



Corso di Data Mining

# Analisi e Classificazione di Dataset mediante Tecniche di Machine Learning

---

Anna Chiara Mameli - Simone Rubiu - Michele Cocco

# Indice

## Analisi esplorativa completa dei 4 dataset

Distribuzione delle feature attraverso istogrammi e boxplot,  
bilanciamento classi e matrice di correlazione

## Preprocessing e preparazione dei dati

Panoramica completa di tutta la pipeline di preprocessing

## Classificazione

Applicazione di alcuni algoritmi di machine learning (metodi ensable, boosting, bagging e geometrici)

## Riduzione delle dimensionalità

Uso di algoritmi di selezione ed estrazione delle feature per ridurre i dati a due dimensioni e confronto tra i due approcci

## Risultati

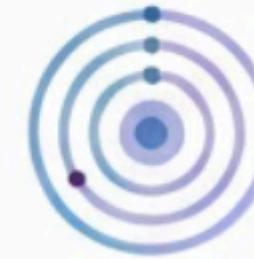
Valutazione dei risultati ottenuti, individuazione dell'algoritmo migliore e confronto finale

# Introduzione ai dataset



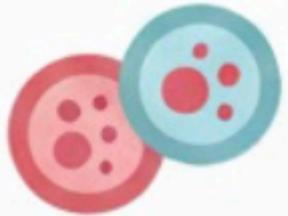
## Banknote Authentication

Immagini in scala di grigi per rilevare contraffazioni.



## Ionosphere

Dati radar per l'analisi della struttura atmosferica.



## Breast Cancer Wisconsin

Immagini di aspirati per diagnosi di malignità.



## Seismic Bumps

Rilevazioni minerarie per predire eventi sismici

# Analisi quantitativa dei Dataset

Dataset	Istanze	Feature	Classe Minoritaria
<b>Banknote</b>	1372	4	44.5%
<b>Ionosphere</b>	351	34	35.9%
<b>Cancer</b>	569	30	37.3%
<b>Seismic</b>	2584	18	<b>6.6%</b>



**Seismic Bumps: forte sbilanciamento (93.4% vs 6.6%)**

1

## Verifica Integrità

Nessun valore mancante nei 4 dataset

2

## One-Hot Encoding

Variabili categoriche (Seismic)

3

## Rimozione Varianza Nulla

VarianceThreshold(threshold=0.0)  
Ionosphere 1  
Sismic Bumps 3

4

## Standardizzazione

Z-score (media = 0, std = 1)

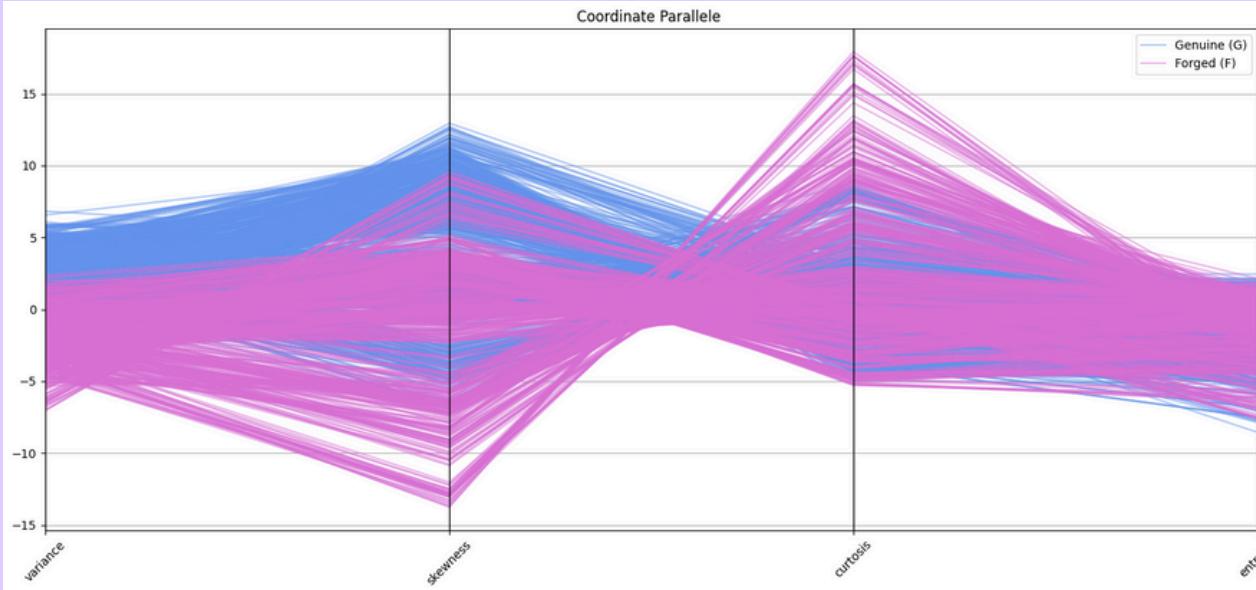
4

## Train/Test Split

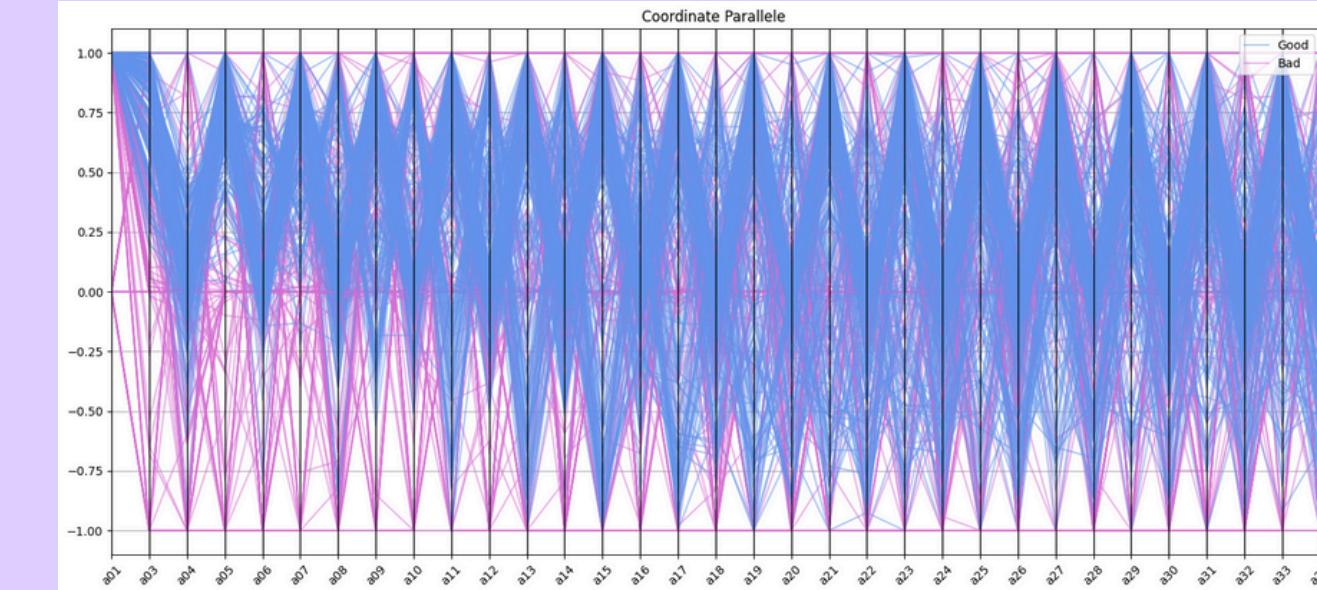
80/20 stratificato per mantenere le proporzioni delle classi,  
random\_state = 42

