

Analisi e Classificazione di Dataset mediante Tecniche di Machine Learning

Anna Chiara Mameli 60/99/00010 - Simone Rubiu 60/68/65278 - Michele Cocco 60/99/00025^a

^aUniversità degli Studi di Cagliari - Corso di Data Mining

Sommario—Il presente progetto ha l'obiettivo di condurre un'analisi preliminare di quattro dataset distinti, applicando tecniche di classificazione supervisionate mediante sette modelli predittivi: Decision Tree, K-Nearest Neighbors, Support Vector Machine, Multi-Layer Perceptron, Random Forest, AdaBoost e XGBoost. Per ciascun dataset sono state effettuate analisi esplorative, operazioni di preprocessing, addestramento dei modelli con cross-validation stratificata e confronto delle prestazioni. È stata inoltre valutata l'efficacia di quattro tecniche di riduzione dimensionale (PCA, Feature Importance, RFE, Mutual Information) proiettando i dati in spazi bidimensionali e confrontando i risultati con quelli ottenuti sui dati originali.

1. INTRODUZIONE

Il presente progetto ha l'obiettivo di condurre un'analisi preliminare di quattro dataset distinti, applicando tecniche di classificazione supervisionate mediante diversi modelli predittivi. L'analisi si propone di confrontare le prestazioni di vari algoritmi di machine learning, valutando la loro efficacia su dati con caratteristiche differenti.

I dataset selezionati provengono dal repository OpenML e rappresentano problemi di classificazione binaria in domini diversi: autenticazione di banconote, rilevamento di segnali radar, diagnosi di tumori e previsione di eventi sismici. Questa varietà consente di analizzare il comportamento dei classificatori su dati con differenti dimensionalità, bilanciamento delle classi e natura delle feature.

Gli obiettivi principali del progetto sono:

- Eseguire un'analisi esplorativa completa dei dataset
- Applicare tecniche di preprocessing appropriate
- Implementare e confrontare 7 algoritmi di classificazione
- Valutare l'impatto della riduzione dimensionale sulle prestazioni
- Analizzare criticamente i risultati ottenuti

1.1. Dataset

I quattro dataset analizzati provengono da domini applicativi diversi tra loro, spaziando dal riconoscimento di immagini alla diagnostica medica, dall'analisi di segnali radar alla previsione di eventi geologici. Questa eterogeneità permette di valutare il comportamento degli algoritmi di classificazione in contesti reali con caratteristiche differenti.

1.1.1. Banknote Authentication

Il dataset Banknote Authentication affronta il problema di distinguere le banconote autentiche da quelle contraffatte. I dati sono stati ottenuti acquisendo immagini in scala di grigi di banconote (sia autentiche che contraffatte) con una risoluzione di 400x400 pixel. A queste immagini è stata applicata la **trasformata wavelet**, una tecnica matematica che scomponete il segnale in diverse componenti frequenziali, permettendo di catturare sia le caratteristiche globali che i dettagli locali dell'immagine. Da questa trasformazione sono state estratte quattro feature statistiche:

- **Variance**: misura la dispersione dei coefficienti wavelet, indicando quanto i valori si discostano dalla media.
- **Skewness**: indica l'asimmetria della distribuzione dei coefficienti.

Progetto di Machine Learning: analisi di quattro dataset (Banknote Authentication, Ionosphere, Breast Cancer Wisconsin, Seismic Bumps) mediante sette algoritmi di classificazione supervisionata.

- **Curtosis**: descrive la “pesantezza” delle code della distribuzione rispetto a una distribuzione normale.
- **Entropy**: quantifica il disordine o la casualità nell'immagine.

Il dataset contiene **1372 istanze**, suddivise in due classi: banconote autentiche (762 campioni, 55.5%) e contraffatte (610 campioni, 44.5%).

1.1.2. Ionosphere

Il dataset Ionosphere proviene da un sistema radar installato a Goose Bay, Labrador (Canada), progettato per studiare la ionosfera terrestre. La ionosfera è uno strato dell'atmosfera, situato tra 60 e 1000 km di altitudine, caratterizzato dalla presenza di elettroni liberi prodotti dalla radiazione solare.

Il sistema radar trasmette impulsi ad alta frequenza verso la ionosfera e analizza i segnali di ritorno. L'obiettivo è classificare questi segnali in due categorie:

- **Good**: segnali che mostrano evidenza di struttura nella ionosfera, indicando la presenza di elettroni liberi che hanno riflesso l'onda radar.
- **Bad**: segnali che passano attraverso la ionosfera senza essere riflessi, suggerendo l'assenza di strutture ionizzate significative.

Il radar utilizza un array di 16 antenne ad alta frequenza, con una potenza trasmessa totale di 6.4 kilowatt. I segnali ricevuti vengono elaborati mediante una funzione di autocorrelazione. Ogni istanza del dataset è descritta da **34 attributi numerici continui**: per ciascuno dei 17 impulsi analizzati, vengono registrati 2 valori (parte reale e immaginaria del segnale complesso). Tutti i valori sono normalizzati nell'intervallo [-1, 1]. La feature *a02* risulta costante (valore zero per tutte le istanze) e viene rimossa in fase di preprocessing.

Il dataset comprende **351 istanze**, con distribuzione: 225 segnali “good” (64.1%) e 126 segnali “bad” (35.9%).

1.1.3. Breast Cancer Wisconsin (WDBC)

Il dataset Wisconsin Diagnostic Breast Cancer (WDBC) contiene dati derivanti dall'analisi di immagini di **aspirati con ago sottile** (**Fine Needle Aspiration - FNA**) di masse mammarie. Le immagini digitalizzate degli aspirati vengono elaborate per identificare e caratterizzare i nuclei cellulari presenti. Per ogni nucleo vengono calcolate 10 caratteristiche morfologiche fondamentali:

- **Radius**: distanza media dal centro ai punti del perimetro
- **Texture**: deviazione standard dei valori di grigio
- **Perimeter** e **Area**: dimensioni geometriche del nucleo
- **Smoothness**: variazione locale nella lunghezza dei raggi
- **Compactness**: calcolata come $(\text{perimetro}^2 / \text{area} - 1)$
- **Concavity** e **Concave points**: caratteristiche delle porzioni concave
- **Symmetry**: simmetria del nucleo
- **Fractal dimension**: complessità del contorno

Per ciascuna di queste 10 caratteristiche vengono calcolate tre statistiche aggregate: la **media (mean)**, l'**errore standard (SE)** e il **valore peggiore (worst)**. Questo porta a un totale di **30 feature** (10×3).

Il dataset contiene **569 casi**, classificati come benigni (357 casi, 62.7%) o maligni (212 casi, 37.3%).

1.1.4. Seismic Bumps

Il dataset Seismic Bumps affronta un problema di sicurezza industriale: la previsione di eventi sismici pericolosi nelle miniere di carbone. Questi eventi, chiamati “bumps” o “rock bursts”, sono improvvise e violente espulsioni di roccia dalle pareti della miniera causate dall’accumulo e rilascio di energia elastica nel sottosuolo.

I dati provengono da due miniere di carbone polacche. Il sistema di rilevamento utilizza diversi tipi di sensori per misurare l’attività sismica e le condizioni geofisiche. Le feature includono:

- **Attributi sismici:** valutazione dell’attività sismica, metodo di valutazione sismoacustica, numero di scosse in diversi intervalli di energia
- **Attributi energetici:** energia totale e massima registrata, deviazione dell’energia rispetto ai valori attesi
- **Attributi operativi:** tipo di turno, presenza di drenaggio del metano

Una particolarità di questo dataset è la presenza di **variabili categoriche** (seismic, seismoacoustic, shift, ghazard) accanto alle variabili numeriche, trasformate mediante One-Hot Encoding.

Il dataset contiene **2584 istanze** ed è caratterizzato da un **forte sbilanciamento**: solo il 6.6% dei casi (170 istanze) corrisponde a eventi pericolosi (hazardous), mentre il 93.4% (2414 istanze) rappresenta situazioni non pericolose.

1.1.5. Riepilogo dei Dataset

Tabella 1. Caratteristiche principali dei quattro dataset analizzati.

| Dataset | Istanze | Feature | Classe Min. |
|------------|---------|---------|-------------|
| Banknote | 1372 | 4 | 44.5% |
| Ionosphere | 351 | 34 | 35.9% |
| Cancer | 569 | 30 | 37.3% |
| Seismic | 2584 | 18 | 6.6% |

2. PREPROCESSING

La fase di preprocessing è fondamentale per garantire la qualità dei dati e l’efficacia dei modelli di classificazione.

2.1. Verifica dell’Integrità e Valori Mancanti

Una fase preliminare essenziale del preprocessing ha riguardato la verifica dell’integrità strutturale dei dataset. È stata condotta una scansione completa di tutte le feature presenti nei dataset selezionati. L’analisi ha confermato la totale assenza di valori nulli; di conseguenza, i dataset si presentano densi e completi, rendendo non necessaria l’applicazione di tecniche di imputazione o la rimozione di istanze.

2.2. Encoding delle Variabili Categorical

Il dataset Seismic Bumps presenta diverse feature di natura categorica nominale: seismic, seismoacoustic, shift e ghazard. Poiché gli algoritmi di Machine Learning selezionati operano su input numerici, è stato necessario convertire tali attributi in un formato idoneo. Per questo, è stata adottata la tecnica del One-Hot Encoding. Operativamente, la trasformazione è stata eseguita tramite la funzione `pd.get_dummies()` della libreria Pandas, che ha espanso lo spazio delle feature creando nuove colonne binarie per ogni livello delle variabili originali.

