# Predizione del Diabete di Tipo 1: Uno Studio sul Ruolo del Genoma per la Costruzione di Modelli di Machine Learning Explainable


Prof. Fabio Palomba
Dott. Antonio della Porta
Dott.ssa Viviana Pentangelo

Rosa Carotenuto
Mat. 0512113246


Tesi

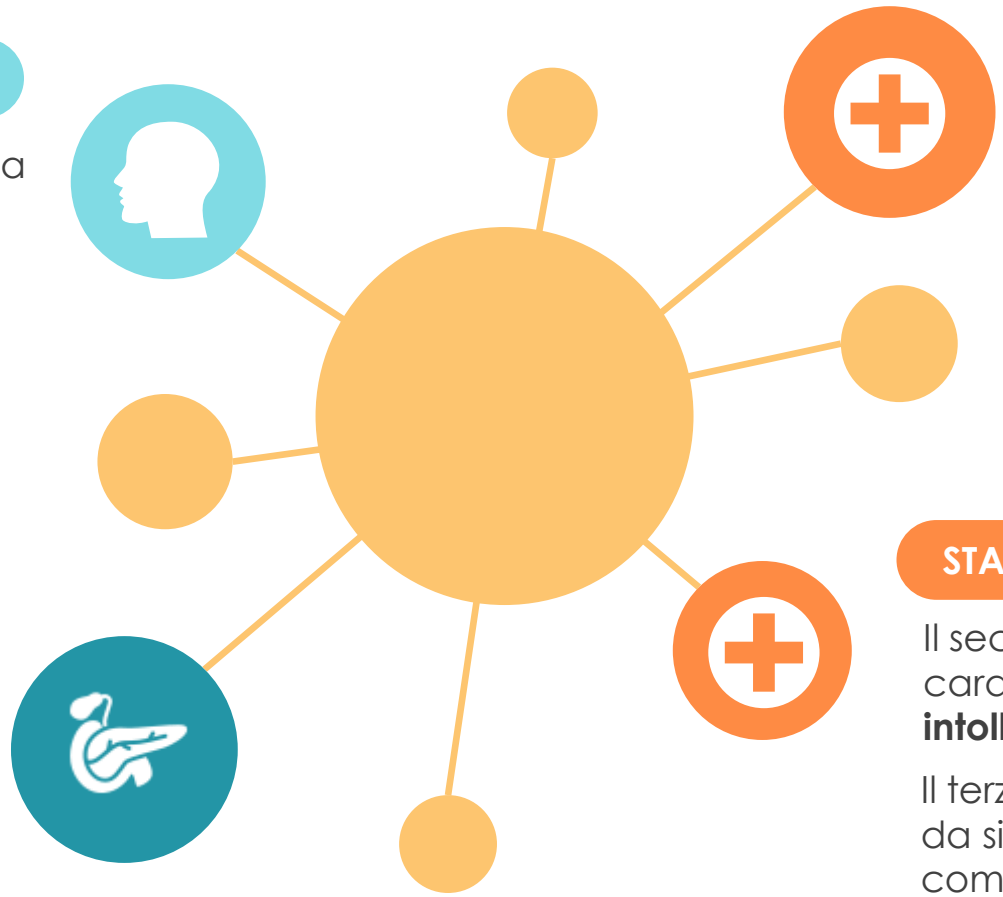
# INTRODUZIONE

## DIABETE DI TIPO 1

Il **diabete di tipo 1 (T1D)** è una malattia autoimmune caratterizzata da distruzione delle cellule β del pancreas

## IL PANCREAS

La componente endocrina è formata dalle **isole di Langherans,** dove troviamo le **cellule β**



## STADI PATOGENESI: I

Primo stadio abbiamo l'**insulite** e di consegenza la comparsa di autoanticorpi anti-isole

## STADI PATOGENESI: II E III

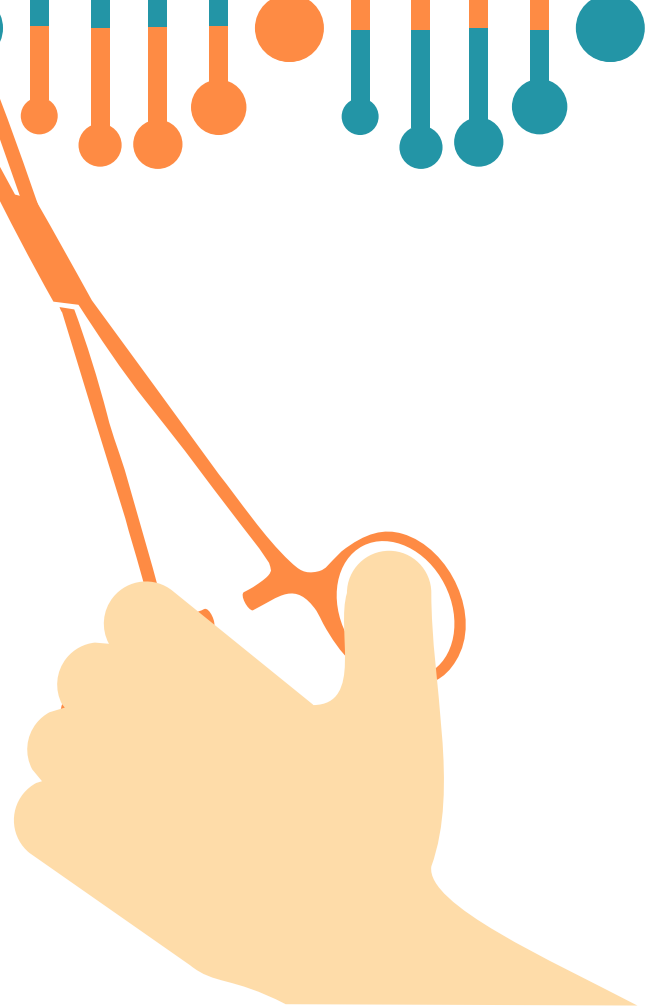
Il secondo stadio è caratterizzato da **disglicemia** o **intolleranza al glucosio**.