# Analisi esplorativa

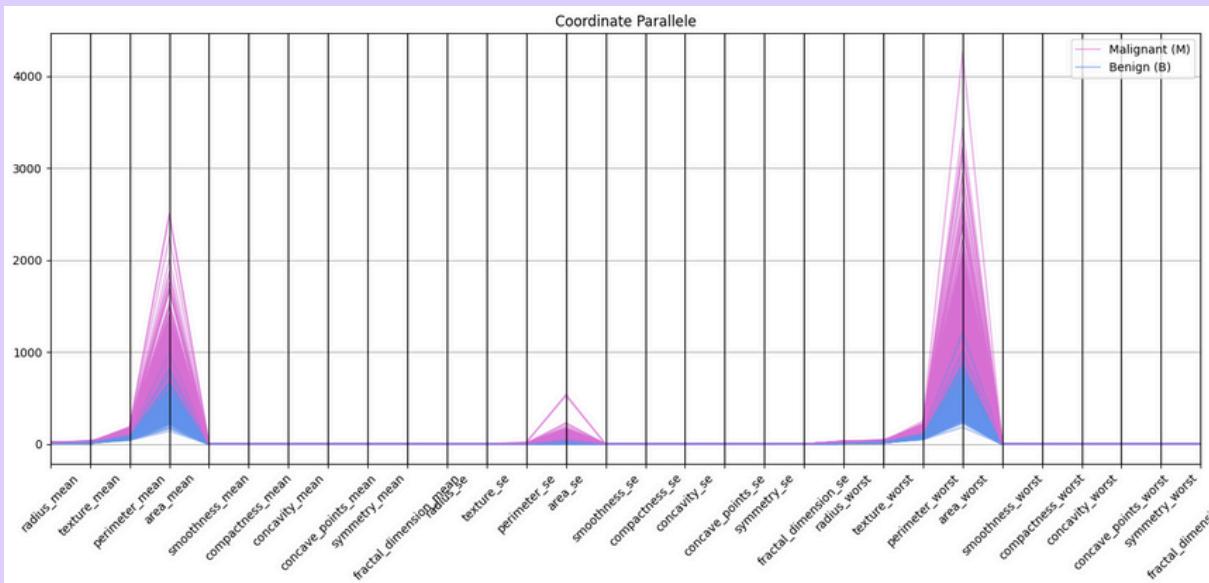
## Coordinate parallele



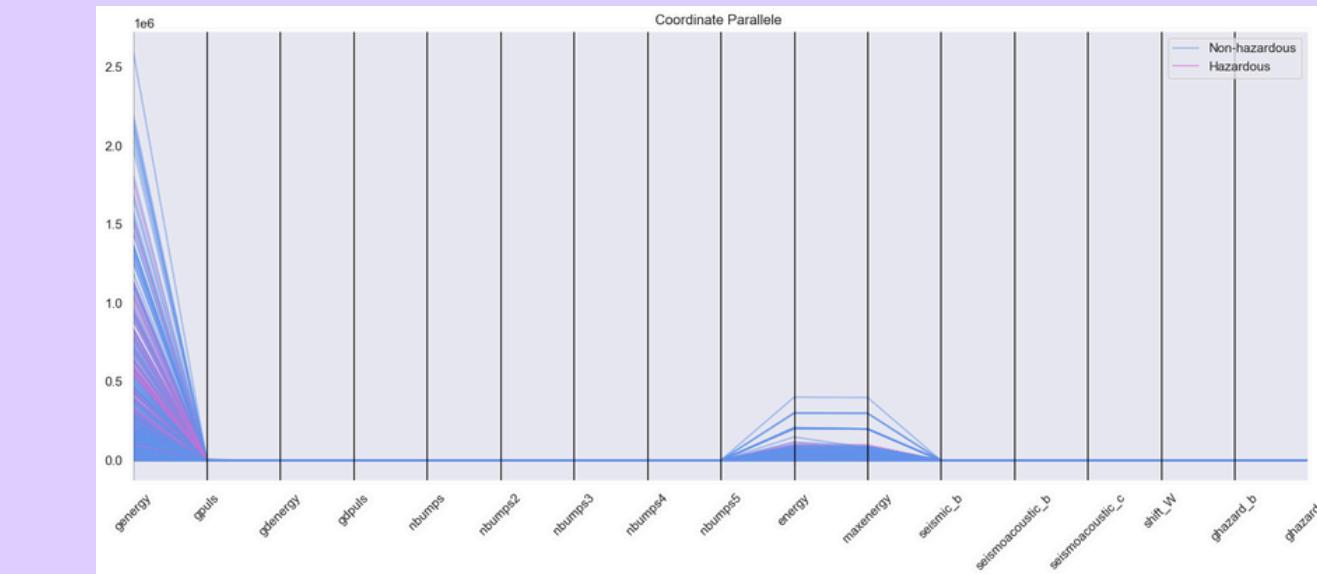
Banknote



Ionosphere



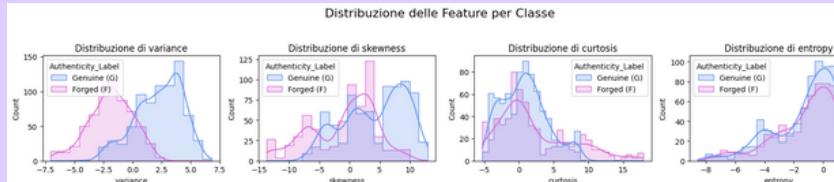
Breast Cancer



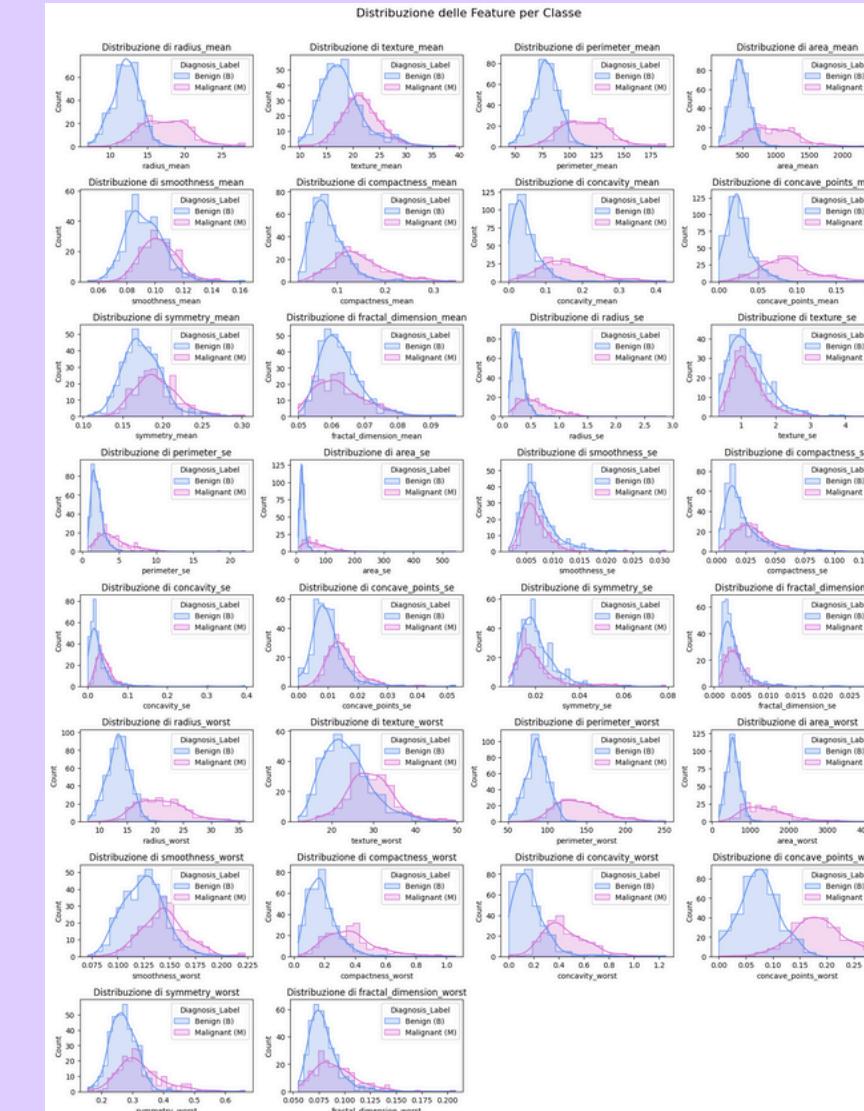
Seismic

