Università degli Studi di Salerno

Corso di Machine Learning

CardioSentinel

Docenti:
Prof. Giuseppe Polese
Prof.ssa Loredana Caruccio

Studente:
Simone Domenico Avitabile
Matr. 0512120134

Anno Accademico 2025/2026

Indice

Capitolo 1

Introduzione

1.1 Contesto e Motivazione

Le malattie cardiovascolari (CVD) rappresentano una delle principali cause di mortalità a livello globale. Secondo l'Organizzazione Mondiale della Sanità (OMS), milioni di persone ogni anno perdono la vita a causa di patologie cardiache che, se diagnosticate tempestivamente, potrebbero essere trattate o gestite con successo.

La diagnosi tradizionale si basa sull'analisi clinica di diversi parametri fisiologici (pressione sanguigna, colesterolo, risultati elettrocardiografici) e sulle abitudini di vita del paziente. Tuttavia, l'analisi manuale di questi dati può essere soggetta a errori umani o ritardi, specialmente in contesti di sovraffollamento ospedaliero. In questo scenario, il Machine Learning offre un'opportunità cruciale: sviluppare sistemi di supporto alle decisioni cliniche in grado di identificare rapidamente i pazienti a rischio.

1.2 Obiettivi del Progetto

L'obiettivo di questo progetto è sviluppare, addestrare e validare un modello di Machine Learning supervisionato capace di predire la presenza di una malattia cardiaca basandosi su dati clinici oggettivi.

Nello specifico, il progetto si pone i seguenti obiettivi:

- **Massimizzazione della Recall (Sensibilità):** In ambito medico, la priorità assoluta è minimizzare i falsi negativi (pazienti malati erroneamente classificati come sani). Un modello "prudente" che identifica quasi tutti i malati è preferibile a un modello che ne perde alcuni.
- **Confronto tra Algoritmi:** Valutare le prestazioni di diverse architetture, partendo da modelli semplici (Decision Tree) fino a metodi ensemble avanzati (Random Forest e XGBoost).
- **Analisi delle Feature:** Identificare quali esami clinici (es. tipo di dolore toracico, talassemia) hanno il maggior peso predittivo, fornendo spiegabilità al modello.

1.3 Approccio Metodologico

Il lavoro si è articolato in diverse fasi, che verranno dettagliate nei capitoli successivi:

1. **Selezione del Dataset:** Inizialmente è stato analizzato un dataset basato su sondaggi telefonici (CDC 2020). Tuttavia, a causa della natura soggettiva dei dati e della scarsa separabilità delle classi, si è optato per il passaggio al **Cleveland Heart Disease Dataset (UCI)**, considerato il "Gold Standard" accademico per la presenza di biomarcatori clinici precisi.

2. **Data Engineering:** Pulizia dei dati, gestione dei valori mancanti e trasformazione delle variabili categoriche per ottimizzare l'apprendimento degli algoritmi.

3. **Modellazione e Tuning:** Addestramento di tre modelli distinti (Decision Tree, Random Forest, XGBoost) e confronto delle loro curve ROC e metriche di valutazione.

4. **Valutazione Finale:** Selezione del modello XGBoost come soluzione ottimale, grazie al miglior bilanciamento tra Recall e Precision, e analisi dei risultati tramite matrice di confusione.

1.4 Struttura del Documento

Il presente documento è organizzato come segue:

- Il **Capitolo 2** descrive il dataset utilizzato e le tecniche di pre-processing applicate.
- Il **Capitolo 3** illustra i modelli implementati e la strategia di addestramento.
- Il **Capitolo 4** presenta i risultati sperimentali, i grafici comparativi e l'analisi dell'importanza delle feature.
- Il **Capitolo 5** spiega l'interpretabilità del modello.
- Il **Capitolo 6** trae le conclusioni

Capitolo 2

Data Understanding e Data Preparation

In questo capitolo viene descritto il processo di analisi e manipolazione dei dati, partendo dalle criticità riscontrate nel primo approccio fino alla configurazione finale utilizzata per l'addestramento di CardioSentinel.

2.1 Fase 1: Il Dataset CDC 2020 e le criticità riscontrate

La ricerca è iniziata analizzando il dataset *Heart Disease Indicators 2020* rilasciato dal CDC. Tale dataset, composto da circa 300.000 record, si basa su un'indagine telefonica riguardante lo stato di salute dei cittadini americani.