Il terzo stadio è caratterizzato da sintomi di i**perglicemia**, come la poliuria.

# PREDISPOSIZIONE GENETICA

- Il rischio di T1D nei fratelli è **15** volte superiore al rischio nella popolazione generale

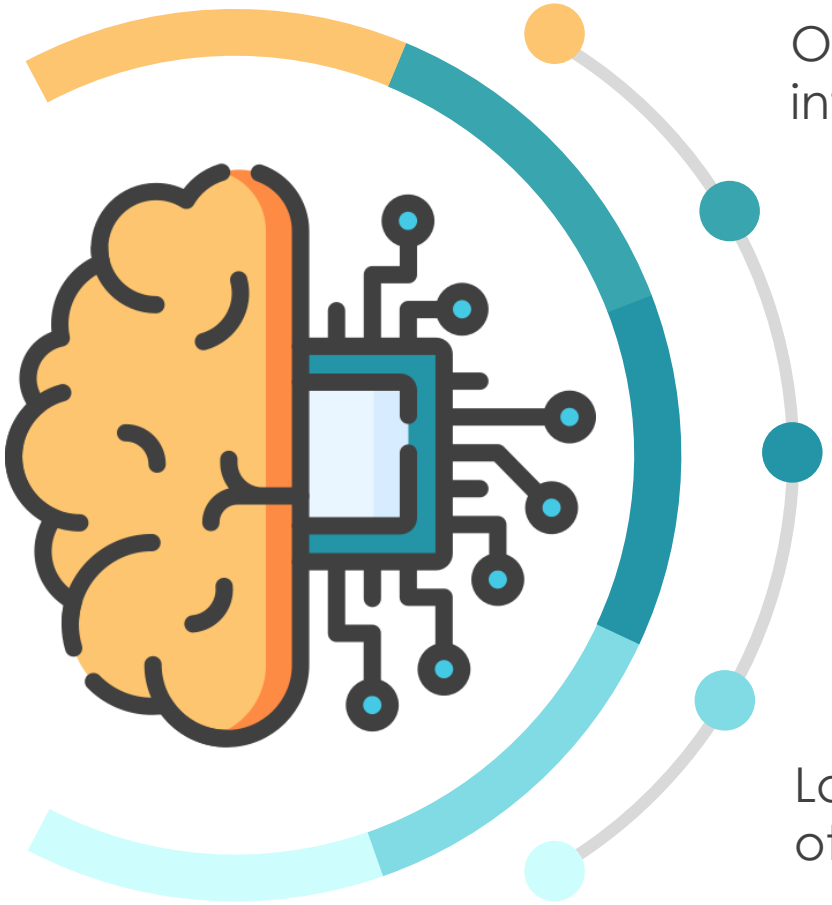
- **SNP - polimorfismi a singolo nucleotide**
  Modifica la struttura e il livello di espressione rendendo il gene unico per quell'individuo



## Geni HLA

## Geni non- HLA

# PERCHÉ IL MACHINE LEARNING?



Offre **strumenti per analizzare** e interpretare dati ad alta dimensionalità

Costruire modelli di classificazione può **facilitare la diagnosi precoce** del T1D

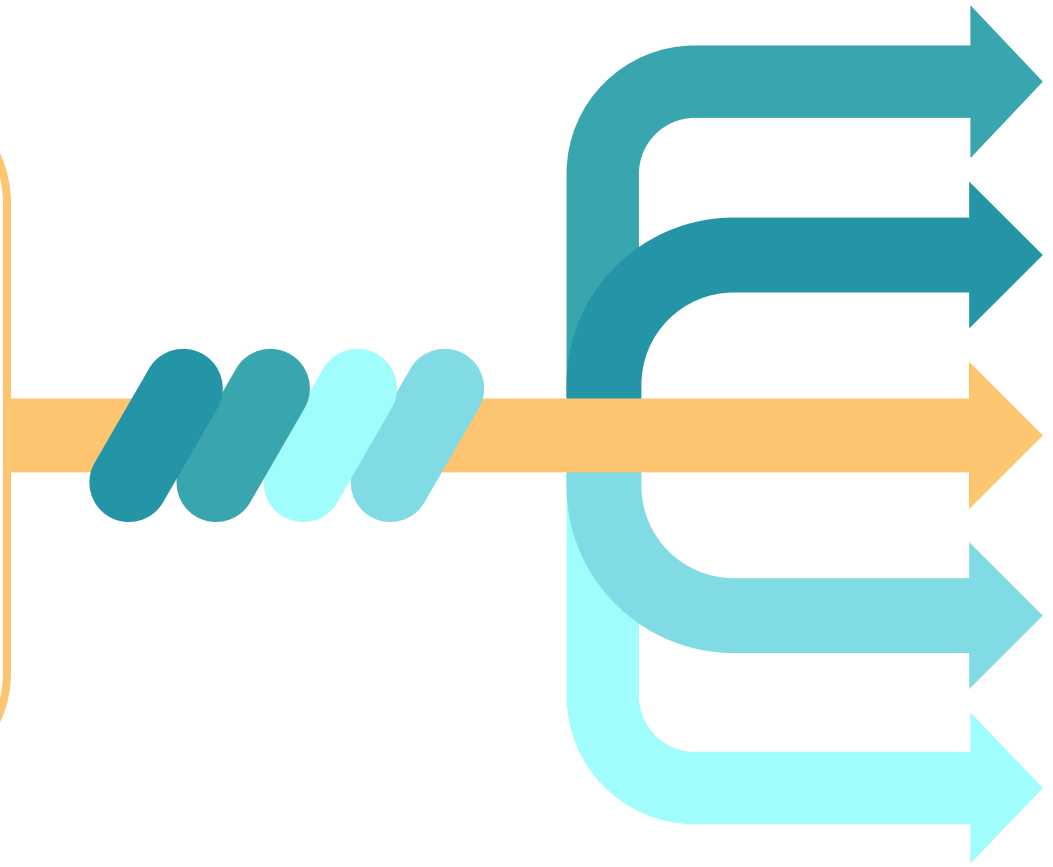
Ma i modelli attuali presentano **limitazioni in termini di interpretabilità e affidabilità**