# Analisi esplorativa

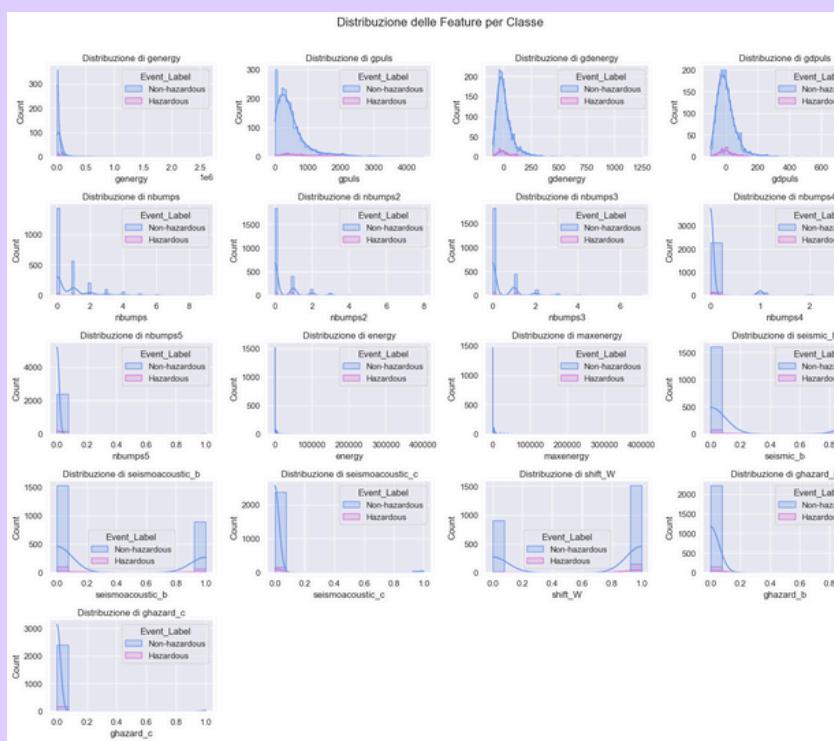
## Iistogrammi feature



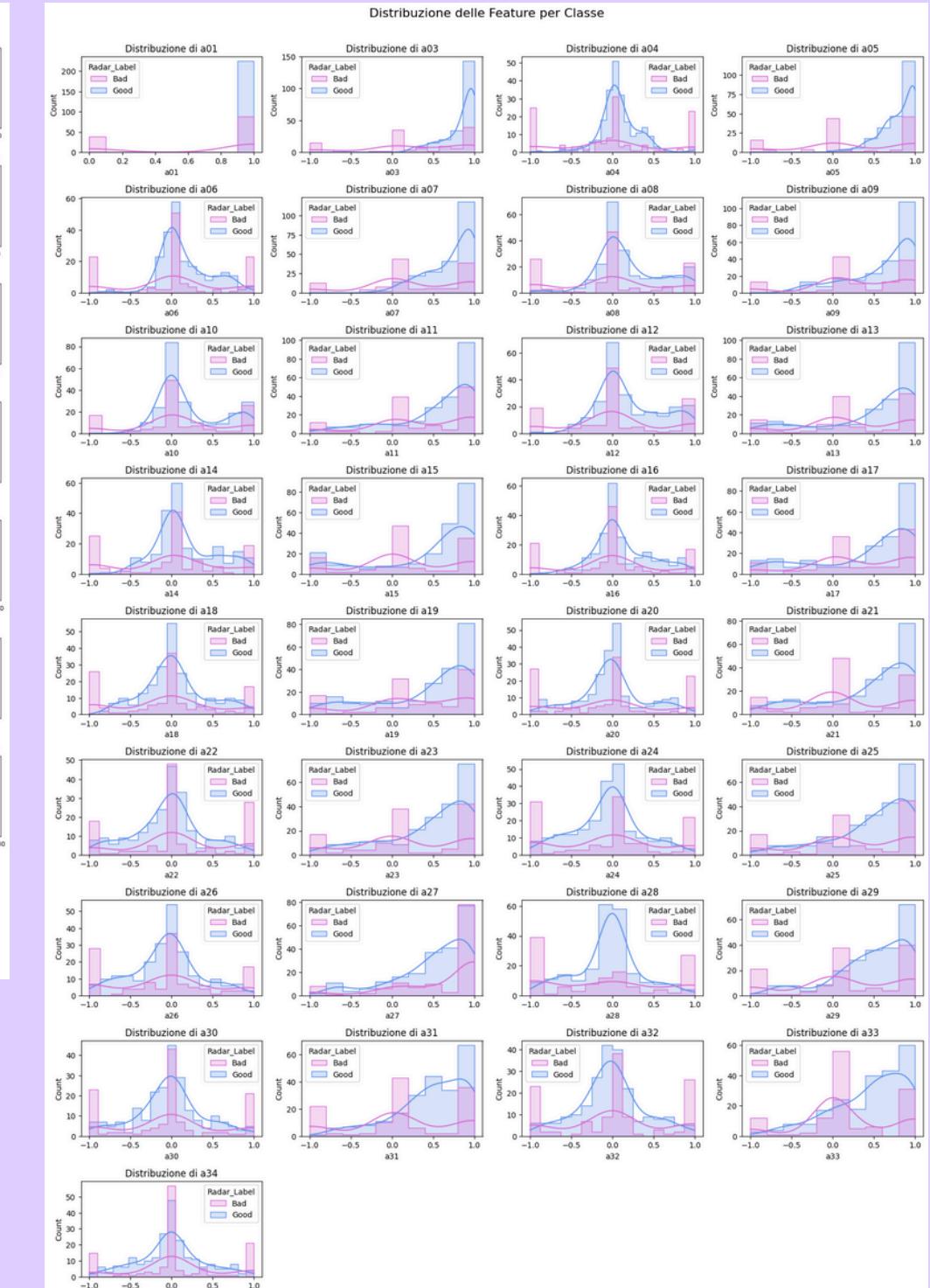
**Banknote**



**Breast Cancer**



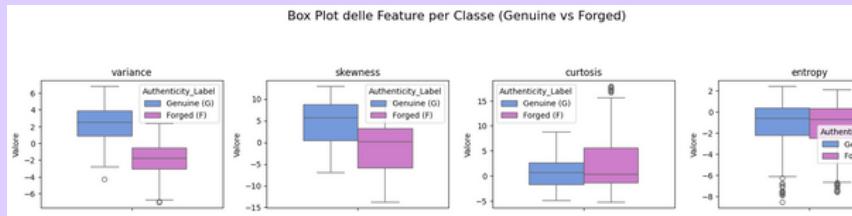
**Seismic**



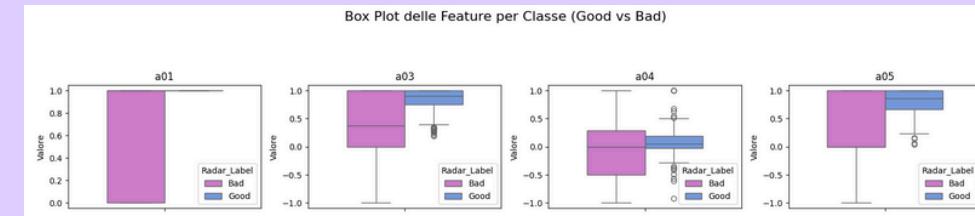
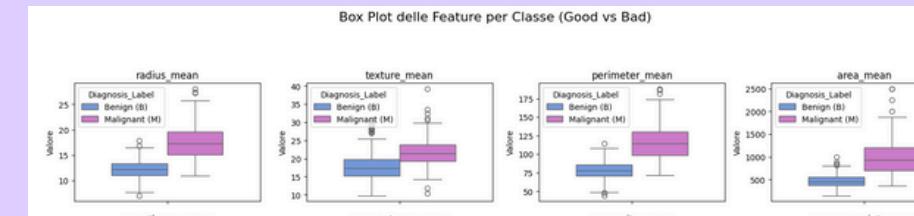
**Ionosphere**