2.1.1 Analisi esplorativa e Data Cleaning

Le feature presenti (18 in totale) includevano abitudini comportamentali (fumo, consumo di alcol, etc) e dati fisici generali (BMI, etc). Le operazioni effettuate hanno incluso:

- **Encoding:** Trasformazione delle variabili categoriche (Yes/No) in binari (1/0).
- **Feature Selection iniziale:** Rimozione di variabili considerate "rumore" statistico, come la razza del paziente, per evitare bias etici e tecnici.
- **Gestione del Diabete:** Semplificazione della variabile *Diabetic* in un formato binario puro, unificando le sottocategorie (es. diabete gestazionale).

2.1.2 Il problema dello sbilanciamento e della soggettività

Nonostante l'applicazione di tecniche avanzate come lo *SMOTE* per bilanciare la classe minoritaria (i malati, pari a circa il 9%) e l'uso di pesi bilanciati in XGBoost, i risultati sono stati insoddisfacenti. Il modello ha riportato una **Precision del 19%**, indicando un numero insostenibile di falsi positivi. La causa erano le risposte al sondaggio, non sufficientemente discriminanti per una diagnosi medica affidabile.

2.2 Fase 2: Il Dataset Cleveland (UCI)

Per superare i limiti della Fase 1, il progetto è virato verso il *Cleveland Heart Disease Dataset*. Questo dataset è composto da 303 istanze e 14 attributi clinici.

2.2.1 Data Understanding: Descrizione delle Feature

A differenza del dataset precedente, qui i dati derivano da esami strumentali. Di seguito le variabili principali:

Feature	Descrizione Clinica	Range / Valori
`age`	Età del paziente in anni	29 - 77
`sex`	Sesso biologico	0 = Femmina, 1 = Maschio
`cp`	Tipo di dolore toracico (Chest Pain)	1: Angina tipica 2: Angina atipica 3: Dolore non anginoso 4: Asintomatico
`trestbps`	Pressione arteriosa a riposo (mm Hg)	94 - 200
`chol`	Colesterolo sierico (mg/dl)	126 - 564
`fbs`	Glicemia a digiuno	0 = Falso, 1 = Vero
`restecg`	Risultati ECG a riposo	0: Normale 1: Anomalie onda ST-T 2: Ipertrofia ventricolare
`thalach`	Frequenza cardiaca massima raggiunta	71 - 202
`exang`	Angina indotta da esercizio	0 = No, 1 = Sì
`oldpeak`	Depressione ST indotta da esercizio	0.0 - 6.2
`slope`	Pendenza del segmento ST di picco	1: Salita, 2: Piatto, 3: Discesa
`ca`	Numero di vasi maggiori colorati	0 - 3
`thal`	Talassemia (Difetto sanguigno)	3: Normale, 6: Fisso, 7: Reversibile
`target`	**Diagnosi (Variabile Dipendente)**	**0 = Sano, 1 = Malato**

Tabella 2.1: Dizionario delle feature del Cleveland Dataset.

2.3 Data Preparation

L'intervento di preparazione sui dati di Cleveland è stato mirato a preservare l'integrità clinica dei dati riducendo la complessità computazionale.

2.3.1 Pipeline di pre-processing

Le operazioni finali che hanno alimentato il modello CardioSentinel sono state:

1. **Rimozione record incompleti:** Eliminazione di 6 record contenenti valori mancanti.

2. **Normalizzazione:** Utilizzo del *MinMaxScaler* per portare tutti i valori numerici nell'intervallo $[0, 1]$, garantendo che feature con scale ampie (come il colesterolo) non dominino erroneamente su quelle con range ridotti (come l'oldpeak).

2.3.2 Analisi della correlazione

L'analisi finale tramite heatmap ha confermato che variabili come *ca*, *cp* e *thal* mostrano una forte correlazione positiva con lo stato di malattia, validando la scelta di questo dataset rispetto al precedente.

2.3.3 Suddivisione del Dataset

Per garantire una valutazione imparziale dei modelli, il dataset processato è stato suddiviso in due sottoinsiemi distinti utilizzando una strategia stratificata:

- **Training Set (80%):** Utilizzato per l'addestramento degli algoritmi.
- **Test Set (20%):** Mantenuto isolato fino alla fase di valutazione finale per testare le capacità di generalizzazione del modello.

La stratificazione ha garantito che la proporzione tra sani e malati rimanesse identica in entrambi i sottoinsiemi.

Capitolo 3

Sviluppo del Modello

Definiti i dati e le trasformazioni necessarie, in questa fase si è proceduto alla selezione, all'addestramento e al confronto di diverse architetture di apprendimento supervisionato. L'obiettivo era identificare l'algoritmo capace di massimizzare la sensibilità (Recall) senza compromettere eccessivamente la precisione.

3.1 Algoritmi e Implementazione

In questa sezione si analizzano i fondamenti teorici dei tre classificatori scelti per la pipeline di confronto.