La comprensione delle decisioni è essenziale per ottenere la **fiducia dei medici e dei ricercatori**

# METODOLOGIA

**OBIETTIVO**

*Fornire agli esperti modelli di machine learning spiegabili basati su dati di espressione genica permettendo di confrontare i risultati con le loro conoscenze*



Analisi e pulizia dei **dati**

Scelta e costruzione dei **modelli**

Analisi dei **risultati** dei modelli

Utilizzo di tecniche di **explainability**

Analisi dei **risultati** delle tecniche utilizzate

# COSTRUZIONE DEL DATASET

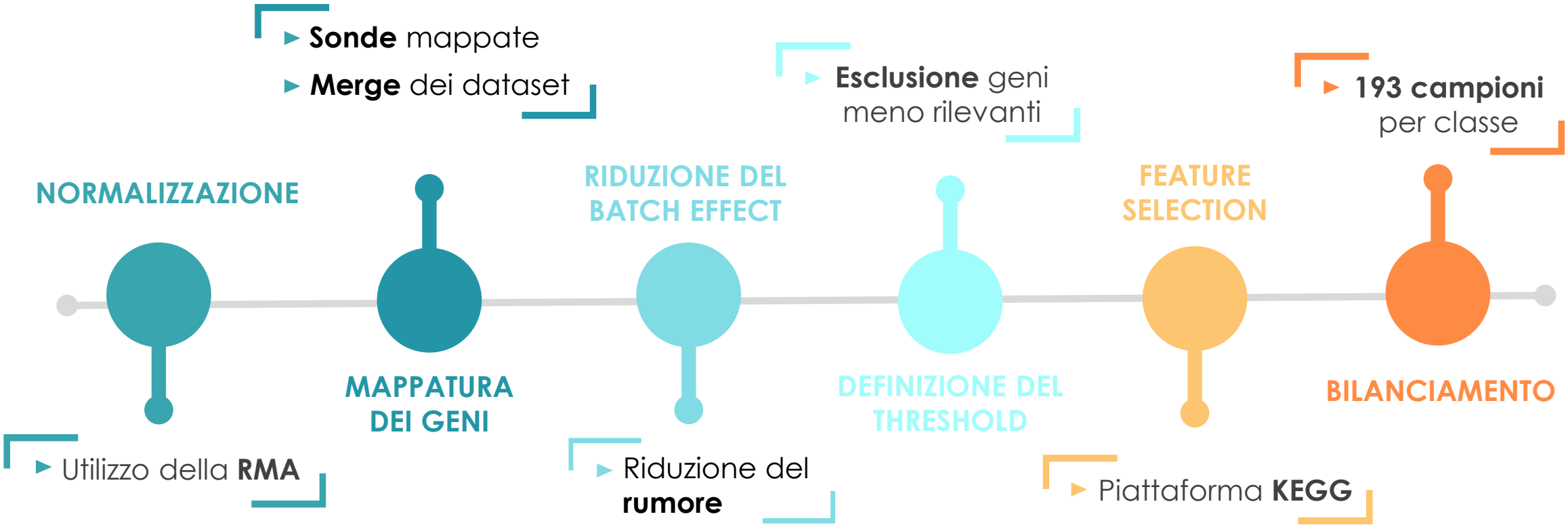
## GSE9006

- **43 pazienti** con diagnosi di **T1D**
- **12 pazienti** con diagnosi di **T2D**
- **24** soggetti **sani**
- Età compresa tra i **2** e i **18 anni**

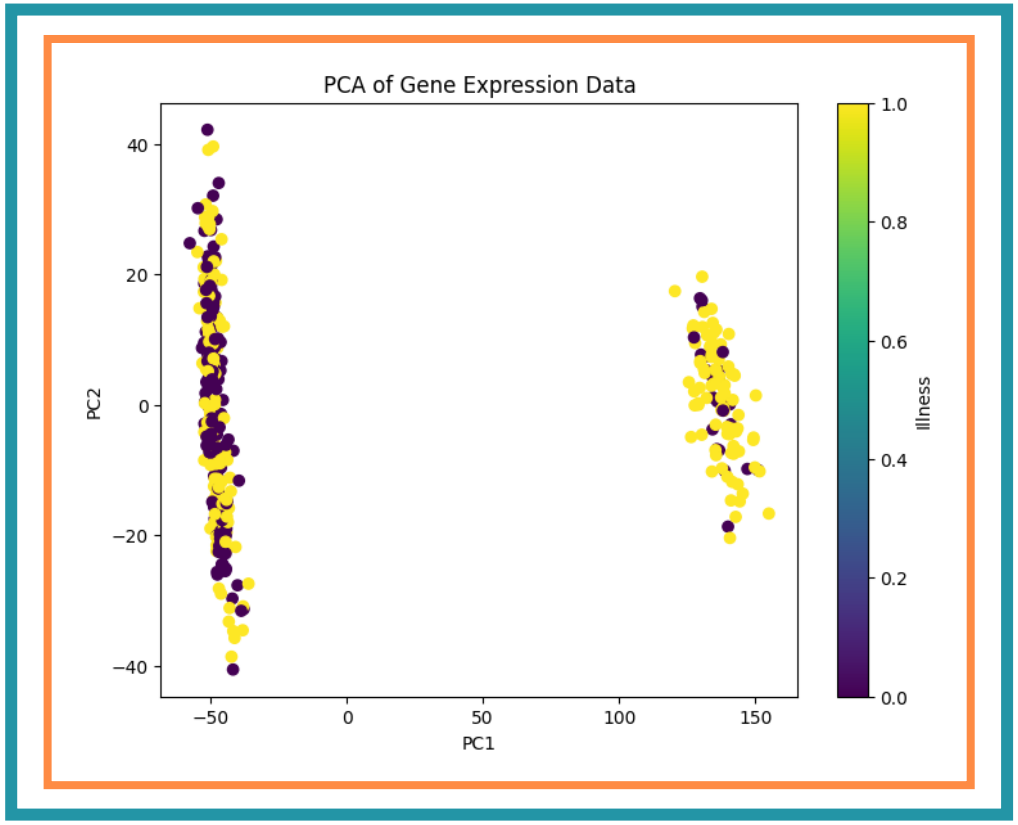
## GSE43488

- **10 pazienti** positivi agli **autoanticorpi**
- **18** pazienti **prediabetici**
- **28 control**, ognuno associato ad un paziente positivo agli autoanticorpi o prediabetico
- Età **infantile** o **adolescenziale**

