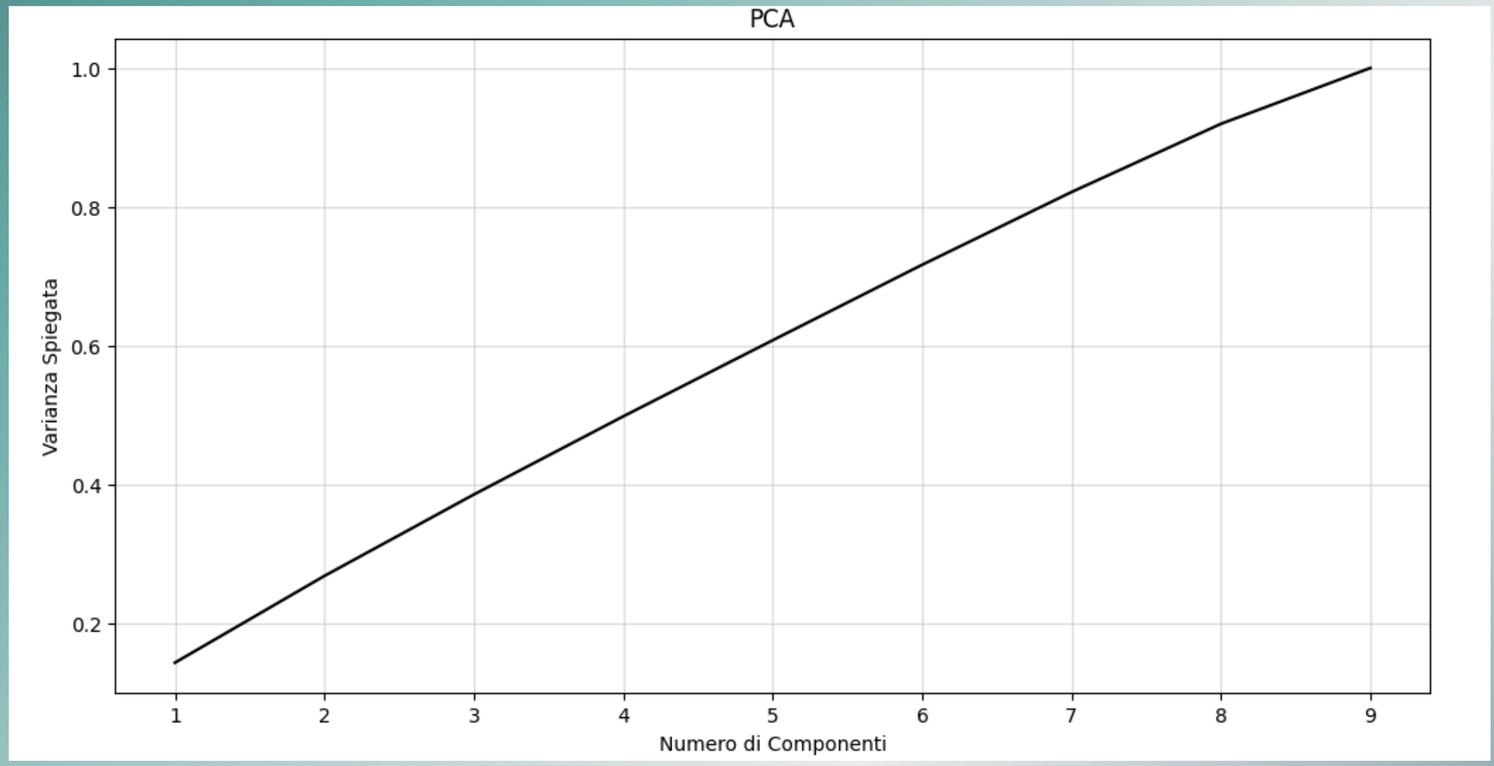


KNN Imputer riesce a conservare più fedelmente (rispetto all'imputazione con mediana) la distribuzione originale dei dati.

Applicata a 3 features:

- ph
- Trihalomethanes
- Sulfate

I test statistici applicati non hanno evidenziato particolari differenze tra le variabili;
anche un approccio meno interpretabile (PCA) non riesce a ridurre la dimensionalità del dataset senza comprometterne l’informatività:



Machine Learning

Obiettivo: Progettare modello di Classificazione Binaria

Metrica scelta: Accuracy, Precisione classe 1

Precisione Classe 1: Importante ridurre al minimo i falsi positivi

Accuracy Benchmark: 51%

(Dummy Classifier, Strategia = Stratificato)

Workflow seguito

1. Train Test split
2. Pipeline:
   2.1 KNN Imputer
   2.2 Standardizzazione
3. Trasformazione dati su pipeline
4. Spotcheck con e senza feature selection

usando Mutual Information per trovare le tre feats più decisive: *Sulfate, Hardness, ph*

5. Grid Search sui due migliori modelli
6. Confronto tra Top - Tutte features e
   Top - Feat Selection

dettagli Spot Check:

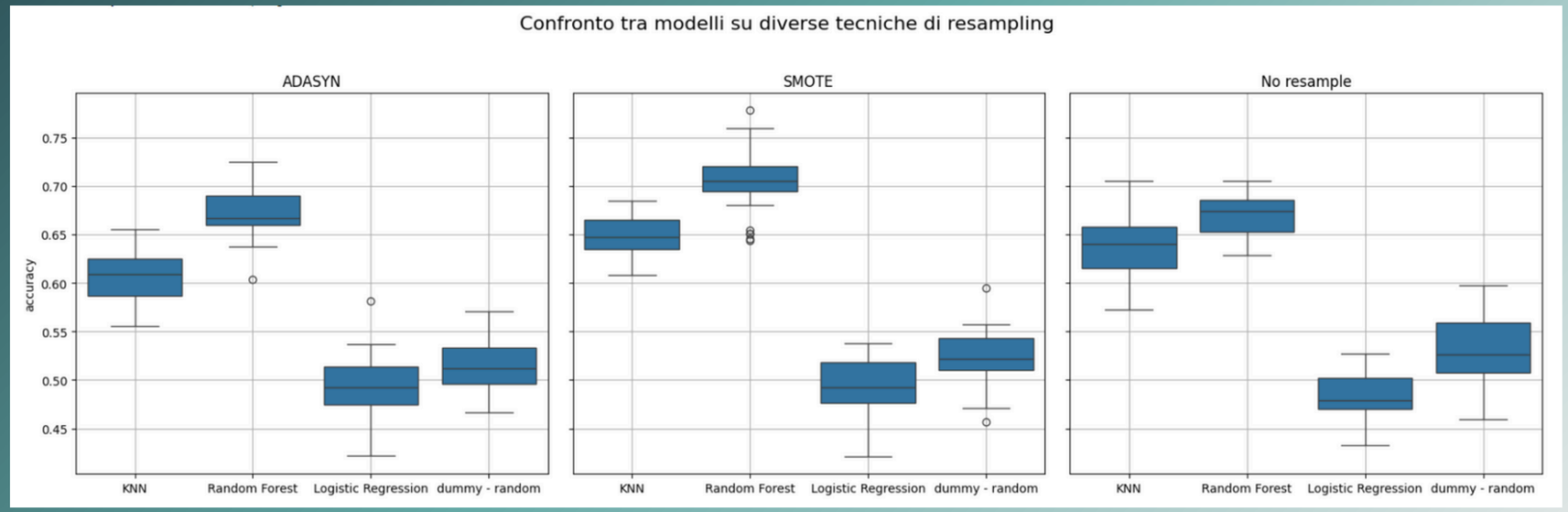
provate tutte le combinazioni possibili con un ‘for loop’

Algoritmi Scelti: Logistic Regression, KNN, Random Forest

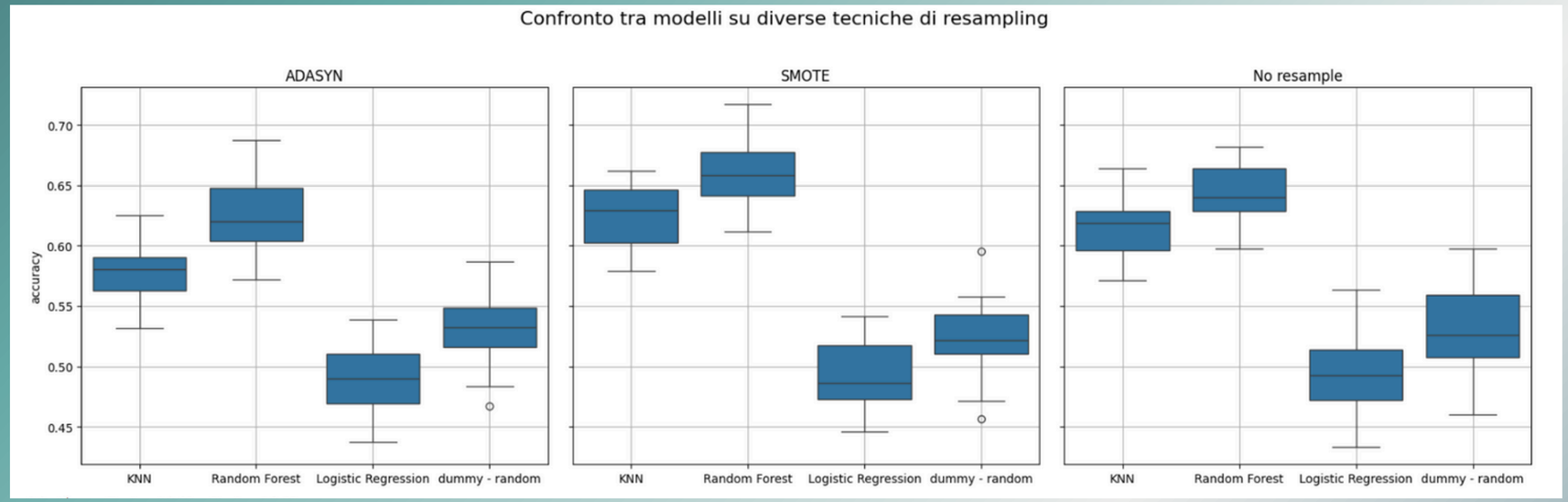
Tecnica di Bilanciamento: Nessuna, SMOTE, ADASYN