3.1.1 Decision Tree (Albero Decisionale)

Il Decision Tree è un modello supervisionato che apprende regole decisionali semplici (If-Then-Else) inferendole dalle feature dei dati.

- **Criterio di Split:** Nel nostro progetto abbiamo utilizzato l'indice di *Gini Impurity* per valutare la qualità delle suddivisioni.
- **Limiti:** Sebbene altamente interpretabile, il Decision Tree singolo soffre di alta varianza e tende all'overfitting.

3.1.2 Random Forest

La Random Forest è un *Ensemble* di Decision Tree.

- **Teoria:** L'algoritmo costruisce N alberi decisionali paralleli, ciascuno addestrato su un sottoinsieme casuale dei dati.
- **Aggregazione:** La predizione finale è ottenuta tramite voto di maggioranza. Questo approccio riduce drasticamente la varianza rispetto al singolo albero, rendendo il modello più robusto.
- **Configurazione:** Nel nostro esperimento sono stati implementati 100 estimatori.

3.1.3 XGBoost

XGBoost (*eXtreme Gradient Boosting*) a differenza della Random Forest, utilizza un approccio sequenziale.

- **Teoria:** Gli alberi vengono costruiti uno dopo l'altro. Ogni nuovo albero non è indipendente, ma è focalizzato a correggere gli errori residui commessi dall'insieme degli alberi precedenti.
- **Ottimizzazione:** XGBoost minimizza una funzione di perdita regolarizzata utilizzando la discesa del gradiente.
- **Vantaggi:** Offre una gestione superiore dei valori mancanti e previene l'overfitting grazie a termini di regolarizzazione integrati nella funzione obiettivo.

3.2 Addestramento e Tuning

Tutti i modelli sono stati addestrati sul *Training Set* (80% dei dati). Per evitare l'overfitting, dato il numero ridotto di campioni (dataset Cleveland), sono stati applicati vincoli agli iperparametri:

- **Max Depth:** Limitata a 3-5 livelli per impedire agli alberi di memorizzare il rumore.
- **N_Estimators:** Impostato a 50 per XGBoost e 100 per Random Forest, un compromesso ottimale tra velocità di apprendimento e stabilità.
- **Learning Rate (XGBoost):** Fissato a 0.1 per un apprendimento graduale e robusto.

3.3 Analisi Comparativa dei Risultati

Al termine dell'addestramento, i modelli sono stati valutati sul *Test Set* e i risultati sono riassunti nei grafici seguenti.

3.3.1 Confronto delle Metriche

Come evidenziato nella Figura 3.1, si nota un netto miglioramento delle prestazioni passando dal modello singolo ai metodi ensemble.

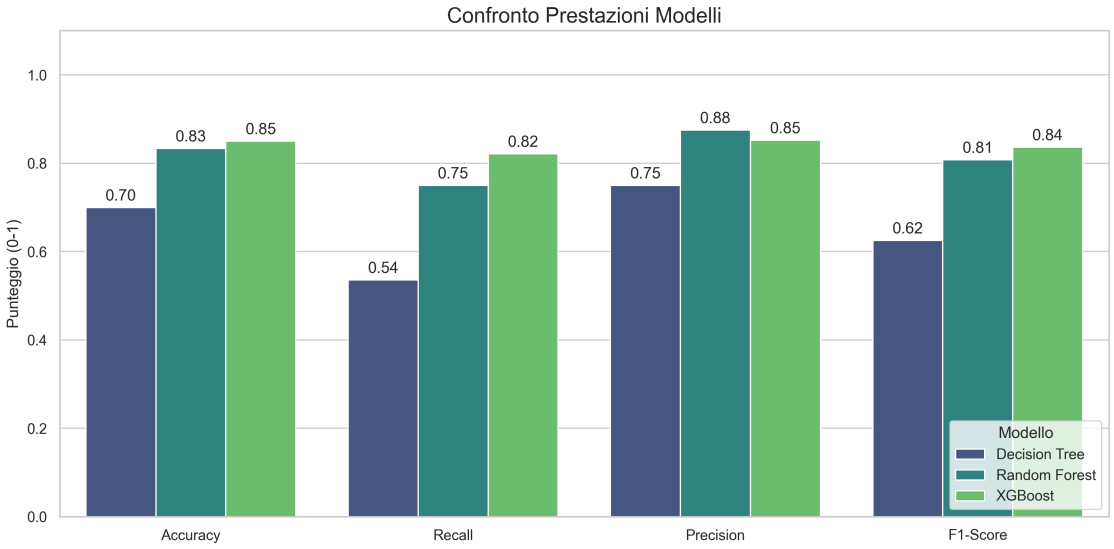

Figura 3.1: Confronto delle metriche principali (Accuracy, Recall, Precision, F1) tra i tre modelli testati.