# PREPROCESSING

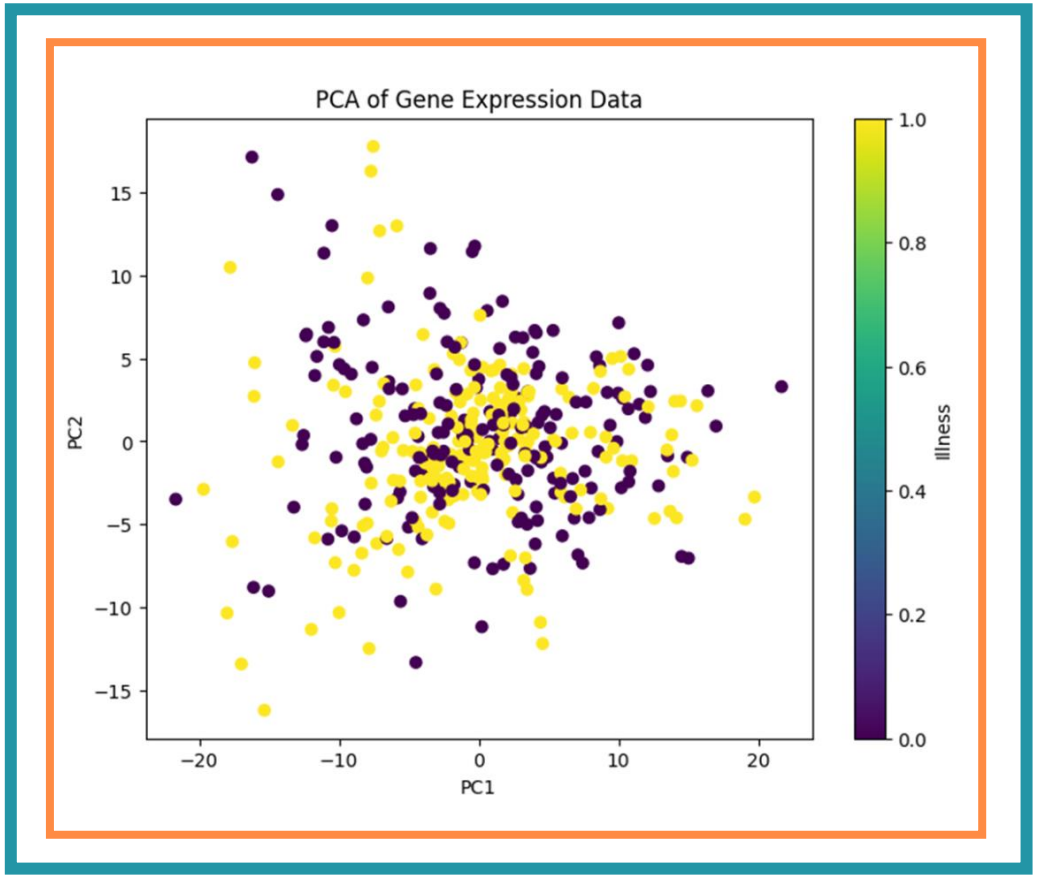


# RIDUZIONE DEL BATCH EFFECT



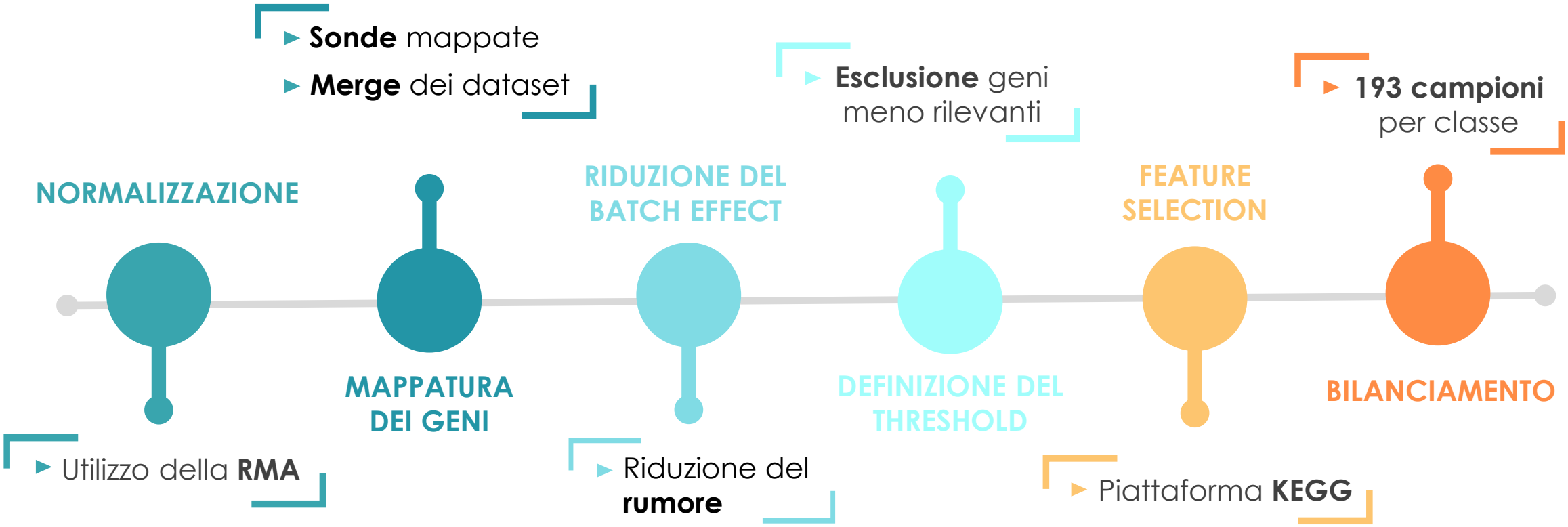

PRIMA

- **PCA (Principal Component Analysis)**
  Tecnica di riduzione della dimensionalità riducendo le variabili a quelle più rilevanti.