2.3. Rimozione Feature a Varianza zero

La fase di pre-elaborazione ha incluso una riduzione dimensionale iniziale basata sulla varianza delle feature. È stata impostata una soglia di taglio pari a zero con l’obiettivo di identificare e rimuovere le variabili costanti.

La logica alla base di questa scelta è che una feature con varianza zero assume lo stesso valore per la totalità dei campioni; in termini pratici, una varianza di zero implica che la variabile è identica in circa il 100% delle osservazioni. Tali feature non apportano alcun contenuto informativo utile al modello, in quanto non possiedono la capacità di discriminare tra classi diverse. La loro rimozione permette quindi di semplificare il dataset eliminando il rumore e la ridondanza, senza rischiare di perdere informazioni significative per la classificazione. L’applicazione di tale filtro si è rivelata determinante per i dataset Ionosphere e Seismic Bumps, permettendo l’identificazione e l’eliminazione delle feature costanti (`a02` per ionosphere e `nbumps6`, `nbumps7`, `nbumps89` per seismic bumps).

2.4. Standardizzazione delle Feature

Per garantire l’omogeneità dei dati e facilitare la convergenza degli algoritmi, ai dataset Banknote, Cancer e Seismic Bumps è stata applicata la tecnica di standardizzazione (z-score normalization) utilizzando la classe `StandardScaler` di Scikit-learn. Tale trasformazione ridimensiona la distribuzione dei valori affinché abbiano media nulla e varianza unitaria, secondo la formula:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

dove x è il valore originale, μ la media del campione e σ la deviazione standard.

Questa operazione è critica per modelli basati sul calcolo delle distanze (come KNN e SVM) o sull’ottimizzazione tramite discesa del gradiente (come MLP). Per il dataset Ionosphere, l’analisi preliminare ha mostrato che i dati erano già normalizzati all’origine (intervallo [-1, 1]), rendendo superfluo un ulteriore ridimensionamento.

2.5. Partizionamento dei Dati (Train/Test Split)

Al fine di valutare le capacità di generalizzazione dei modelli su dati non visti, è stata adottata una strategia di validazione di tipo Hold-out. Tutti i dataset sono stati partizionati in due sottoinsiemi disgiunti: Training Set (80%) utilizzato per l’addestramento dei modelli e la selezione delle feature, e Test Set (20%) riservato esclusivamente per la valutazione finale delle performance.

L’operazione è stata eseguita tramite la funzione `train_test_split` di Scikit-learn, impostando un parametro `random_state` fisso per garantire la riproducibilità degli esperimenti. Inoltre, è stato attivato il parametro di stratificazione (`stratify`), che assicura che la proporzione originale tra le classi venga mantenuta inalterata sia nel training che nel test set.

2.6. Analisi Esplorativa dei Dati

Prima di procedere alla modellazione, è stata condotta una fase di esplorazione visiva per comprendere la struttura dei dati e identificare potenziali criticità. Per ogni dataset sono state prodotte le seguenti visualizzazioni diagnostiche:

- **Grafico a torta:** Utilizzato per visualizzare la proporzione tra le classi target, quantificando il grado di sbilanciamento.
- **Coordinate Parallele:** Una visualizzazione che permette di osservare contemporaneamente tutte le feature di un’istanza, utile per valutare la separabilità delle classi.
- **Istogrammi:** Generati per ogni feature suddivisa per classe target, permettono di analizzare la forma delle distribuzioni e il grado di sovrapposizione tra le classi.
- **Box Plot:** Utili per studiare la variabilità dei dati attraverso i quartili e per l’identificazione degli outlier.
- **Matrice di Correlazione:** Rappresentata tramite heatmap, permette di quantificare le relazioni lineari tra le variabili (Indice di Pearson) e individuare fenomeni di collinearità.

2.6.1. Grafici: Banknote Authentication

Distribuzione delle classi (Fig. 1): Il grafico a torta mostra la distribuzione tra le classi: 55.5% di banconote autentiche (G) contro 44.5% di banconote contraffatte (F).

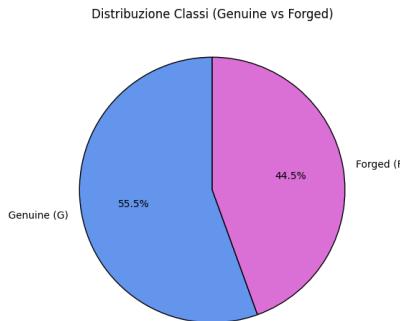


Figura 1. Distribuzione delle classi nel dataset Banknote: 55.5% Genuine vs 44.5% Forged.

Coordinate Parallelle (Fig. 2): La rappresentazione mediante coordinate parallele mostra il comportamento delle quattro feature rispetto alle due classi. La feature *variance* evidenzia una distinzione tra le classi: le banconote autentiche (blu) si concentrano nella regione dei valori positivi (circa 0–7), mentre quelle contraffatte (rosa) occupano prevalentemente l'intervallo dei valori negativi (circa –7–0).

Per la feature *skewness*, le istanze autentiche tendono ad assumere valori positivi più elevati (5–13) rispetto alle contraffatte. La feature *curtosis* presenta maggiore sovrapposizione tra le classi, pur mantenendo tendenze opposte. L'*entropy* risulta la variabile con maggiore sovrapposizione lungo l'intero intervallo osservato (–8–2).

È inoltre evidente il pattern di incrocio osservabile tra *skewness* e *curtosis*, in corrispondenza del quale le traiettorie delle due classi tendono visivamente a invertirsi.

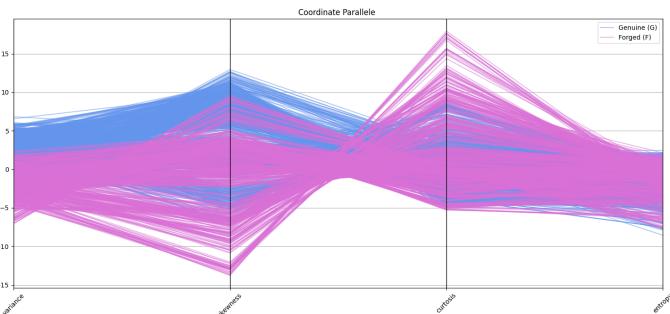


Figura 2. Coordinate parallele per Banknote. Le banconote autentiche (blu) e contraffatte (rosa) mostrano pattern distinguibili, specialmente in *variance* e *skewness*.

Istogrammi (Fig. 3): L'analisi delle distribuzioni mostra che la *variance* presenta la minore sovrapposizione tra le classi: le banconote autentiche si concentrano su valori positivi (picco intorno a 4–5), mentre le contraffatte mostrano una distribuzione prevalentemente negativa e bimodale.

Per la *skewness*, le istanze autentiche occupano valori più elevati (circa 7–9) rispetto alle contraffatte, con una regione di parziale sovrapposizione nell'intervallo 0–5. La *curtosis* presenta ampie sovrapposizioni tra le distribuzioni. L'*entropy* mostra distribuzioni quasi coincidenti tra le due classi.

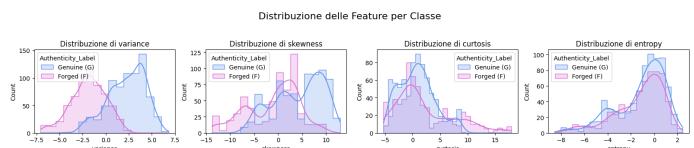


Figura 3. Distribuzione delle feature per classe. *Variance* e *skewness* mostrano la migliore separazione, mentre *entropy* presenta sovrapposizione quasi totale.

Box Plot (Fig. 4): I box plot confermano le osservazioni precedenti. Per la *variance*, la mediana delle banconote autentiche (circa 2.5) è superiore a quella delle contraffatte (circa –1.5), con scarsa sovrapposizione tra i box. Per la *skewness*, le mediane sono distanziate (circa 5 per le autentiche contro 1 per le contraffatte) con parziale sovrapposizione degli intervalli interquartili.

La *curtosis* mostra significativa sovrapposizione tra le distribuzioni e presenza di outlier. L'*entropy* presenta box quasi coincidenti e mediane simili (circa –1).

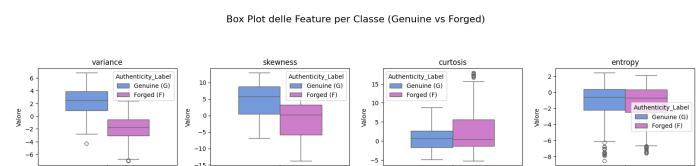


Figura 4. Box plot delle feature per classe. *Variance* mostra la separazione più netta tra le classi Genuine (blu) e Forged (rosa).

Matrice di Correlazione (Fig. 5): La heatmap evidenzia una correlazione negativa moderata (circa –0.4) tra *skewness* e *curtosis*. La *variance* presenta correlazione negativa con la *curtosis* (circa –0.3) e correlazione positiva debole con la *skewness*. L'*entropy* mostra correlazioni deboli con tutte le altre feature. L'assenza di correlazioni elevate (in valore assoluto > 0.8) indica che le feature forniscono informazioni complementari.

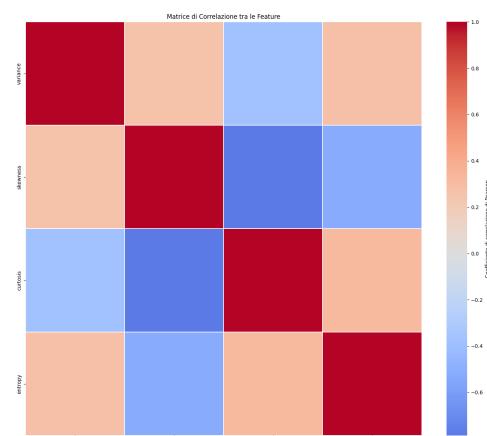


Figura 5. Matrice di correlazione del dataset Banknote. Si osserva la correlazione negativa tra *skewness* e *curtosis*.