# Analisi esplorativa

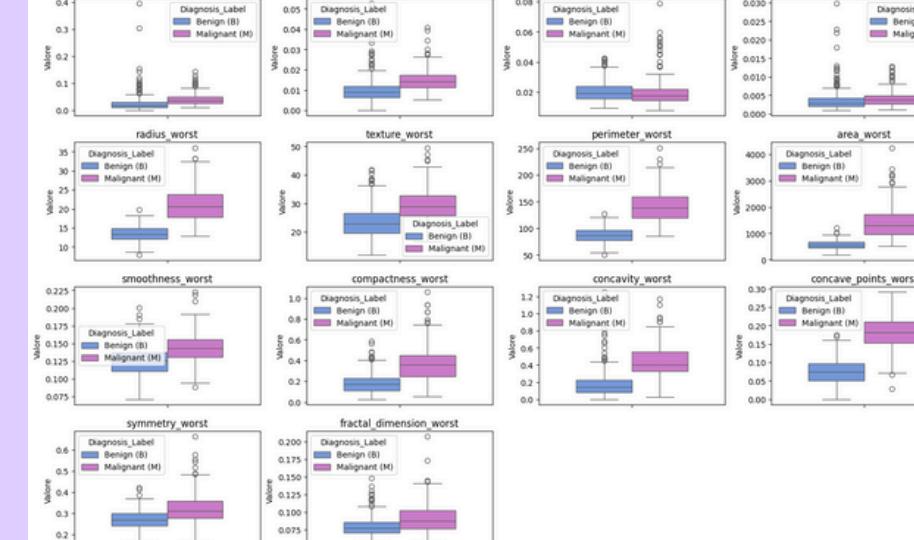
## Box plot



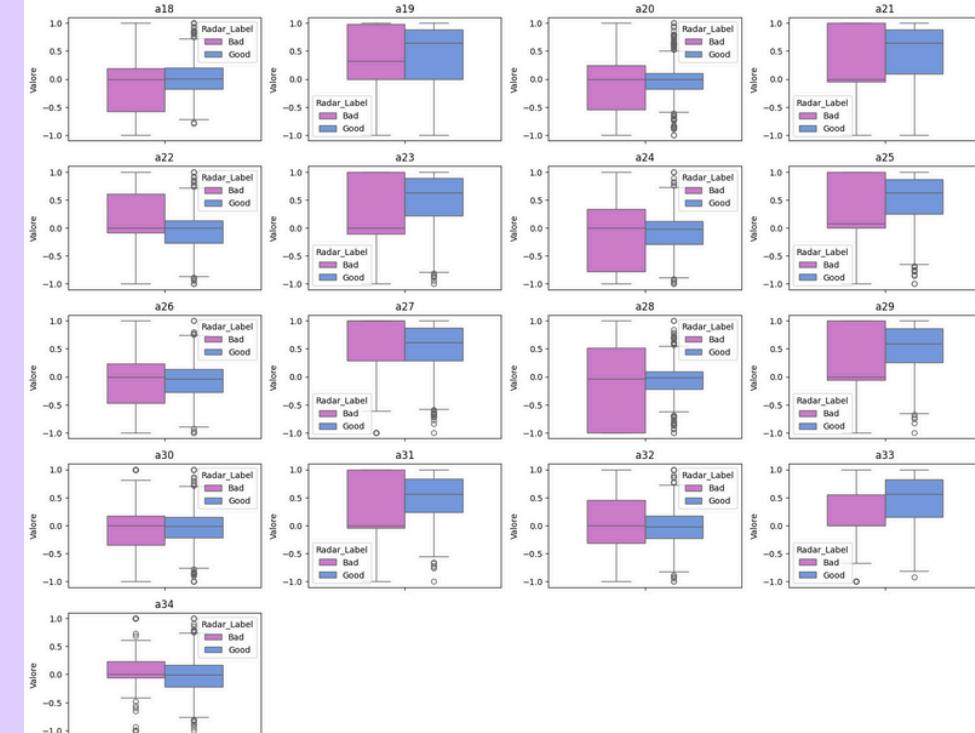
**Banknote**



**Seismic**



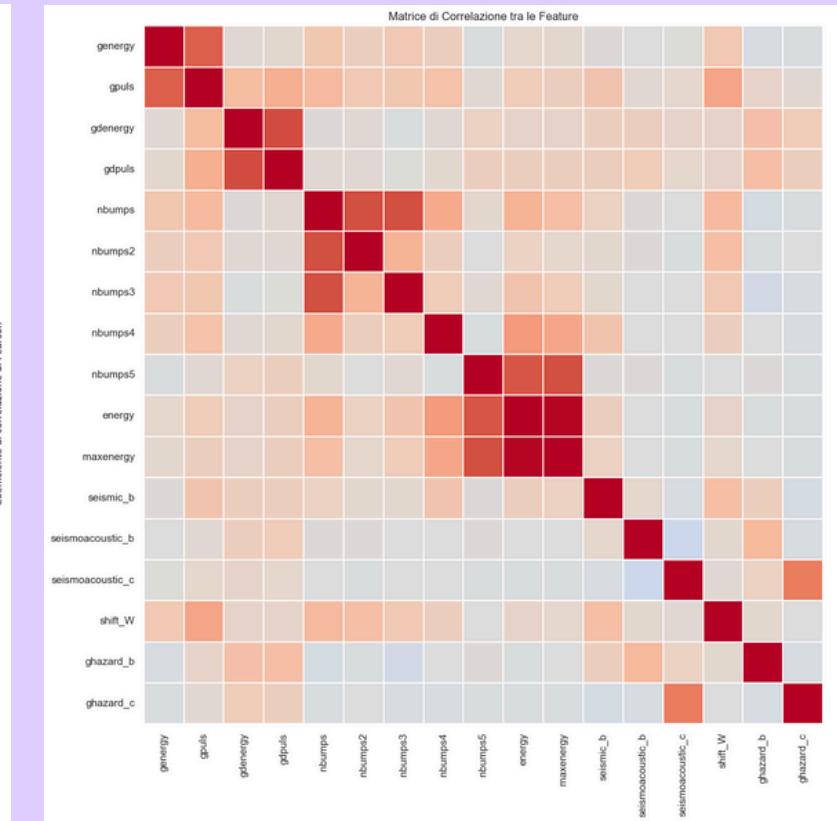
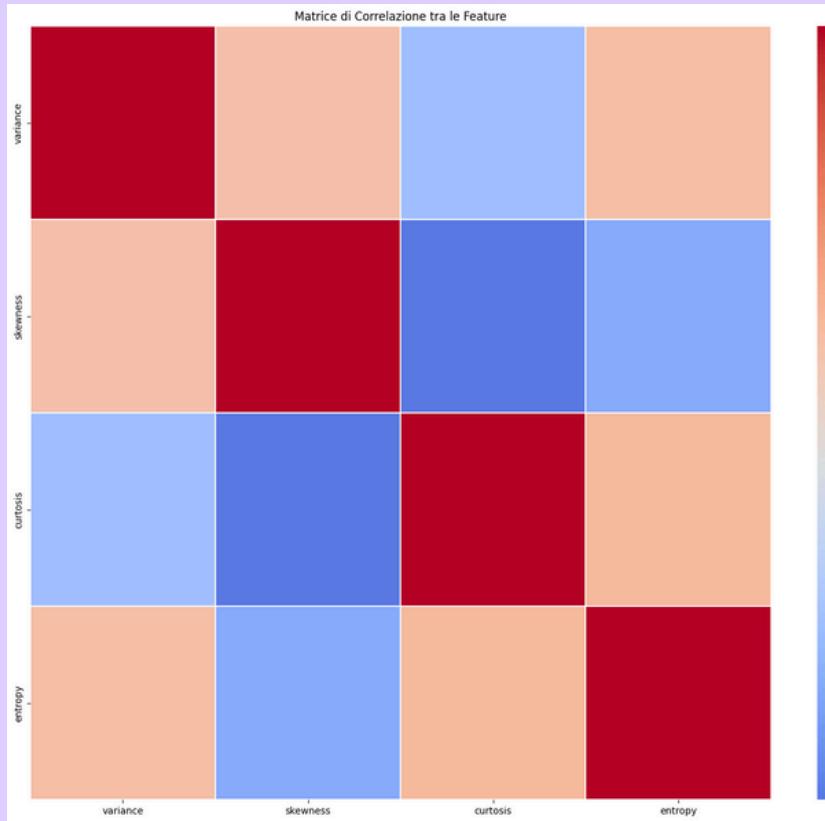
**Breast Cancer**