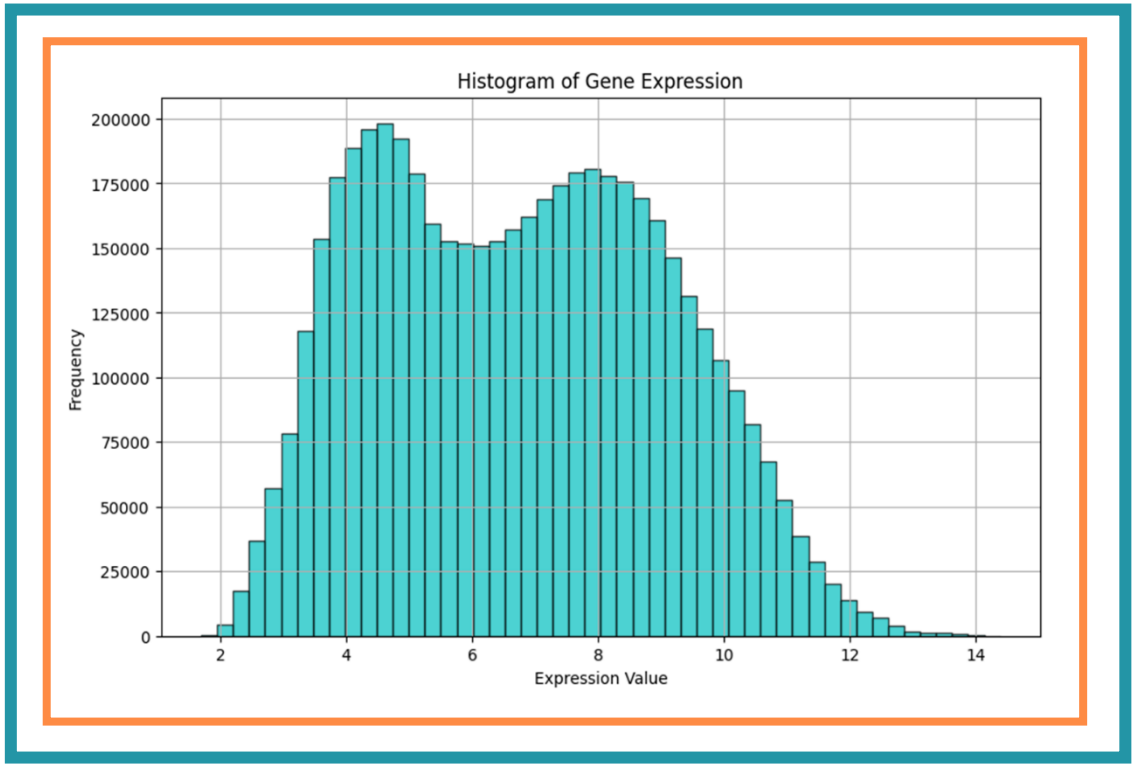



DOPO

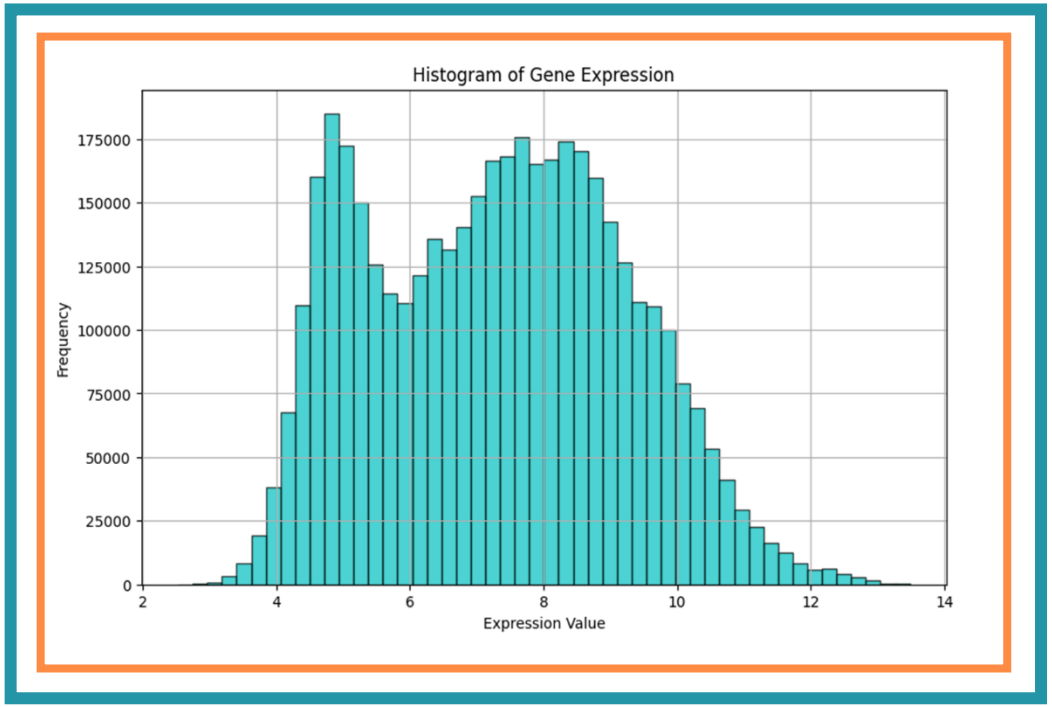
# PREPROCESSING



# DEFINIZIONE DEL THRESHOLD