2.6.2. Grafici: Ionosphere

Distribuzione delle classi (Fig. 6): Il grafico a torta mostra la distribuzione tra le classi: 64.1% di segnali *Good* e 35.9% di segnali *Bad*.

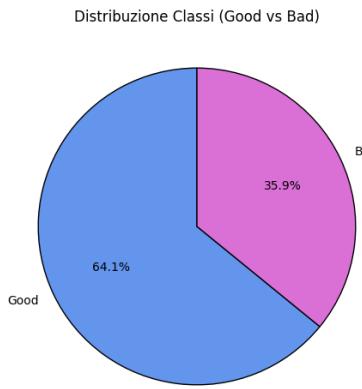
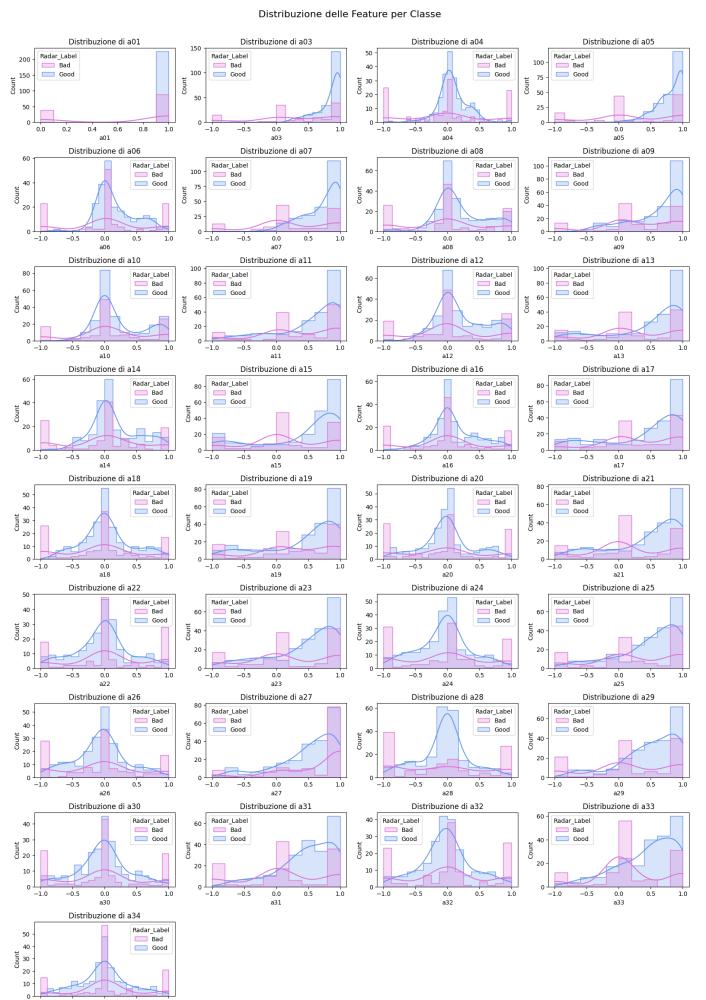


Figura 6. Distribuzione delle classi nel dataset Ionosphere: 64.1% Good vs 35.9% Bad.



Coordinate Parallelle (Fig. 7): L'analisi delle coordinate parallele sulle 34 feature, normalizzate nell'intervallo $[-1, 1]$, evidenzia che la feature $a01$ mostra un comportamento discriminativo: i segnali *Good* (blu) risultano concentrati su valori prossimi a 1, mentre i segnali *Bad* (rosa) presentano una distribuzione più dispersa.

L'andamento complessivo delle traiettorie rivela un pattern alternato tra feature dispari e pari, coerente con la rappresentazione delle componenti reale e immaginaria del segnale radar. I segnali *Good* presentano oscillazioni più regolari, mentre i segnali *Bad* appaiono caratterizzati da maggiore variabilità.

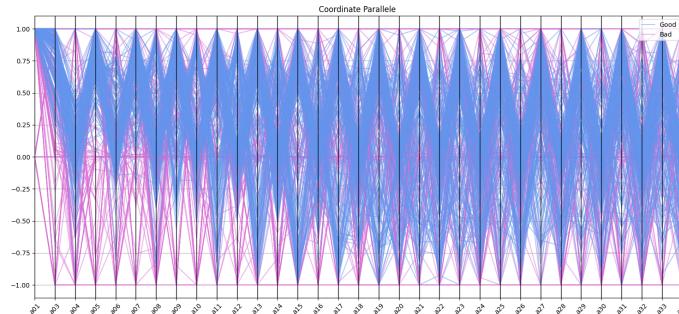


Figura 7. Coordinate parallele per il dataset Ionosphere. I segnali Good (blu) mostrano pattern oscillatori più regolari rispetto ai Bad (rosa).

Istogrammi (Fig. 8): La feature $a01$ risulta discriminativa, con i segnali *Good* concentrati su valori prossimi a 1 e i *Bad* distribuiti in modo più uniforme. Le feature dispari (parte reale del segnale) tendono a presentare distribuzioni bimodali con accumuli agli estremi, mentre le feature pari (parte immaginaria) mostrano profili più simmetrici e centrati attorno allo zero. È presente sovrapposizione tra le distribuzioni delle due classi per molte feature.

Box Plot (Fig. 9): La feature $a01$ mostra separazione tra le classi: la mediana dei segnali *Good* è prossima a 1, mentre quella dei *Bad* è inferiore. La feature $a03$ presenta mediane lievemente differenti tra le classi, ma è caratterizzata da un'elevata variabilità. Per le altre feature si osserva sovrapposizione tra le distribuzioni. La presenza complessiva di outlier risulta limitata, coerentemente con la normalizzazione nell'intervallo $[-1, 1]$.

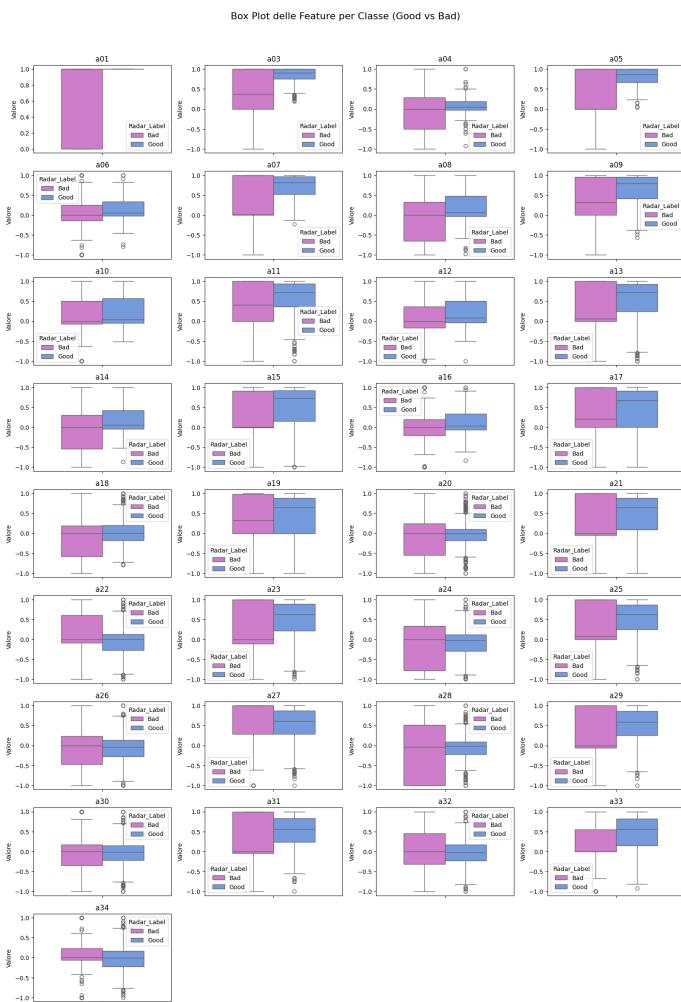


Figura 9. Box plot delle feature per classe. La feature *a01* evidenzia la separazione più netta tra segnali Good (blu) e Bad (rosa).

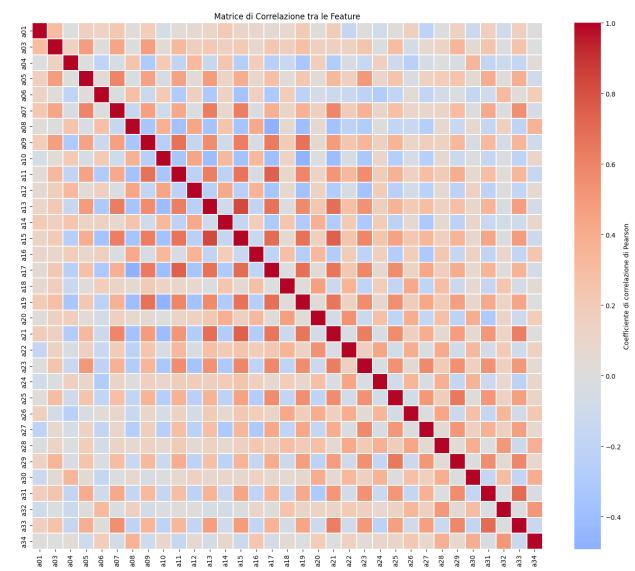


Figura 10. Matrice di correlazione del dataset Ionosphere. È visibile una struttura a blocchi lungo la diagonale.

2.6.3. Grafici: Breast Cancer Wisconsin

Distribuzione delle classi (Fig. 11): Il grafico a torta mostra la distribuzione tra le classi: 62.7% di casi benigni (B) e 37.3% di casi maligni (M).