**Ionosphere**

# Analisi esplorativa

## Matrice di correlazione

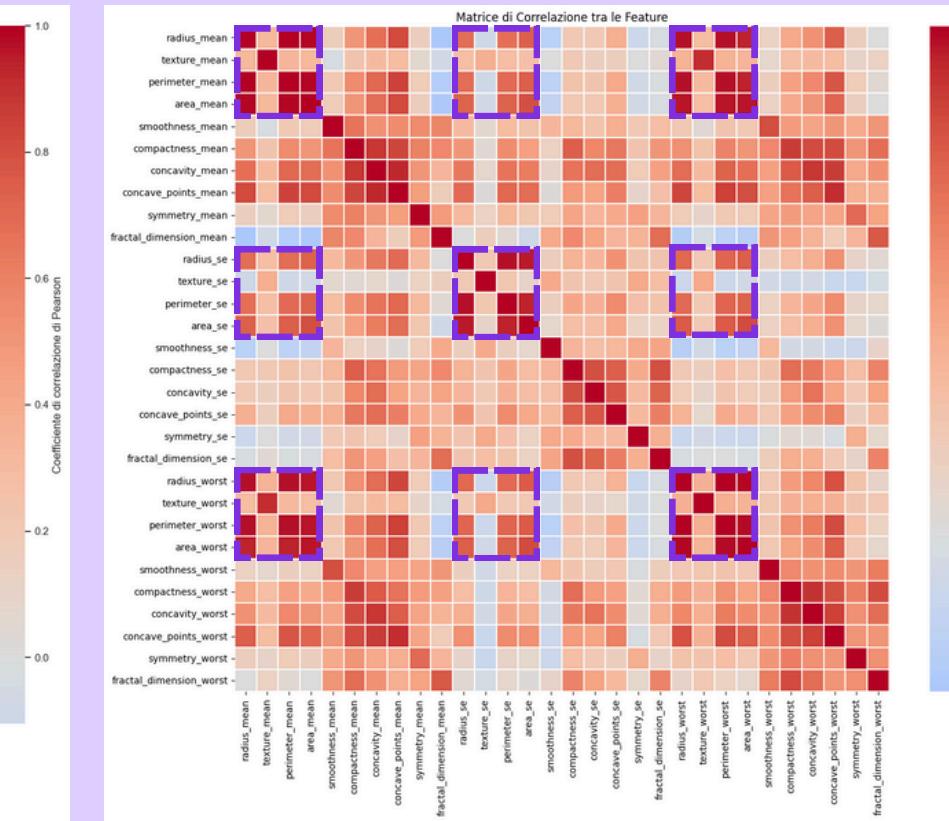


**Banknote**

**Seismic**

**Breast Cancer**

**Ionosphere**



# Metodologia Classificazione

## 7 Algoritmi Testati

Decision Tree

K-Nearest Neighbors

Support Vector Machine

Multi-Layer Perceptron

Random Forest

AdaBoost

XGBoost

## Validazione e Metriche

### 5-Fold Stratified Cross-Validation

#### Metriche:

- Accuracy, Precision, Recall
- F1 score (metrica principale)
- AUC-ROC

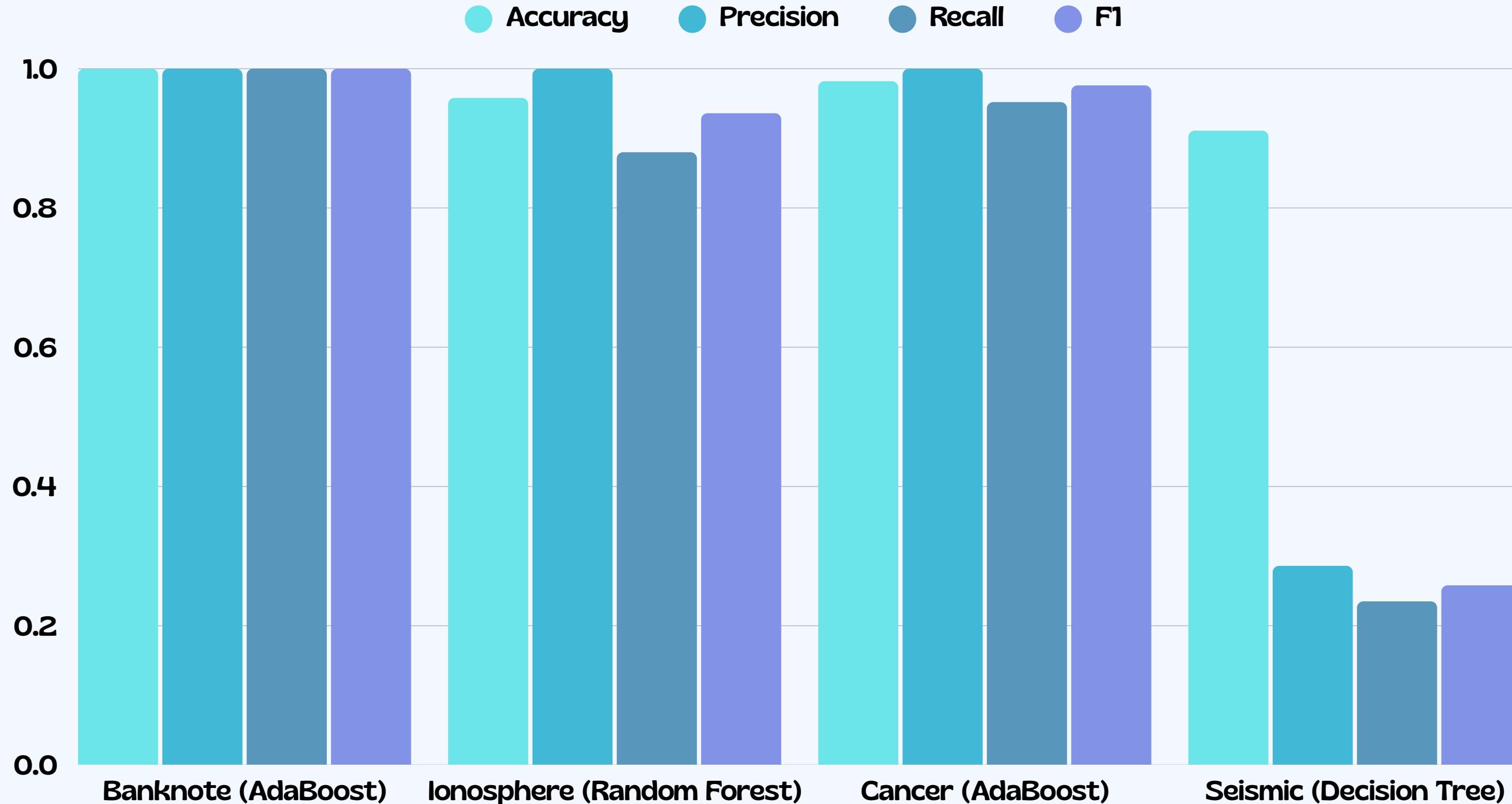
#### Gestione sbilanciamento:

- class\_weight = 'balanced'
- scale\_pos\_weight (XGBoost)