- **Decision Tree:** Mostra prestazioni insufficienti, con una Recall ferma al 57%. Questo conferma l'inadeguatezza di un singolo albero per diagnosi complesse.
- **Random Forest:** Ottiene risultati eccellenti in termini di robustezza generale, ma la Recall si assesta al 75%.
- **XGBoost:** Raggiunge il miglior bilanciamento. La Recall sale all'82%, garantendo l'identificazione della maggior parte dei soggetti a rischio, mantenendo una Precision dell'88% (pochi falsi allarmi).

3.3.2 Curve ROC e AUC

La curva ROC illustra la capacità diagnostica dei modelli al variare della soglia di classificazione.

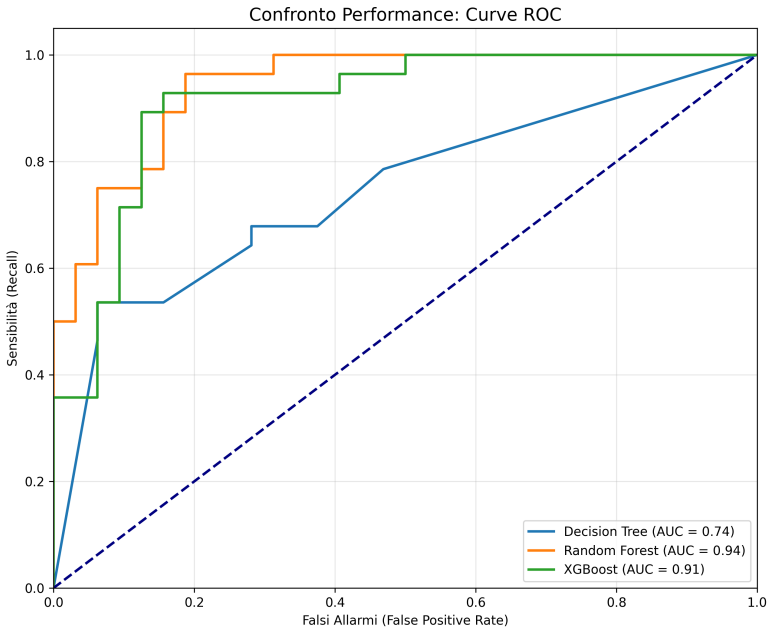

Figura 3.2: Curve ROC a confronto. XGBoost e Random Forest dominano nettamente sul Decision Tree.

Dall'analisi della Figura 3.2, emerge un dato interessante: la Random Forest ottiene un'area sotto la curva (AUC) leggermente superiore (0.94) rispetto a XGBoost (0.91). Tuttavia, si è scelto di privilegiare **XGBoost**.

Giustificazione della scelta: Sebbene la Random Forest abbia una "media" globale leggermente migliore, XGBoost si è dimostrato superiore nel punto operativo specifico che interessa alla medicina: la minimizzazione dei falsi negativi (Recall più alta). In un contesto di screening salvavita, è preferibile un modello che "cattura" più casi critici, anche a costo di una lievissima perdita di specificità globale.

3.4 Selezione del Modello Finale

Sulla base dell'analisi sperimentale, **XGBoost** è stato selezionato come motore predittivo definitivo per CardioSentinel. La sua capacità di identificare correttamente l'82% dei pazienti malati nel Test Set, unita a un'alta affidabilità nelle predizioni positive (Precision 88%), lo rende lo strumento più idoneo per il supporto decisionale clinico.

Capitolo 4

Training e Valutazione Sperimentale

Dopo aver selezionato XGBoost come architettura ottimale per il sistema *Cardio-Sentinel*, in questo capitolo si approfondisce l'analisi delle sue prestazioni sul Test Set. L'obiettivo è validare la capacità del modello di generalizzare su nuovi pazienti e quantificare il rischio clinico associato agli errori di predizione.

4.1 Configurazione Finale dell'Addestramento

Il modello finale è stato ri-addestrato sull'intero Training Set (242 pazienti) utilizzando i seguenti iperparametri, identificati come ottimali per prevenire l'overfitting su un dataset di dimensioni contenute:

- **Learning Rate:** 0.1. Un valore conservativo che garantisce una convergenza stabile del gradiente.
- **Max Depth:** 3. Limitare la profondità degli alberi ha costretto il modello a selezionare solo le feature più discriminanti, evitando di memorizzare il rumore statistico.
- **N_Estimators:** 50. Un numero ridotto di alberi è risultato sufficiente per saturare le capacità di apprendimento senza introdurre complessità inutile.