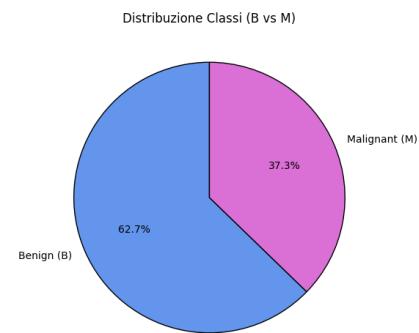


Figura 11. Distribuzione delle classi nel dataset Breast Cancer Wisconsin: 62.7% Benigni vs 37.3% Maligni.

Matrice di Correlazione (Fig. 10): La matrice di correlazione mostra una struttura con pattern a scacchiera. Le correlazioni più elevate si concentrano prevalentemente nella regione centrale della matrice, mentre ampie aree con correlazioni prossime allo zero indicano che molte coppie di variabili mantengono indipendenza lineare. Questa struttura suggerisce che i diversi ritardi temporali catturano informazioni in parte complementari.

Coordinate Parallele (Fig. 12): La rappresentazione mediante coordinate parallele mostra che i casi maligni (linee rosa) tendono ad assumere valori più elevati rispetto ai benigni (linee blu), in particolare nelle feature dimensionali quali *area_mean* e *area_worst*. Un comportamento analogo si osserva per *perimeter_mean* e *perimeter_worst*. Le feature con valori più contenuti, come *radius_se*, *texture_se*, *compactness_worst* e *concavity_worst*, mostrano una separazione visiva meno marcata tra le classi.

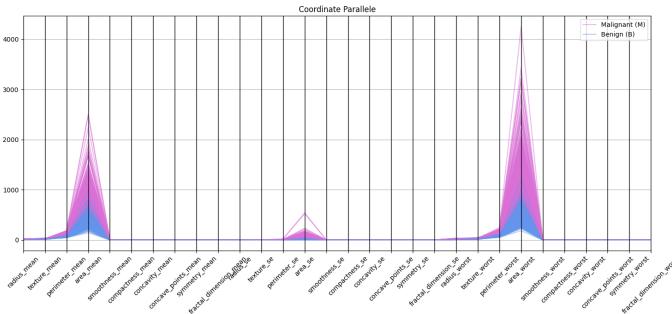


Figura 12. Coordinate parallele per il dataset Breast Cancer Wisconsin. I casi maligni (rosa) presentano valori più elevati nelle feature dimensionali.

Istogrammi (Fig. 13): Le feature del gruppo *worst* (*radius_worst*, *perimeter_worst*, *area_worst*) mostrano distribuzioni distinte tra le classi, con i casi maligni spostati verso valori più elevati. Le feature *mean* presentano capacità discriminativa intermedia, con picchi delle distribuzioni ancora distinguibili ma con parziale sovrapposizione. Le metriche basate sull'errore standard ("SE") mostrano spesso distribuzioni quasi coincidenti. Le feature morfologiche quali *compactness_worst* indicano che i tumori maligni tendono ad associarsi a forme più irregolari.

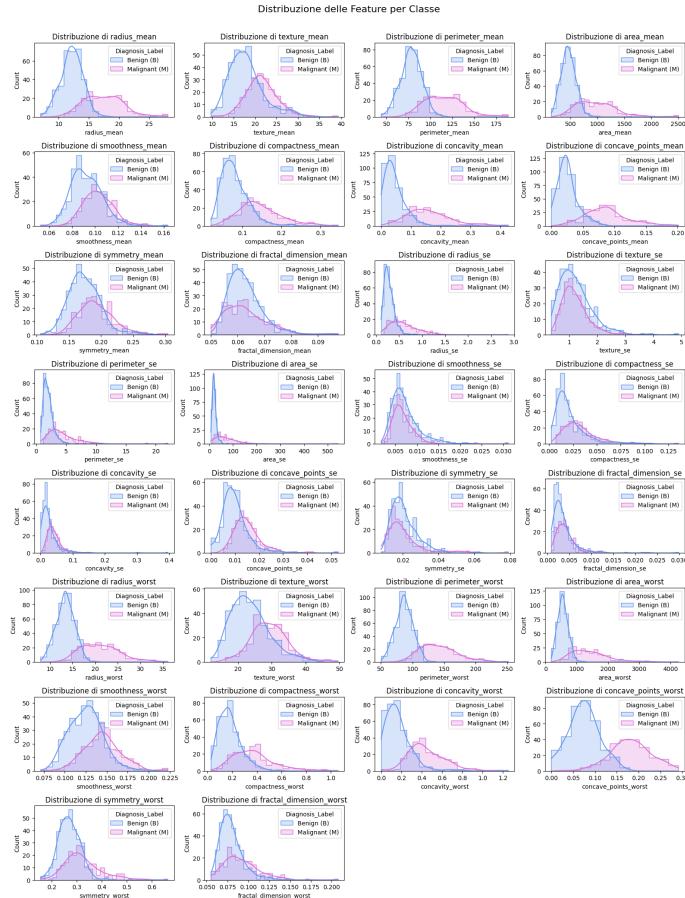


Figura 13. Distribuzione delle feature per classe. Le feature del gruppo *worst* mostrano la migliore separazione tra benigni (blu) e maligni (rosa).

Box Plot (Fig. 14): In quasi tutte le feature, la mediana dei casi maligni risulta superiore a quella dei benigni, con differenze marcate nelle variabili *radius*, *perimeter* e *area*. I tumori maligni mostrano inoltre maggiore variabilità, evidenziata da box più ampi e da una presenza più frequente di outlier. La feature *area_worst* presenta separazione tra le distribuzioni delle due classi, con il terzo quartile dei benigni inferiore al primo quartile dei maligni.

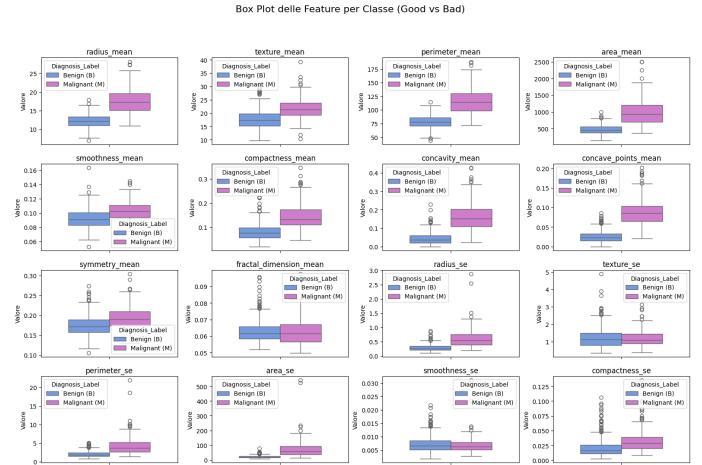


Figura 14. Box plot delle feature per classe. I casi maligni (rosa) mostrano mediane più elevate e maggiore dispersione rispetto ai benigni (blu).

Matrice di Correlazione (Fig. 15): La matrice di correlazione mostra blocchi di elevata correlazione tra feature geometricamente interdipendenti. Variabili come *radius*, *perimeter* e *area* mostrano correlazioni elevate (> 0.9) sia nelle versioni *mean* che *worst*, dato atteso vista la loro relazione matematica diretta.

Le feature basate sull'errore standard ("SE") presentano correlazioni più deboli con il resto del dataset. Un comportamento di relativa indipendenza emerge anche per *texture_mean* e *texture_worst*, poco correlate con le metriche dimensionali. L'elevata ridondanza informativa tra alcune feature suggerisce che tecniche di riduzione dimensionale come la PCA potrebbero comprimere l'informazione in un numero ridotto di componenti.

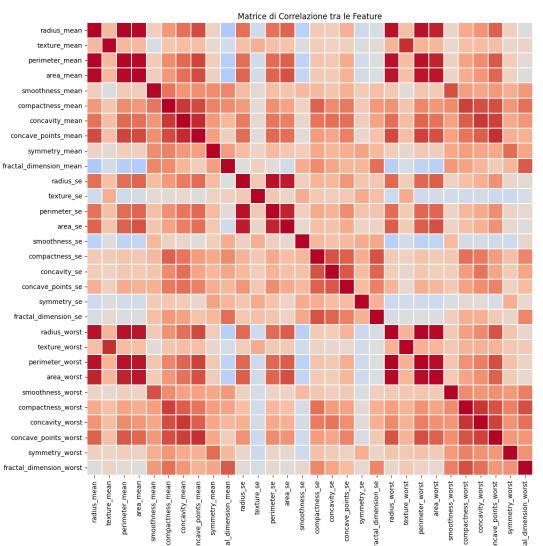


Figura 15. Matrice di correlazione del dataset Breast Cancer Wisconsin. I blocchi ad alta correlazione evidenziano la dipendenza tra le feature geometriche.

2.6.4. Grafici: Seismic Bumps

Distribuzione delle classi (Fig. 16): Il grafico a torta evidenzia uno sbilanciamento marcato tra le classi, con il 93.4% di eventi *Non-hazardous* e il 6.6% di eventi *Hazardous*. Tale configurazione rende necessario l'impiego di metriche quali *Precision*, *Recall* e *F1-score* per una valutazione delle prestazioni dei modelli.

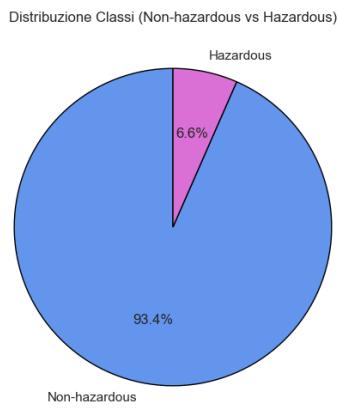


Figura 16. Distribuzione delle classi nel dataset Seismic Bumps: 93.4% Non-hazardous vs 6.6% Hazardous.