Iperparametri default Scikit-learn

Riproducibilità: random\_state = 42

# Risultati: F1-Score per Dataset



## Osservazioni

### AdaBoost

Vincitore su 2/4 dataset

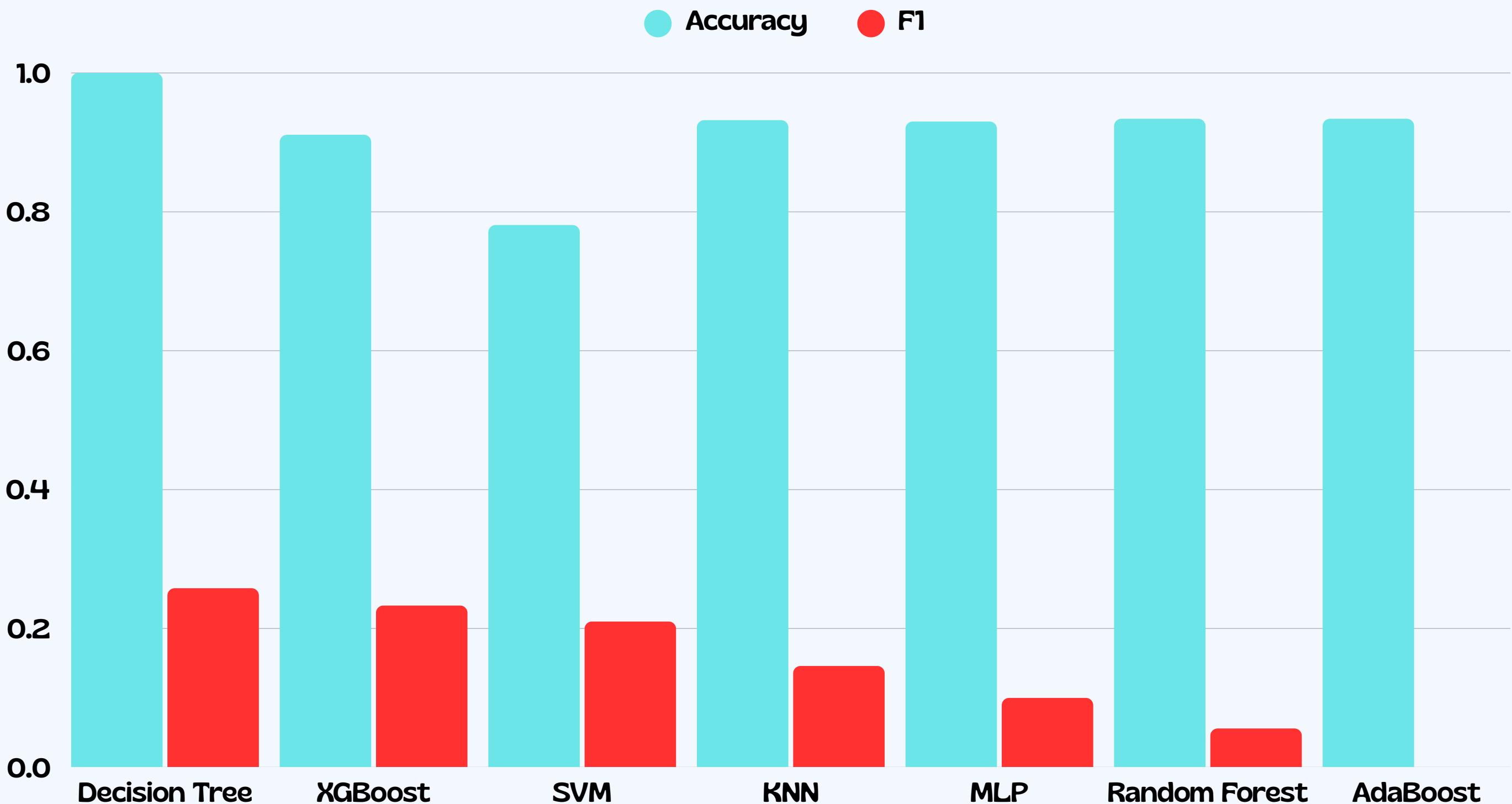
**Ensemble methods** i più scelti.

### Seismic Bumps

F1 = 0.258

Sbilanciamento critico

# Caso Critico: Seismic Bumps



Perchè l'accuracy non è la metrica ottimale?

**Il problema:**

93.4% classe maggioritaria.  
Accuracy alta perchè predice tutto “non-hazardous”.

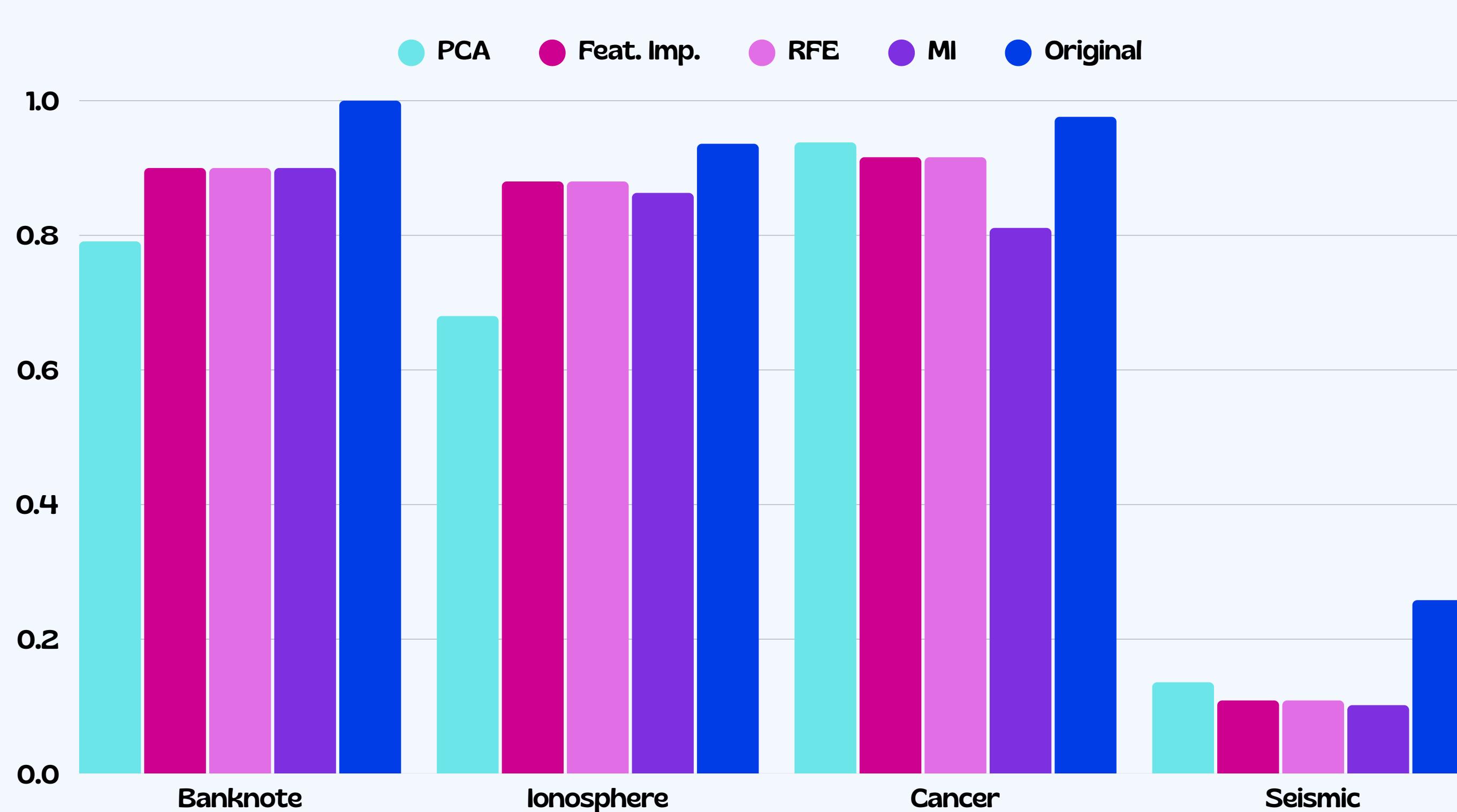
**La soluzione:**

F1-Score bilancia Precision e Recall



L'Accuracy può essere fuorviante con classi sbilanciate!