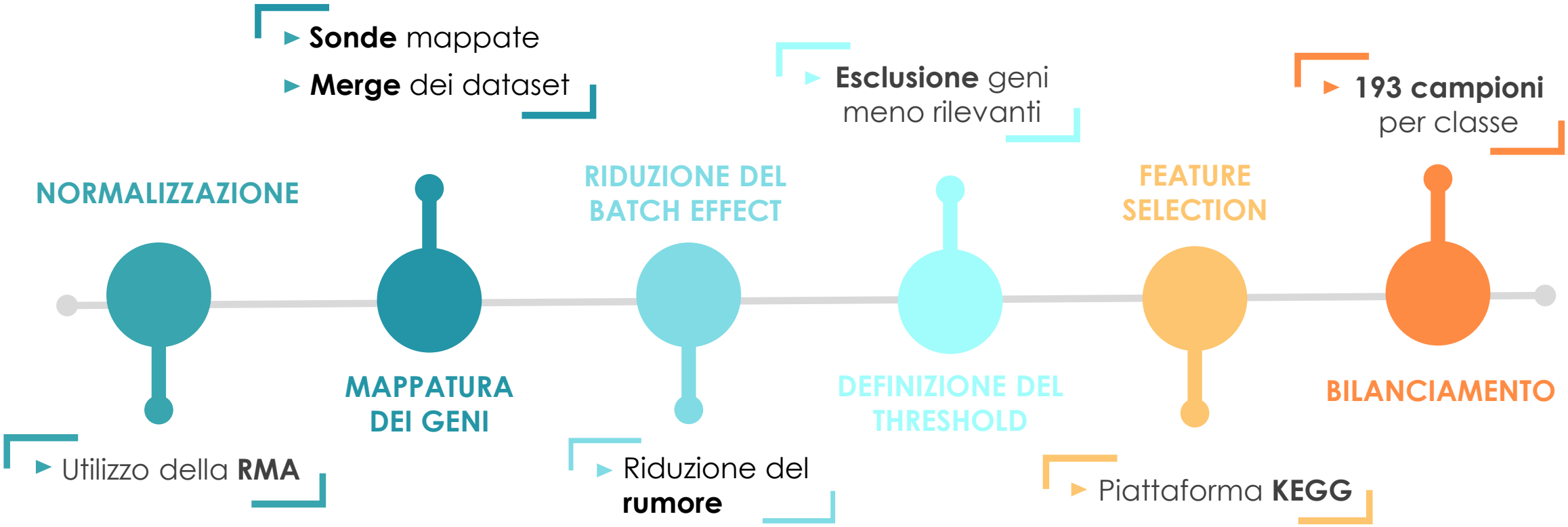

PRIMA



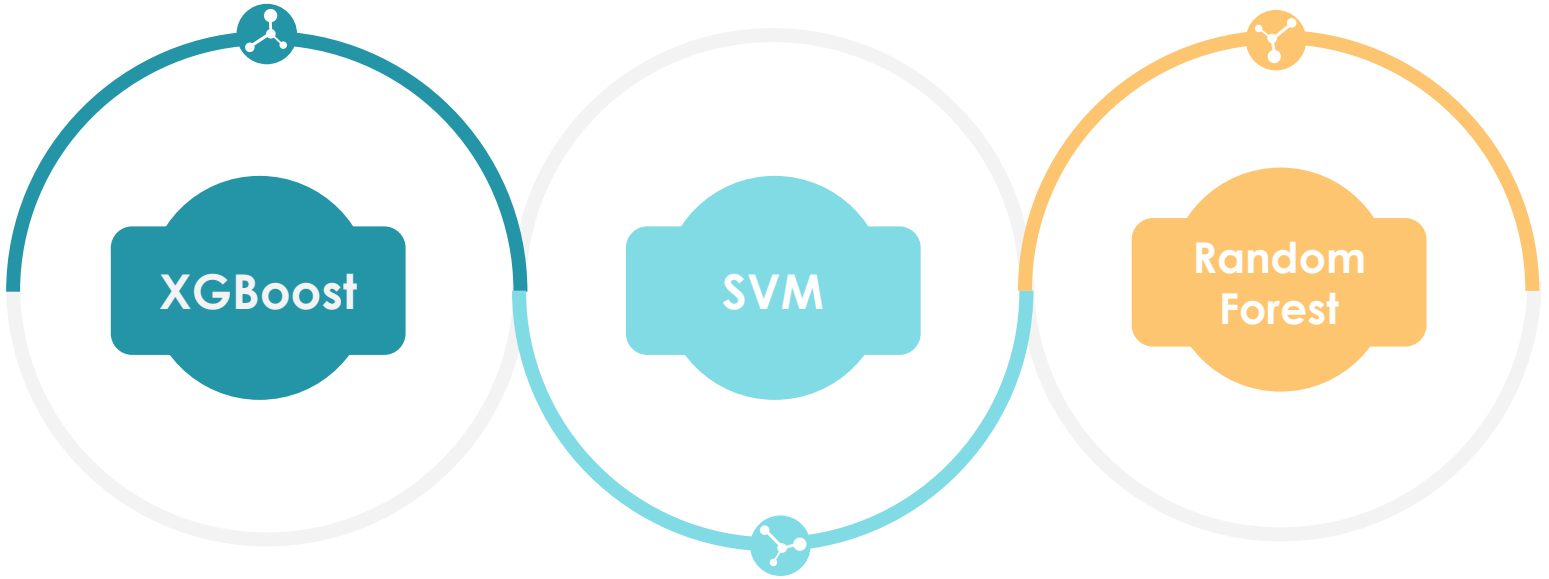

DOPO

- È stata impostata una **soglia pari a 5**.