Coordinate Parallele (Fig. 17): La visualizzazione risulta dominata dalla feature *genergy*, che assume valori dell'ordine di 10^6 e presenta una marcata asimmetria, insieme ai picchi isolati osservabili in *energy* e *maxenergy*. Tale predominanza delle variabili energetiche comprime visivamente verso lo zero le feature di conteggio e le variabili categoriche codificate. Si osserva sovrapposizione tra le traiettorie delle classi *Non-hazardous* (blu) e *Hazardous* (rosa).

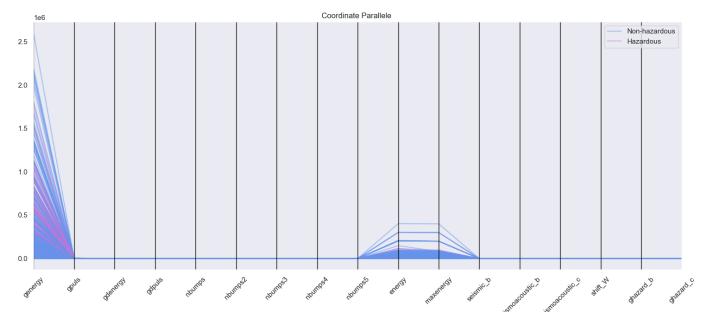


Figura 17. Coordinate parallele per il dataset Seismic Bumps. La scala è dominata dalle feature energetiche.

Istogrammi (Fig. 18): Le feature energetiche (ad esempio *genergy*, *gdenergy*) presentano distribuzioni fortemente asimmetriche, caratterizzate da lunghe code di valori estremi; la concentrazione della maggior parte delle osservazioni su valori bassi determina sovrapposizione tra le classi. Le feature di conteggio (come *nbumps*) mostrano distribuzioni discrete concentrate sui valori 0–2. Le feature binarie ottenute tramite One-Hot Encoding evidenziano proporzioni simili tra le classi.

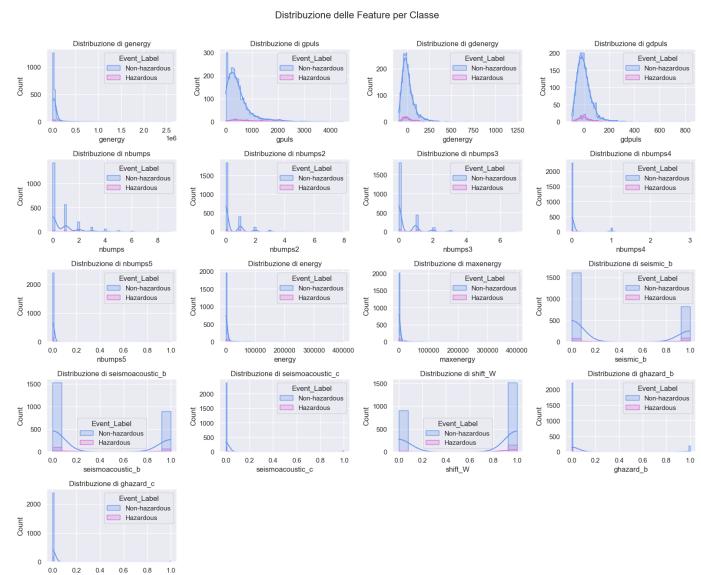


Figura 18. Distribuzione delle feature per classe nel dataset Seismic Bumps. Si osserva sovrapposizione tra eventi Non-hazardous (blu) e Hazardous (rosa).

Box Plot (Fig. 19): I box plot evidenziano la presenza di numerosi outlier nelle feature energetiche. Dal confronto tra le classi emerge che gli eventi *Hazardous* tendono a presentare mediane leggermente più elevate in variabili quali *genergy*, *gpuls* e *nbumps*. Le variabili binarie mostrano mediane stabili e pressoché identiche tra le classi. Le feature *energy* e *maxenergy* risultano concentrate vicino allo zero, con pochi outlier estremi.

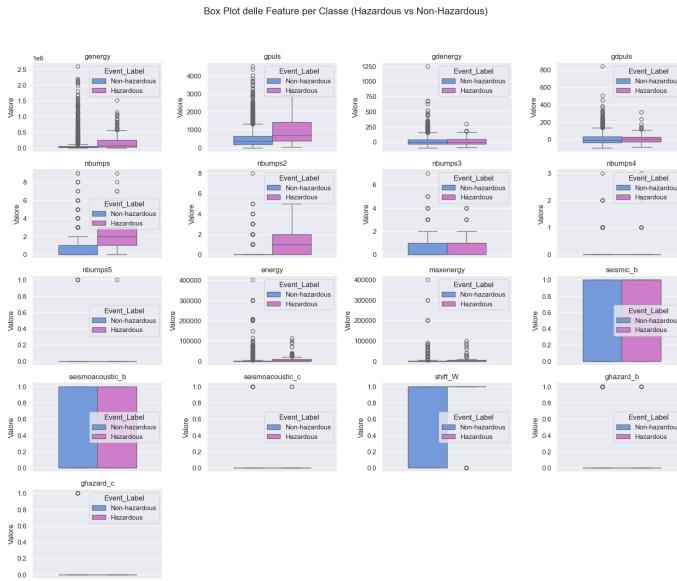


Figura 19. Box plot delle feature per classe nel dataset Seismic Bumps. Gli eventi Hazardous (rosa) mostrano mediane leggermente più elevate nelle feature energetiche.

Matrice di Correlazione (Fig. 20): La heatmap delle correlazioni evidenzia una struttura a blocchi. Si osserva forte correlazione positiva all'interno del gruppo delle feature energetiche (ad esempio *genergy/gpuls* ed *energy/maxenergy*), nonché tra le feature di conteggio (*nbumps* e varianti). Le feature categoriche binarie risultano sostanzialmente isolate, mostrando correlazioni deboli o nulle con le altre variabili. Questa struttura indica ridondanza all'interno dei blocchi e complementarietà tra gruppi diversi.

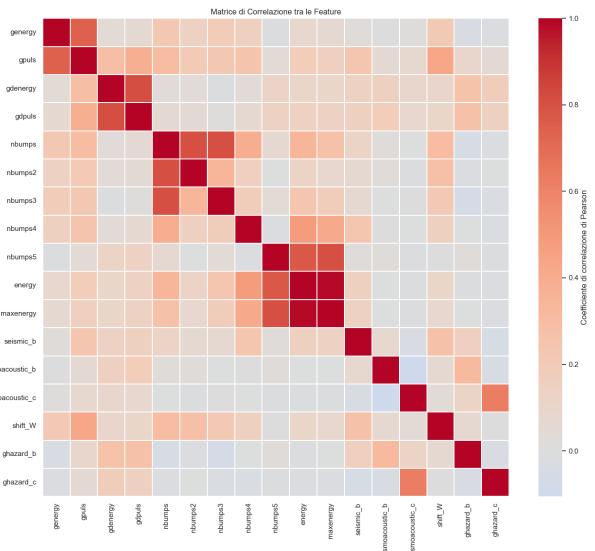


Figura 20. Matrice di correlazione del dataset Seismic Bumps. Sono evidenti blocchi di alta correlazione tra feature energetiche e tra feature di conteggio.

3. CLASSIFICAZIONE

La fase di classificazione ha previsto l'addestramento e la valutazione di sette algoritmi di machine learning su ciascun dataset. I modelli selezionati rappresentano diverse famiglie di approcci: modelli basati su istanze (K-Nearest Neighbors), modelli lineari con kernel (Support Vector Machine), reti neurali (Multi-Layer Perceptron), modelli basati su alberi decisionali (Decision Tree) e metodi ensemble sia di tipo bagging (Random Forest) che boosting (AdaBoost, XGBoost).

3.1. Metodologia

Per ciascun modello è stata effettuata una validazione mediante **5-Fold Stratified Cross-Validation** sul training set (80% dei dati), seguita da una valutazione finale sul test set (20%). La stratificazione garantisce che ogni fold mantenga la proporzione originale delle classi, aspetto cruciale per i dataset sbilanciati.

Le metriche di valutazione utilizzate sono: **Accuracy**, **Precision**, **Recall** e **F1-Score**. Quest'ultimo è stato scelto come criterio principale per la selezione del miglior modello, in quanto bilancia Precision e Recall ed è più informativo dell'Accuracy in presenza di sbilanciamento tra le classi. È stata inoltre calcolata l'**AUC-ROC** (Area Under the ROC Curve) per valutare la capacità discriminativa complessiva dei classificatori.

Per gestire lo sbilanciamento delle classi, i modelli che lo supportano (Decision Tree, SVM, Random Forest) sono stati configurati con il parametro `class_weight='balanced'`, che assegna pesi inversamente proporzionali alla frequenza delle classi. Per XGBoost è stato utilizzato il parametro `scale_pos_weight`, calcolato come rapporto tra la classe maggioritaria e quella minoritaria.