# Riduzione Dimensionale: Confronto Tecniche



## Risultato

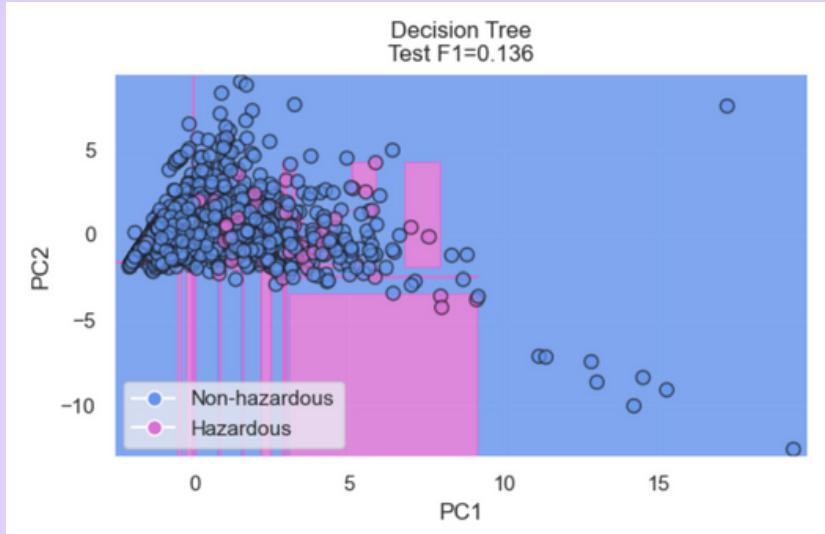
**PCA & Collinearità (Cancer/Seismic):** Le matrici mostrano forti blocchi di correlazione (rosso intenso). La PCA eccelle qui perché sintetizza variabili ridondanti in componenti più informative.

**Feature Selection & Indipendenza (Banknote):** La matrice "fredda" indica variabili indipendenti. In questo caso, mantenere le feature originali (RFE/Feat. Imp.) è meglio che mescolarle, poiché ogni variabile ha un valore predittivo unico.

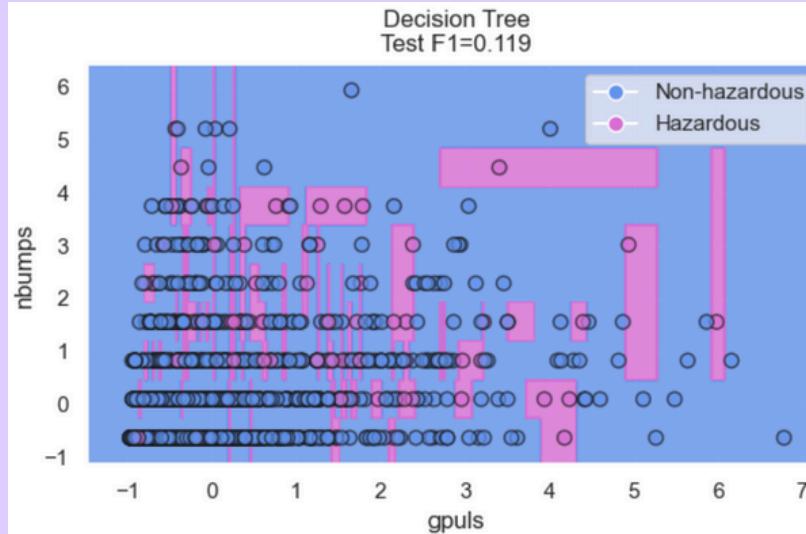
**Efficienza della Riduzione:** Il passaggio a sole 2 feature mantiene l'accuratezza vicina all'originale (calo minimo). Questo conferma che la riduzione ha eliminato il rumore evidenziato nelle zone a bassa correlazione.