Validazione: RepeatedStratifiedKFold (10 splits, 3 ripetizioni)

Risultati Spotcheck



tutte le features



feature selection

- in entrambi i casi, KNN e Random Forest emergono come Top 2;
- SMOTE dà risultati più stabili, scelto come tecnica di oversampling

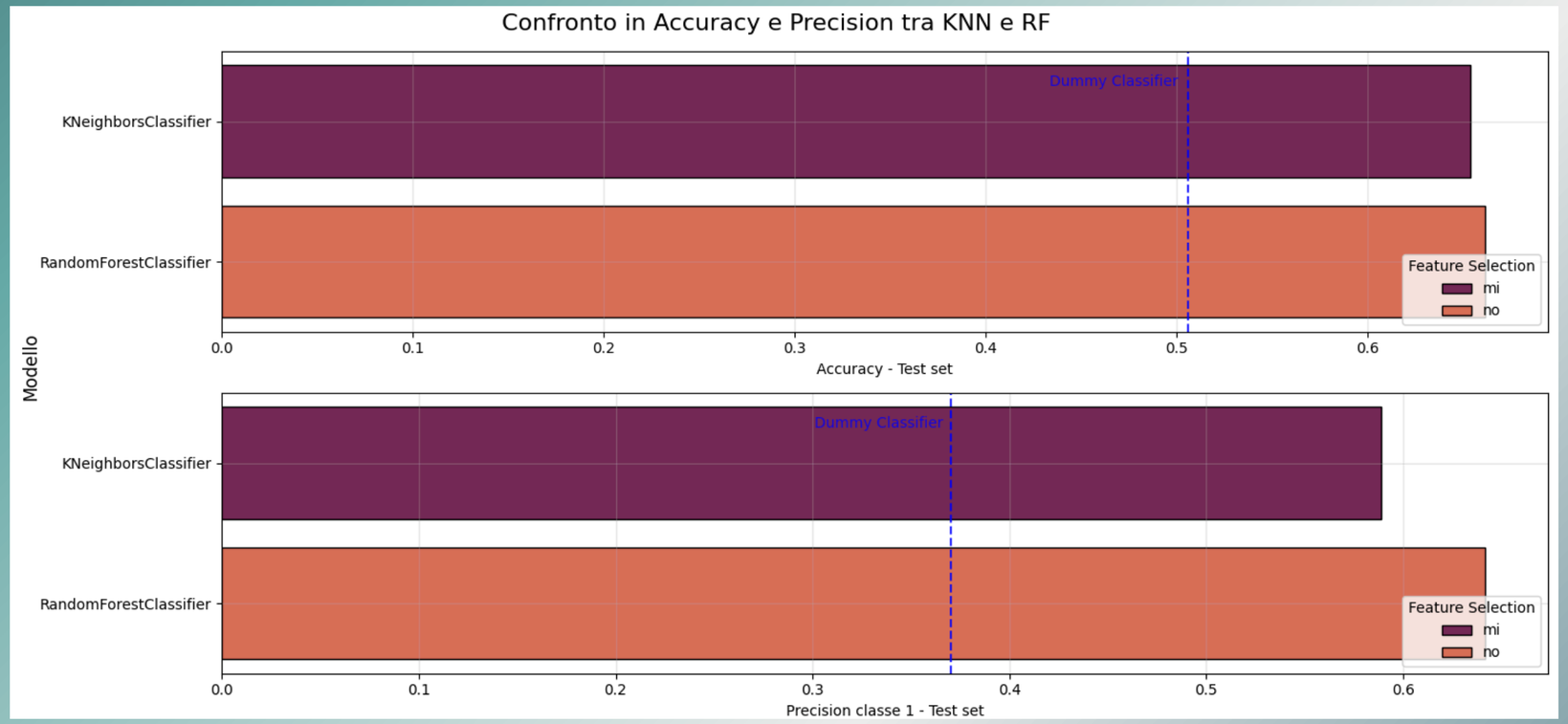
Miglior Modello:

senza feature selection:

Random Forest

con feature selection:

KNN



Conclusioni

- Random Forest addestrato su dataset senza oversampling è il miglior modello per performance attese (sia Accuracy che Precision 1);
- i tempi di training ed inferenza sono praticamente uguali tra KNN e RF, mentre la GridSearch di RF è considerevolmente più lunga;
- se vi fosse esigenza di un traninig più rapido ma usando comunque una GridSearch, la scelta dovrebbe andare su KNN;