3.2. Risultati: Banknote Authentication

Tabella 2. Metriche di classificazione sul test set per Banknote Authentication.

| Modello | Accuracy | Precision | Recall | F1 |
|---------------|----------|-----------|--------|-------|
| KNN | 1.000 | 1.000 | 1.000 | 1.000 |
| SVM | 1.000 | 1.000 | 1.000 | 1.000 |
| MLP | 1.000 | 1.000 | 1.000 | 1.000 |
| AdaBoost | 1.000 | 1.000 | 1.000 | 1.000 |
| Random Forest | 0.996 | 0.992 | 1.000 | 0.996 |
| Decision Tree | 0.993 | 0.984 | 1.000 | 0.992 |
| XGBoost | 0.993 | 1.000 | 0.984 | 0.992 |

Il dataset Banknote rappresenta un problema di classificazione relativamente semplice grazie alla buona separabilità delle classi osservata nell'analisi esplorativa. Quattro modelli (KNN, SVM, MLP, AdaBoost) hanno raggiunto un F1-score pari a 1.0, mentre i restanti si attestano sopra 0.99. Le curve ROC (Fig. 21) confermano questa tendenza, con valori di AUC prossimi a 1.0 per tutti i classificatori. Il **miglior modello selezionato è in questo caso AdaBoost**.

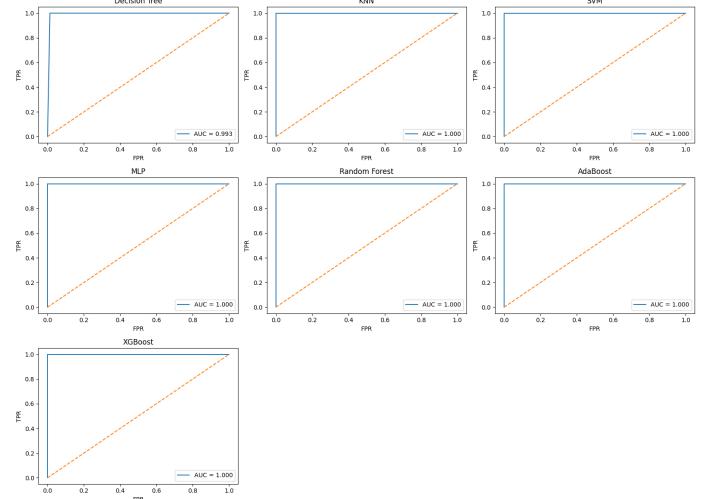


Figura 21. Curve ROC per il dataset Banknote Authentication. Tutti i modelli mostrano AUC ≥ 0.993 .

3.3. Risultati: Ionosphere

Tabella 3. Metriche di classificazione sul test set per Ionosphere.

| Modello | Accuracy | Precision | Recall | F1 |
|---------------|----------|-----------|--------|-------|
| Random Forest | 0.958 | 1.000 | 0.880 | 0.936 |
| XGBoost | 0.958 | 1.000 | 0.880 | 0.936 |
| SVM | 0.944 | 0.957 | 0.880 | 0.917 |
| AdaBoost | 0.944 | 1.000 | 0.840 | 0.913 |
| MLP | 0.930 | 1.000 | 0.800 | 0.889 |
| Decision Tree | 0.915 | 0.852 | 0.920 | 0.885 |
| KNN | 0.845 | 1.000 | 0.560 | 0.718 |

Il dataset Ionosphere, con le sue 33 (una rimossa nelle fasi di pre-processing) feature derivanti da segnali radar, presenta una complessità maggiore. I modelli ensemble (Random Forest e XGBoost) ottengono i risultati migliori con F1-score di 0.936. È interessante notare come KNN mostri prestazioni nettamente inferiori (F1 = 0.718, Recall = 0.56), soffrendo della cosiddetta “curse of dimensionality” tipica degli algoritmi basati su distanze in spazi ad alta dimensionalità. Le curve ROC (Fig. 22) evidenziano questa differenza: KNN ha AUC = 0.851, mentre SVM raggiunge AUC = 0.997. Il **miglior modello selezionato in questo caso è Random Forest**.

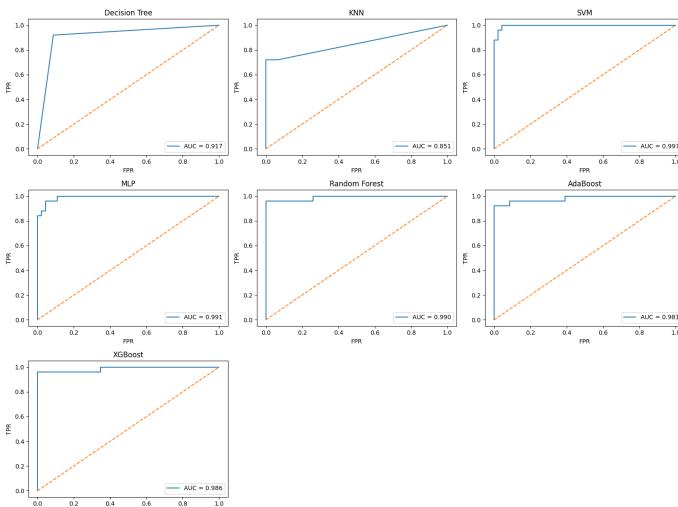


Figura 22. Curve ROC per il dataset Ionosphere. KNN mostra AUC inferiore (0.851) rispetto agli altri modelli.

3.4. Risultati: Breast Cancer Wisconsin

Tabella 4. Metriche di classificazione sul test set per Breast Cancer Wisconsin.

| Modello | Accuracy | Precision | Recall | F1 |
|---------------|----------|-----------|--------|-------|
| SVM | 0.982 | 0.976 | 0.976 | 0.976 |
| AdaBoost | 0.982 | 1.000 | 0.952 | 0.976 |
| MLP | 0.974 | 1.000 | 0.929 | 0.963 |
| Random Forest | 0.974 | 1.000 | 0.929 | 0.963 |
| XGBoost | 0.965 | 1.000 | 0.905 | 0.950 |
| KNN | 0.956 | 0.974 | 0.905 | 0.938 |
| Decision Tree | 0.904 | 0.897 | 0.883 | 0.864 |

Il dataset Breast Cancer Wisconsin presenta 30 feature morfologiche con elevata correlazione tra loro. Tutti i modelli ottengono F1-score superiori a 0.86, con SVM al primo posto (F1 = 0.976). In questo contesto, anche AdaBoost mostra le medesime prestazioni (F1 = 0.976), beneficiando probabilmente della ridondanza informativa tra le feature che mitiga gli effetti negativi dell'alta dimensionalità. Le curve ROC (Fig. 23) mostrano AUC elevate per tutti i modelli, con valori compresi tra 0.889 (Decision Tree) e 0.997 (Random Forest). Il **miglior modello selezionato in questo caso è AdaBoost**.

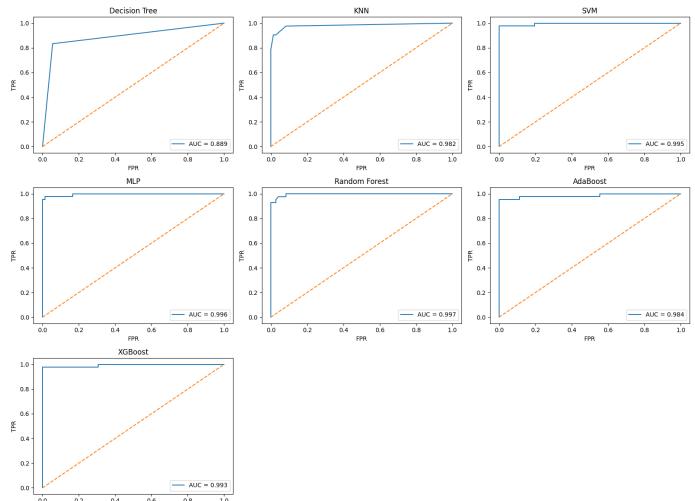


Figura 23. Curve ROC per il dataset Breast Cancer Wisconsin. Tutti i modelli mostrano AUC ≥ 0.889 .

3.5. Risultati: Seismic Bumps

Tabella 5. Metriche di classificazione sul test set per Seismic Bumps.

| Modello | Accuracy | Precision | Recall | F1 |
|---------------|----------|-----------|--------|-------|
| Decision Tree | 0.911 | 0.286 | 0.235 | 0.258 |
| XGBoost | 0.911 | 0.269 | 0.206 | 0.233 |
| SVM | 0.781 | 0.138 | 0.441 | 0.210 |
| KNN | 0.932 | 0.429 | 0.088 | 0.146 |
| MLP | 0.930 | 0.333 | 0.059 | 0.100 |
| Random Forest | 0.934 | 0.500 | 0.029 | 0.056 |
| AdaBoost | 0.934 | 0.000 | 0.000 | 0.000 |

Il dataset Seismic Bumps rappresenta la sfida più complessa a causa del forte sbilanciamento (93.4% vs 6.6%) e della sovrapposizione quasi totale delle classi osservata nell'EDA. I risultati riflettono questa difficoltà: nonostante valori di Accuracy elevati (fino a 0.934 per Random Forest e AdaBoost), gli F1-score rimangono bassi per tutti i modelli. Il Decision Tree ottiene il miglior F1-score (0.258), seguito da XGBoost (0.233).

È significativo il comportamento di AdaBoost, che non riesce a identificare alcun campione della classe minoritaria (Precision, Recall e F1 pari a 0). SVM mostra il Recall più alto (0.441) ma con Precision molto bassa (0.140), indicando un elevato numero di falsi positivi.

Le curve ROC (Fig. 24) confermano la difficoltà del problema: i valori di AUC sono compresi tra 0.597 (Decision Tree) e 0.736 (MLP), indicando una capacità discriminativa limitata, di poco superiore al caso casuale (AUC = 0.5). Il **miglior modello selezionato in questo caso è Decision Tree**.