# PREPROCESSING



# MODELLI DI PREDIZIONE




- Basato sul **gradient boosting**
- Migliora correggendo errori precedenti

- Cerca un **iperpiano** per separare i dati
- Utilizza il **kernel trick**

- Costruisce una collezione di **alberi decisionali**
- **Combina** il risultato

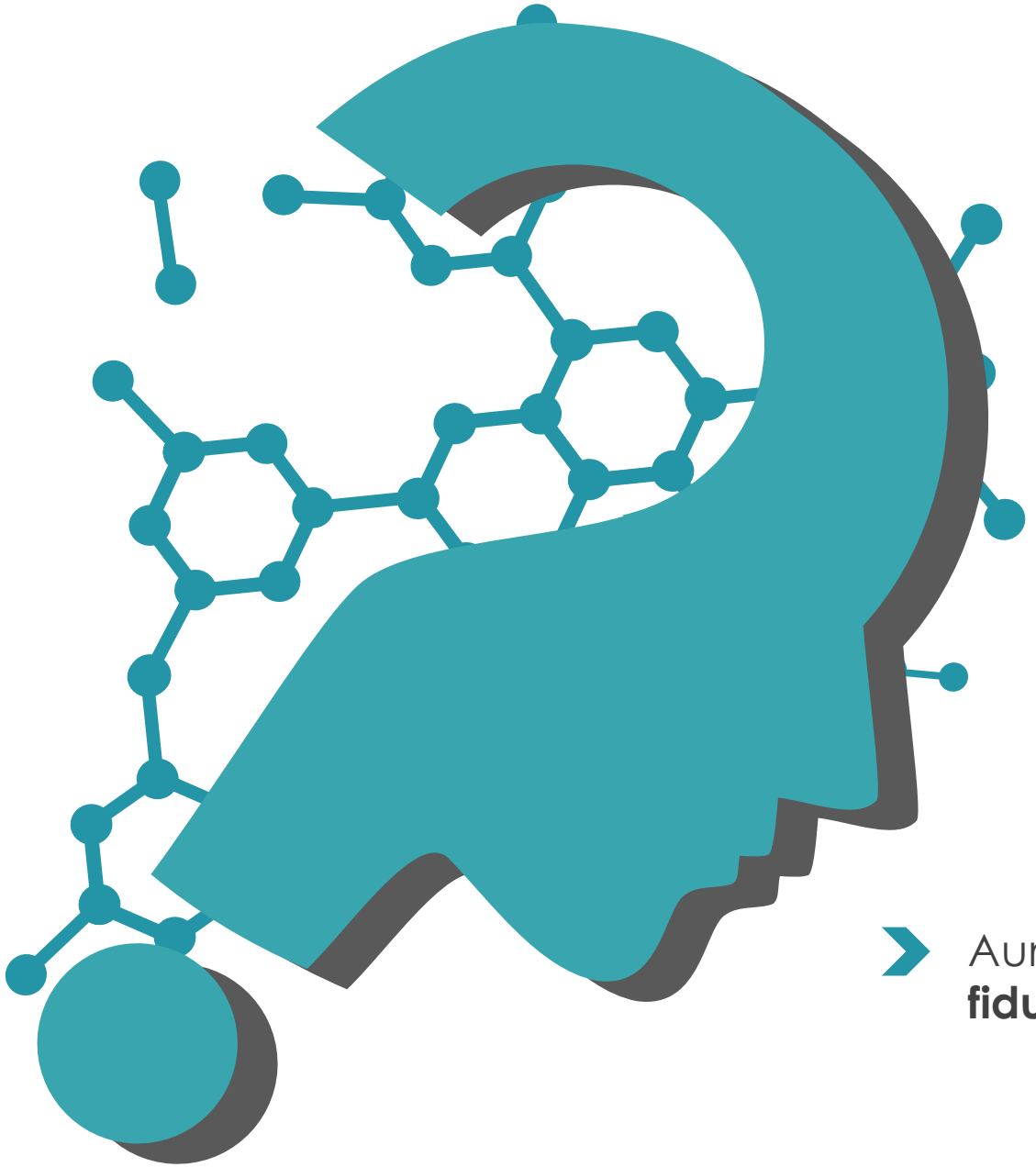
# RISULTATI

## CONSIDERAZIONI

**SVM** risulta essere il **migliore**, nonostante XGBoost e Random Forest siano valide alternative

- Ha **migliori capacità** nel **classificare** i campioni
- Mantiene un giusto **compromesso tra accuracy e recall**, confermato dall'F1
- Riesce a **separare bene** classi positive e negative

| | XGBoost | SVM | Random Forest |
|---|---|---|---|
| Accuracy | **0.794872** | **0.807692** | **0.794872** |
| Precision | **0.798007** | **0.812834** | **0.798007** |
| Recall | **0.794872** | **0.807692** | **0.794872** |
| F1 Score | **0.7943309** | **0.806899** | **0.794331** |
| AUC-ROC | **0.871794** | **0.876397** | **0.873767** |