# Riduzione Dimensionale: Boundary plot

Caso Peggiose: Seismic bumps

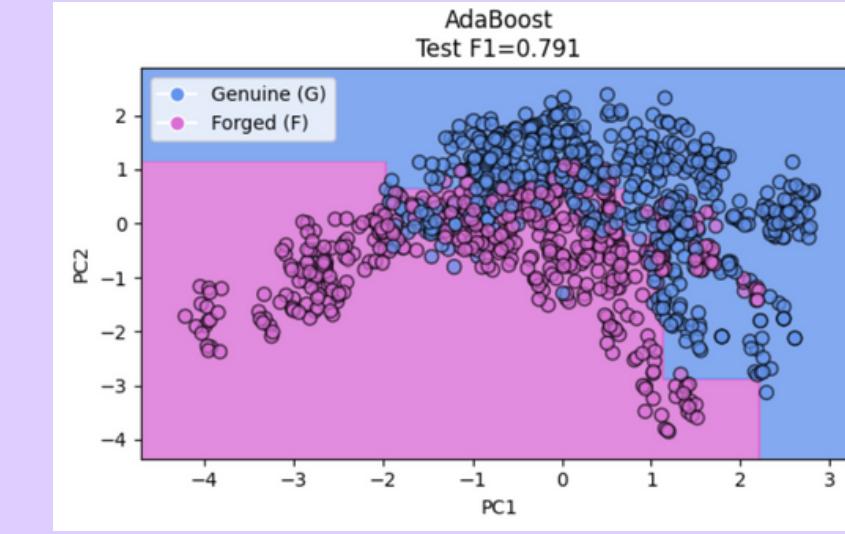


PCA

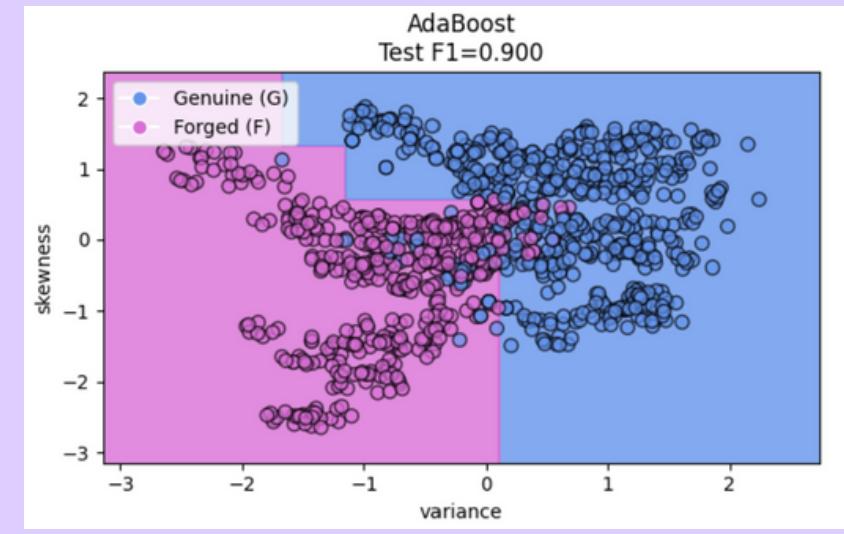


Feature Importance

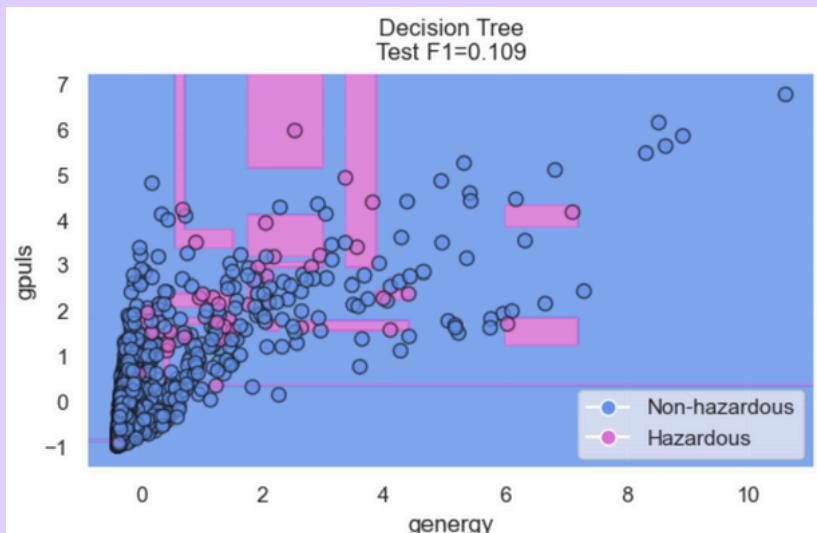
Caso Migliore: Banknote Authentication



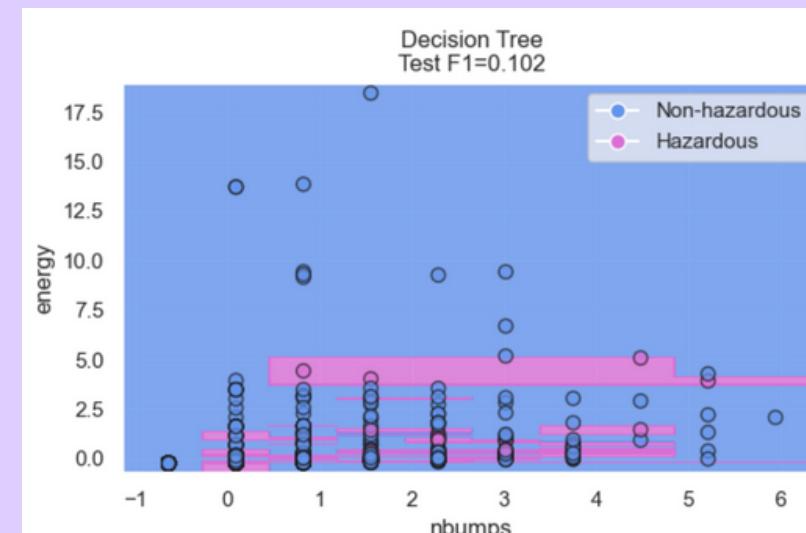
PCA



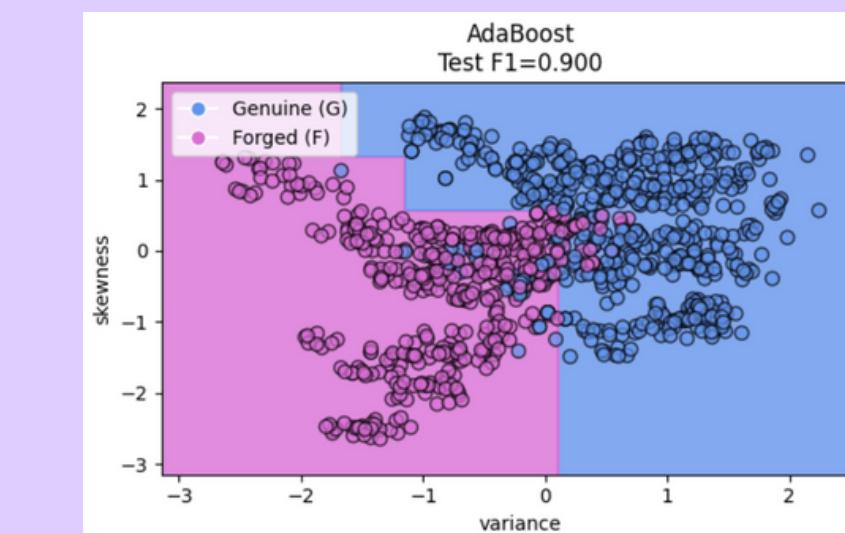
Feature Importance



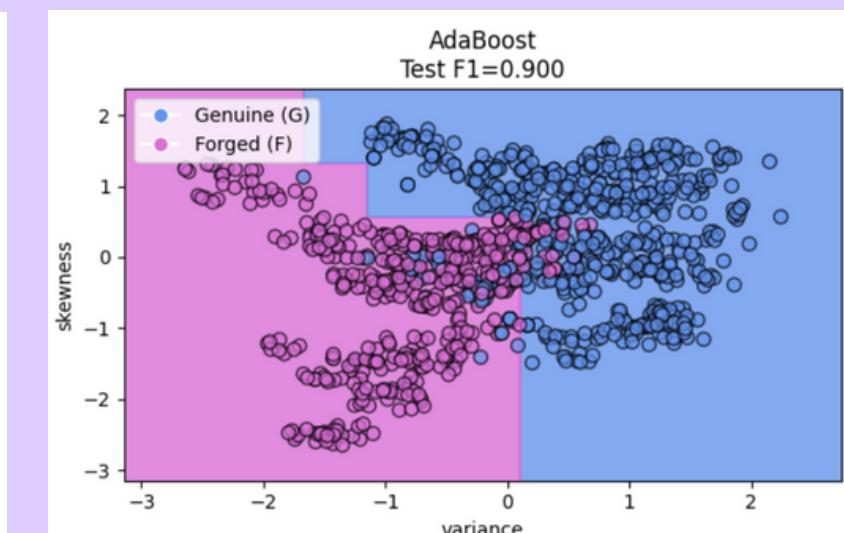
RFE



Mutual Information



RFE



Mutual Information

# PCA vs Feature Selection: Quando Usarle?

## PCA

L'informazione utile è distribuita e si manifesta come varianza globale.

### Quando usarla

- Feature altamente correlate
- Nessuna singola feature domina
- La separazione emerge da combinazioni lineari
- La varianza spiega bene la struttura dei dati

### Cosa fa

- Proietta i dati lungo le direzioni di massima varianza
- Comprime l'informazione ridondante
- Ignora la label

### Limite

- Alta varianza  $\neq$  alta capacità discriminante

## Feature Selection

L'informazione discriminante è concentrata in poche feature.

### Quando usarla

- Presenza di feature con varianza altamente informativa
- La separazione è guidata da variabili specifiche
- Serve interpretabilità

### Cosa fa

- Mantiene le feature con varianza rilevante per il task
- Riduce rumore e ridondanza
- Preserva il significato fisico delle variabili

### Limite

- Sensibile al rumore e al metodo di selezione
- Può scartare feature deboli ma complementari
- Le interazioni complesse tra feature possono non emergere



# Conclusioni



- Gli algoritmi basati su **alberi di decisione** hanno mostrato una **maggior robustezza**, gestendo meglio le non-linearietà.
- Le tecniche di **Feature Selection** eccellono in presenza di **feature indipendenti** con alto potere predittivo individuale (es. dataset Banknote).
- La **PCA** si conferma la strategia ottimale nei dataset ad **alta collinearità** (es. Cancer, Seismic). La proiezione in uno spazio latente permette di condensare l'informazione distribuita su più variabili ridondanti in pochi componenti principali.
- L'**accuratezza** si è rivelata una metrica parziale e **furbante**, specialmente su dataset fortemente sbilanciati come Seismic (6.6% classe minoritaria).



# Sviluppi futuri e ottimizzazioni



- **Gestione dello Sbilanciamento delle Classi:** Implementare tecniche di **Data Augmentation** (es. SMOTE) o **algoritmi cost-sensitive** come Balanced Random Forest per migliorare il richiamo (recall) sulla classe minoritaria.
- **Ottimizzazione del Trade-off Precision/Recall:** Calibrazione della soglia di decisione mediante l'analisi della curva ROC-AUC, al fine di massimizzare l'F1-Score in contesti critici.
- **Fine-tuning del Modello:** Introduzione di procedure di **Hyperparameter Tuning** sistematico (es. GridSearch o RandomizedSearch) per identificare la configurazione ottimale dei parametri.
- **Analisi della Dimensionalità** Ottima: Estendere il testing su un range più ampio di componenti/feature (testando varie soglie), valutando l'impatto della riduzione dimensionale sulla stabilità del modello.



THANK YOU  
SO MUCH

Merci



GRAZIE