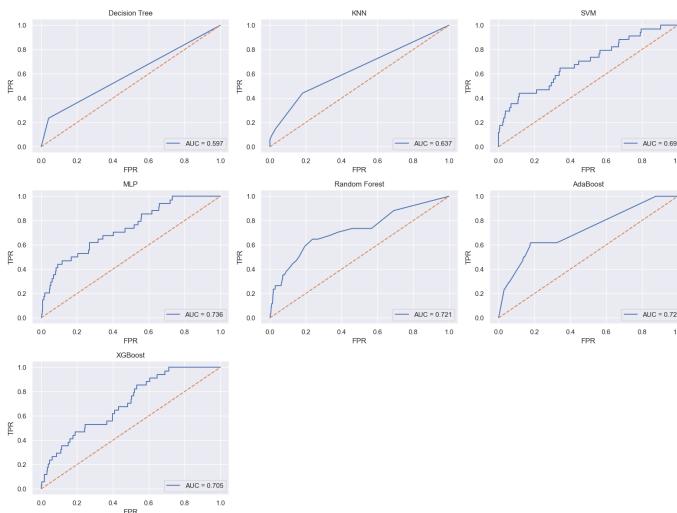


Figura 24. Curve ROC per il dataset Seismic Bumps. I valori di AUC (0.6–0.74) indicano una capacità discriminativa limitata.

3.6. Riepilogo

Tabella 6. Miglior modello selezionato per ciascun dataset sulla base dell’F1-score.

| Dataset | Miglior Modello | F1-Score | AUC |
|------------|-----------------|----------|-------|
| Banknote | AdaBoost | 1.000 | 1.000 |
| Ionosphere | Random Forest | 0.936 | 0.990 |
| Cancer | AdaBoost | 0.976 | 0.984 |
| Seismic | Decision Tree | 0.258 | 0.597 |

L’analisi evidenzia una netta superiorità dei modelli ensemble (Random Forest, AdaBoost e XGBoost) rispetto ai singoli classificatori. In particolare, AdaBoost si è dimostrato l’algoritmo più versatile e consistente, ottenendo il primato nei dataset Banknote ($F_1 = 1.000$) e Cancer ($F_1 = 0.976$). La sua forza risiede nella natura iterativa del boosting, che corregge progressivamente gli errori concentrandosi sulle istanze più difficili. Al contrario, nel dataset Seismic Bumps, il più complesso per via del forte sbilanciamento delle classi, è emerso il Decision Tree ($F_1 = 0.258$). Al contrario, nel dataset Seismic Bumps, il più compleso a causa del marcato sbilanciamento delle classi, il **Decision Tree** è risultato il modello più efficace ($F_1 = 0.258$). In questo scenario, **AdaBoost** ha registrato le performance peggiori proprio a causa dell’assenza del parametro `class_weight`.

Nello specifico, dato che solo il 6.6% del dataset è composto da campioni anomali, la natura iterativa dell’algoritmo ha finito per penalizzare la classe minoritaria. Non potendo assegnare un peso maggiore *a priori* alle anomalie, AdaBoost ha focalizzato il processo di correzione degli errori principalmente sulle istanze della classe maggioritaria (93.4% dei dati), in quanto statisticamente più frequenti. Di conseguenza, il modello ha ottimizzato la capacità di riconoscere correttamente i casi “normali”, fallendo però nel generalizzare sulla classe positiva, che è stata erroneamente trattata come rumore o come un insieme di casi trascurabili rispetto alla massa dei campioni negativi.

3.7. Riduzione Dimensionale

Per valutare l’impatto della riduzione dimensionale sulle prestazioni e permettere la visualizzazione dei decision boundary, sono state applicate quattro tecniche per ridurre lo spazio delle feature a sole 2 dimensioni. Il miglior modello selezionato per ciascun dataset è stato riaddestrato sui dati ridotti.

Le tecniche utilizzate sono:

- **PCA (Principal Component Analysis)**: tecnica di feature extraction non supervisionata che proietta i dati sulle direzioni di massima varianza.
- **Feature Importance**: selezione delle 2 feature con maggiore importanza secondo il criterio del modello.
- **RFE (Recursive Feature Elimination)**: algoritmo wrapper che rimuove iterativamente le feature meno importanti.
- **Mutual Information**: metodo filter-based che seleziona le feature con maggiore informazione mutua rispetto al target.

3.7.1. Risultati

Tabella 7. Confronto F1-Score tra dati originali e tecniche di riduzione a 2 feature.

| Dataset | Originale | PCA | Feat. Imp. | RFE | MI |
|------------|-----------|-------|------------|-------|-------|
| Banknote | 1.000 | 0.791 | 0.900 | 0.900 | 0.900 |
| Ionosphere | 0.936 | 0.680 | 0.880 | 0.880 | 0.863 |
| Cancer | 0.976 | 0.938 | 0.916 | 0.916 | 0.811 |
| Seismic | 0.258 | 0.136 | 0.119 | 0.109 | 0.102 |

Tabella 8. Dettagli della riduzione dimensionale: varianza spiegata da PCA e feature selezionate.

| Dataset | Var. PCA | Feature Importance | RFE | Mutual Info. |
|------------|----------|-----------------------------------|-----------------------------------|-------------------------------|
| Banknote | 86.9% | variance, skewness | variance, skewness | variance, skewness |
| Ionosphere | 43.7% | a05, a27 | a05, a27 | a06, a05 |
| Cancer | 63.1% | perim_worst, concave_points_worst | perim_worst, concave_points_worst | radius_worst, perimeter_worst |
| Seismic | 36% | gpus, nbumps | gpus, genergy | nbumps, energy |

I risultati mostrano comportamenti diversi a seconda delle caratteristiche dei dataset. Per **Banknote** e **Ionosphere**, le tecniche di selezione (Feat. Imp., RFE, MI) preservano meglio le prestazioni rispetto alla PCA: il calo è contenuto (circa il 10% per Banknote e 6% per Ionosphere), mentre la PCA subisce perdite molto più marcate (rispettivamente del 21% e 27%), pur catturando gran parte della varianza.

Il dataset **Cancer** presenta uno scenario differente: qui la PCA ($F_1 = 0.938$) supera le tecniche di selezione, che mostrano cali più evidenti (in particolare la MI scende a 0.811). Questo riflette l’elevata correlazione tra le feature morfologiche, che permette alla PCA di comprimere efficacemente l’informazione discriminativa nelle prime componenti principali.

Per **Seismic**, si osserva lo stesso risultato del Cancer dataset dove si ha che la PCA ottiene risultati migliori rispetto alle tecniche di selezione delle feature. Anche in questo caso, la presenza di diverse feature con elevata correlazione permette alla PCA di comprimere l’informazione nelle prime componenti principali.

3.7.2. Decision Boundary

Le figure seguenti mostrano i decision boundary ottenuti con ciascuna tecnica di riduzione dimensionale. La visualizzazione permette di osservare come le diverse proiezioni influenzino la separabilità delle classi.

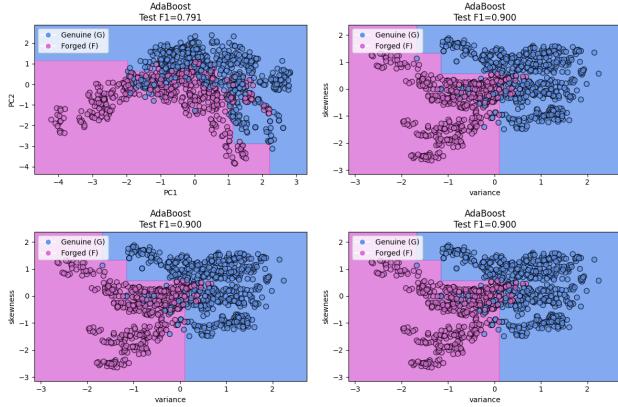


Figura 25. Decision boundary per Banknote: PCA (F1=0.791) (alto a sinistra), Feature Importance (F1=0.900) (alto a destra), RFE (F1=0.900) (basso a sinistra), Mutual Information (F1=0.900) (basso a destra).

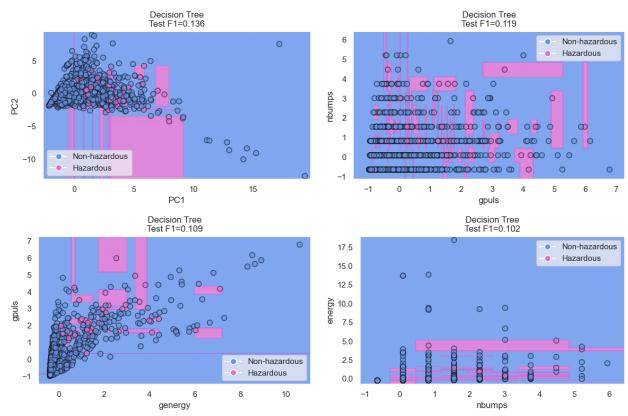


Figura 28. Decision boundary per Seismic: PCA (F1=0.136) (alto a sinistra), Feature Importance (F1=0.119) (alto a destra), RFE (F1=0.109) (basso a sinistra), Mutual Information (F1=0.102) (basso a destra).

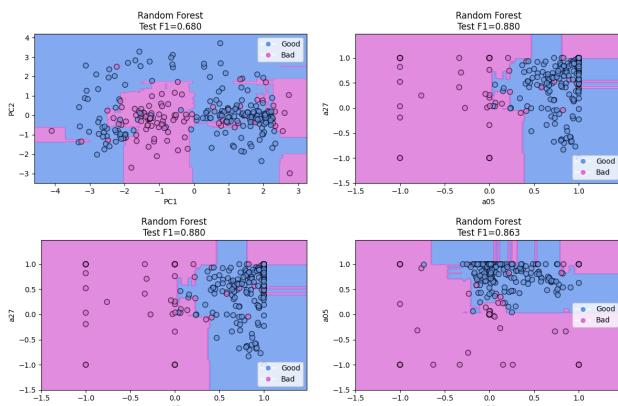


Figura 26. Decision boundary per Ionosphere: PCA (F1=0.680) (alto a sinistra), Feature Importance (F1=0.880) (alto a destra), RFE (F1=0.880) (basso a sinistra), Mutual Information (F1=0.863) (basso a destra).