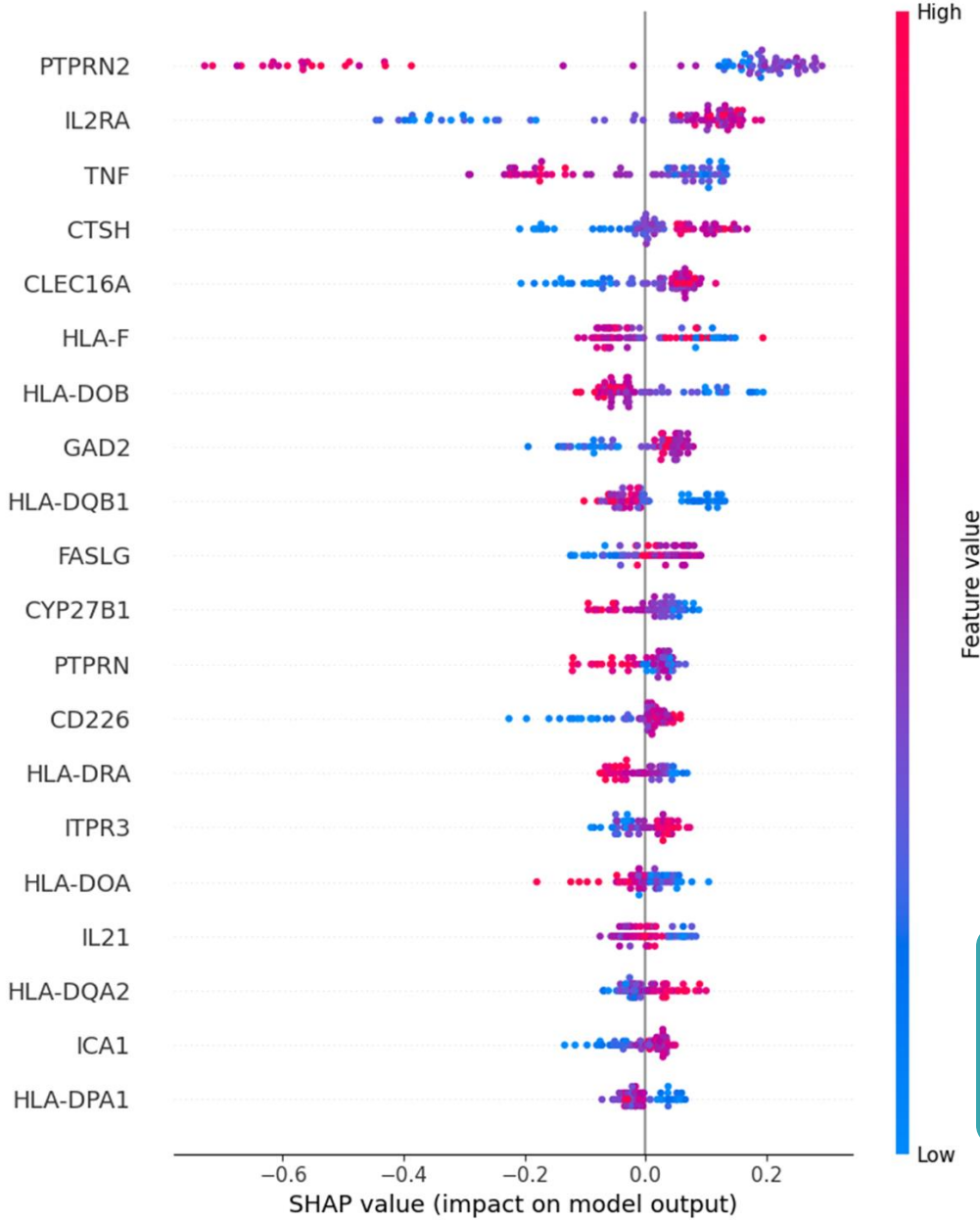


# CHE COS'È L'EXPLAINABILITY?

Possibilità di comprendere e interpretare le decisioni prese da un modello di machine learning

# PERCHÉ EXPLAINABILITY?

- Migliorare la **trasparenza**
- Aumentare la **fiducia** degli utenti finali
- Confronti con le **conoscenze degli esperti**
- Facilita l'**integrazione** nella pratica clinica




# EXPLAINABILITY XGBOOST

## Principali feature di Anchor

- PTPRN2
- IL2RA
- TFN
- CTSH

## REGOLA ANCHOR

['PTPRN2 <= 6.68', 'TNF <= 7.46', 'CTSH > 9.96']

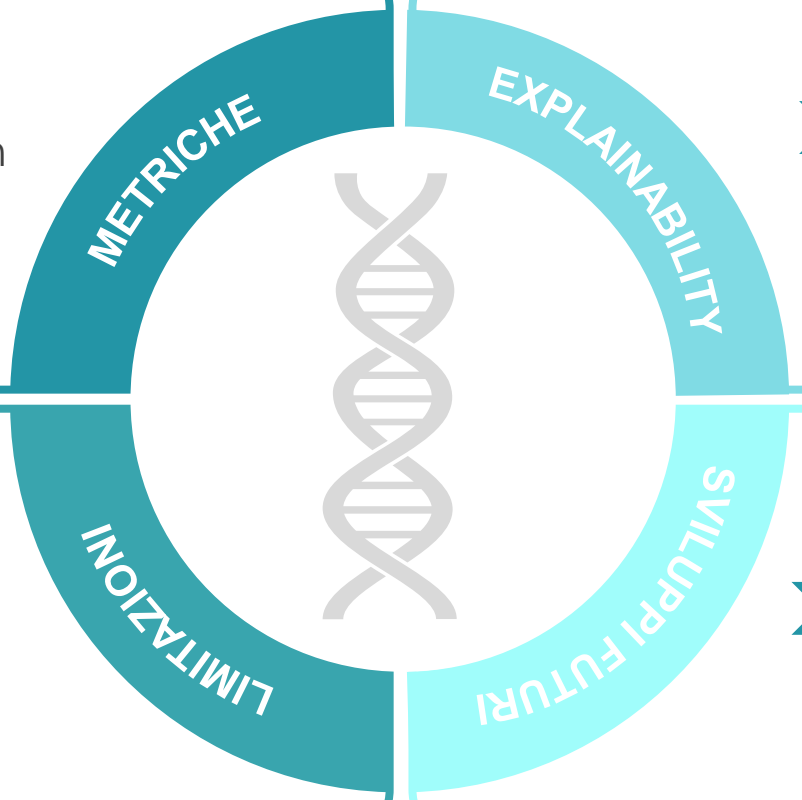
# CONCLUSIONI

- SVM ha raggiunto le **migliori prestazioni** in termini di **metriche**

- XGBoost ha raggiunto le **migliori prestazione** per l'**explainability**

- La **disponibilità di dati pubblici** nell'età dell'insorgenza è altamente limitata

- Integrazione di **dati temporali** per sviluppare un'analisi di serie temporali



# GRAZIE PER L'ATTENZIONE!