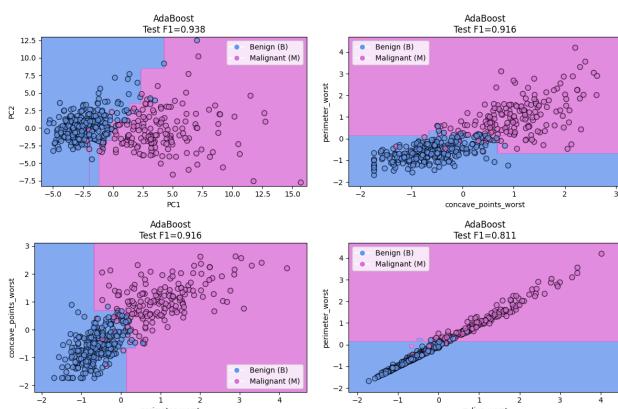


Figura 27. Decision boundary per Cancer: PCA (F1=0.938) (alto a sinistra), Feature Importance (F1=0.916) (alto a destra), RFE (F1=0.916) (basso a sinistra), Mutual Information (F1=0.911) (basso a destra).

L'analisi evidenzia che non esiste una tecnica di riduzione universalmente migliore: la scelta ottimale dipende dalla struttura dei dati. La PCA risulta efficace quando le feature sono altamente correlate (Cancer), mentre le tecniche di selezione sono preferibili quando l'informazione discriminativa è concentrata in poche feature specifiche (Banknote, Ionosphere).

È interessante notare come Feature Importance e RFE tendano a selezionare le stesse feature nella maggior parte dei casi (Banknote, Ionosphere, Breast Cancer Wisconsin), producendo risultati identici. Questo accade quando le feature più importanti sono nettamente dominanti rispetto alle altre: RFE, pur rimuovendo iterativamente le feature meno rilevanti e ri-addestrando il modello ad ogni step, converge sulla stessa selezione di Feature Importance, che si basa su un ranking statico. La divergenza tra i due metodi emerge invece nel dataset Seismic Bumps. Qui, l'eliminazione iterativa delle variabili operata dalla RFE modifica l'importanza relativa delle feature rimanenti, portando alla selezione di genergy al posto di nbumps. Questo comportamento suggerisce che nel contesto sbilanciato del Seismic, le interazioni tra le feature energetiche sono più complesse e meno lineari rispetto agli altri dataset.

In tutti i casi, la riduzione a sole 2 feature comporta una perdita di prestazioni rispetto ai dati originali, ma permette la visualizzazione dei decision boundary e una migliore interpretabilità del modello.

4. DISCUSSIONE E CONCLUSIONI

4.1. Sintesi dei Risultati

Tra i sette classificatori testati, AdaBoost si è dimostrato l'algoritmo più consistente, risultando il migliore su due dataset (Banknote con $F1 = 1.000$ e Cancer con $F1 = 0.984$). La sua efficacia deriva dalla natura iterativa del boosting: l'algoritmo costruisce un "strong classifier" combinando in sequenza diversi modelli deboli. Il funzionamento di AdaBoost si basa su un processo iterativo in cui, a ogni passaggio, l'algoritmo assegna un peso maggiore alle istanze classificate erroneamente in precedenza. Questo meccanismo forza il modello a concentrarsi sui "casi difficili", riducendo progressivamente il bias complessivo.

Tuttavia, questa strategia si rivela efficace solo quando le classi non sono eccessivamente sbilanciate. Nel caso specifico del dataset Seismic Bumps, caratterizzato da uno sbilanciamento estremo (93.4% vs 6.6%), AdaBoost finisce per dare troppo peso alla classe maggioritaria. Non supportando nativamente il bilanciamento dei pesi (*class_weight*) nella configurazione adottata, il modello ha fallito completamente nell'identificare i campioni della classe minoritaria, ottenendo un F1-score pari a zero.

Il Decision Tree si è distinto sul dataset più difficile (Seismic Bumps, $F1 = 0.258$), grazie all'utilizzo del parametro *class_weight*. In dataset

fortemente sbilanciati come questo, il parametro è cruciale perché pondera il calcolo dell'impurità, penalizzando maggiormente gli errori di classificazione sulle classi minoritarie. AdaBoost ha invece ottenuto il miglior risultato su Banknote ($F1 = 1.0$), un dataset con buona separabilità dove il boosting sequenziale riesce a raffinare progressivamente il decision boundary.

Per quanto riguarda la riduzione dimensionale, le tecniche di **Feature Selection** (Feature Importance, RFE, Mutual Information) hanno superato la PCA su due dataset su quattro. Questo risultato si spiega con la natura delle feature: quando l'informazione discriminativa è concentrata in poche variabili specifiche (come *variance* per Banknote o *a05* per Ionosphere), selezionarle direttamente preserva meglio la capacità predittiva rispetto alla proiezione PCA, che mescola tutte le feature.

Una eccezione è Cancer, dove la PCA ha ottenuto $F1 = 0.938$ contro $F1$ compreso tra 0.811 e 0.916 delle tecniche di selezione. In questo caso, l'elevata correlazione tra le 30 feature morfologiche (radius, perimeter e area sono geometricamente legate) fa sì che le prime componenti principali catturino efficacemente l'informazione discriminativa distribuita su più variabili.

Infine, anche nel dataset complesso Seismic Bumps, la PCA ha superato gli algoritmi di feature selection, suggerendo che la proiezione sulle componenti principali ha aiutato a ridurre il rumore meglio della semplice selezione delle variabili.

4.2. Scelte Ottimali e Motivazioni

L'uso dell' $F1$ -score come metrica principale si è rivelato fondamentale, in particolare per Seismic Bumps. Random forest ha ottenuto un'Accuracy di 0.934 ma un $F1$ -score di soli 0.056, poiché classificava quasi tutte le istanze come "non-hazardous". L' $F1$ -score, bilanciando Precision e Recall, ha correttamente identificato Decision Tree come miglior modello nonostante la sua Accuracy inferiore (0.911). Questo esempio dimostra come l'Accuracy possa essere fuorviante in presenza di sbilanciamento tra le classi.

La Stratified Cross-Validation ha garantito che ogni fold mantenesse la proporzione originale delle classi, aspetto critico per Seismic dove un fold non stratificato avrebbe potuto contenere pochissimi o nessun esempio della classe minoritaria, producendo stime inaffidabili.

Il bilanciamento dei pesi attraverso `class_weight='balanced'` ha permesso ai modelli di non ignorare completamente la classe minoritaria. L'importanza di questo accorgimento è evidente nel caso di AdaBoost, che non supportando nativamente questo parametro nella configurazione utilizzata, ha ottenuto $F1 = 0$ su Seismic, fallendo completamente nell'identificare qualsiasi evento hazardous.

4.3. Possibili Miglioramenti

Per il dataset Seismic Bumps, che ha mostrato le maggiori criticità, si potrebbero adottare tecniche di oversampling come SMOTE (Synthetic Minority Over-sampling Technique) per generare campioni sintetici della classe minoritaria e migliorare la capacità del modello di apprendere pattern dagli eventi hazardous. In alternativa, algoritmi come BalancedRandomForest o EasyEnsemble, progettati specificamente per dataset fortemente sbilanciati, potrebbero offrire risultati migliori. Un'altra strategia sarebbe l'ottimizzazione della soglia di decisione: invece di usare il valore standard 0.5, si potrebbe identificare sulla curva ROC la soglia che massimizza l' $F1$ -score o che meglio bilancia Precision e Recall in base alle esigenze applicative.

Per tutti i dataset, una ricerca sistematica degli iperparametri tramite Grid Search, Random Search o Bayesian Optimization potrebbe migliorare le prestazioni, specialmente per modelli sensibili alla configurazione come SVM (parametri C e gamma) e XGBoost (learning rate, max depth, n_estimators). La creazione di nuove feature derivate attraverso rapporti, interazioni o trasformazioni non lineari potrebbe inoltre catturare pattern non visibili nelle variabili originali.

Infine, invece di ridurre drasticamente a sole 2 feature per la visualizzazione, testare un range di dimensionalità (ad esempio 2, 5, 10 componenti) permetterebbe di identificare il compromesso ottimale tra interpretabilità e prestazioni.

4.4. Conclusioni

Il progetto ha analizzato quattro dataset di classificazione binaria mediante sette algoritmi di machine learning. AdaBoost emerge come l'algoritmo più robusto e versatile, rappresentando una scelta consigliata per problemi di classificazione con caratteristiche non note a priori e senza un estremo sbilanciamento. Le tecniche di Feature Selection risultano preferibili alla PCA quando l'informazione discriminativa è concentrata in poche variabili, mentre la PCA si dimostra più efficace in presenza di alta correlazione tra feature, come osservato nel dataset Cancer e Seismic Bumps.

Lo sbilanciamento delle classi richiede particolare attenzione: il caso Seismic Bumps ha mostrato come metriche tradizionali come l'Accuracy possano essere fuorvianti, e come tecniche di bilanciamento standard potrebbero non essere sufficienti per casi estremi. L'analisi esplorativa preliminare si conferma uno strumento prezioso per anticipare le difficoltà e guidare la scelta delle tecniche più appropriate. In conclusione, non esiste un algoritmo o una tecnica universalmente superiore: la scelta ottimale dipende dalle caratteristiche specifiche del dataset e deve essere validata empiricamente attraverso metriche appropriate al problema.