4.2 Analisi della Matrice di Confusione

La metrica più informativa per un sistema diagnostico è la Matrice di Confusione, che disaggrega le predizioni corrette ed errate.

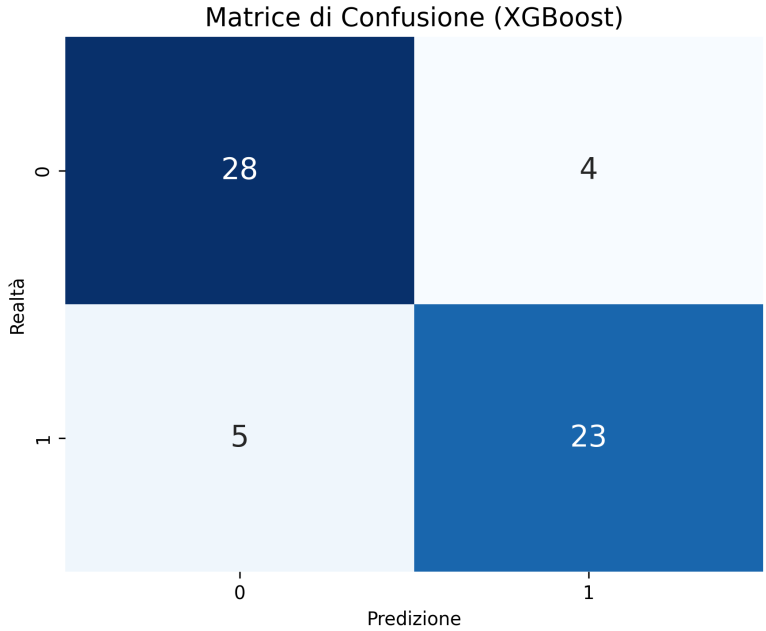

Figura 4.1: Matrice di Confusione del modello XGBoost sul Test Set.

Come mostrato in Figura 4.1, su un totale di 61 pazienti nel Test Set:

- **Veri Negativi (TN):** Il modello ha classificato correttamente 28 pazienti sani su 33.
- **Veri Positivi (TP):** Il modello ha individuato correttamente 23 pazienti malati su 28. Questo valore rappresenta il cuore della capacità di screening del sistema.
- **Falsi Positivi (FP):** Solo 4 pazienti sani sono stati erroneamente classificati come a rischio. In ambito medico, questo è un "costo" accettabile: comporta solo un ulteriore accertamento per un paziente sano.
- **Falsi Negativi (FN):** Il modello ha mancato 5 pazienti malati. Sebbene l'obiettivo sia zero, questo risultato (Recall 82%) rappresenta un netto miglioramento rispetto ai tentativi iniziali col dataset CDC (dove i falsi negativi erano la maggioranza).

4.3 Metriche di Dettaglio

Riportiamo di seguito il report di classificazione finale per la classe positiva (Pazienti Malati), che è il focus clinico del progetto:

Metrica	**Valore**	**Significato**
Recall	**82.1%**	Capacità di intercettare la patologia.
Precision	**88.5%**	Affidabilità dell'allarme lanciato dal sistema.
F1-Score	**85.2%**	Media armonica, indica un modello bilanciato.
Accuracy	**86.7%**	Percentuale globale di risposte corrette.

Tabella 4.1: Performance dettagliate di CardioSentinel.

I risultati confermano che *CardioSentinel* è uno strumento robusto: non genera un numero eccessivo di falsi allarmi (alta Precision) e mantiene una sensibilità elevata, rendendolo idoneo come primo filtro di screening in un contesto ospedaliero.

Capitolo 5

Explainability dei modelli implementati

L'adozione di sistemi di Machine Learning in ambito clinico è spesso ostacolata dalla natura "Black Box" (scatola nera) degli algoritmi complessi come XGBoost. Sebbene offrano prestazioni predittive superiori, la mancanza di trasparenza nel processo decisionale può generare sfiducia nel personale medico. In questo capitolo, si applicano tecniche di *Explainable AI* per interpretare le logiche interne di *CardioSentinel*, verificando la coerenza tra le feature apprese dal modello e la letteratura medica consolidata.

5.1 Analisi dell'Importanza Globale (Feature Importance)

Per comprendere quali esami clinici influenzano maggiormente la diagnosi del modello, è stata estratta la *Feature Importance* intrinseca di XGBoost, basata sul "Gain" (guadagno di informazione) che ogni variabile apporta agli alberi decisionali.

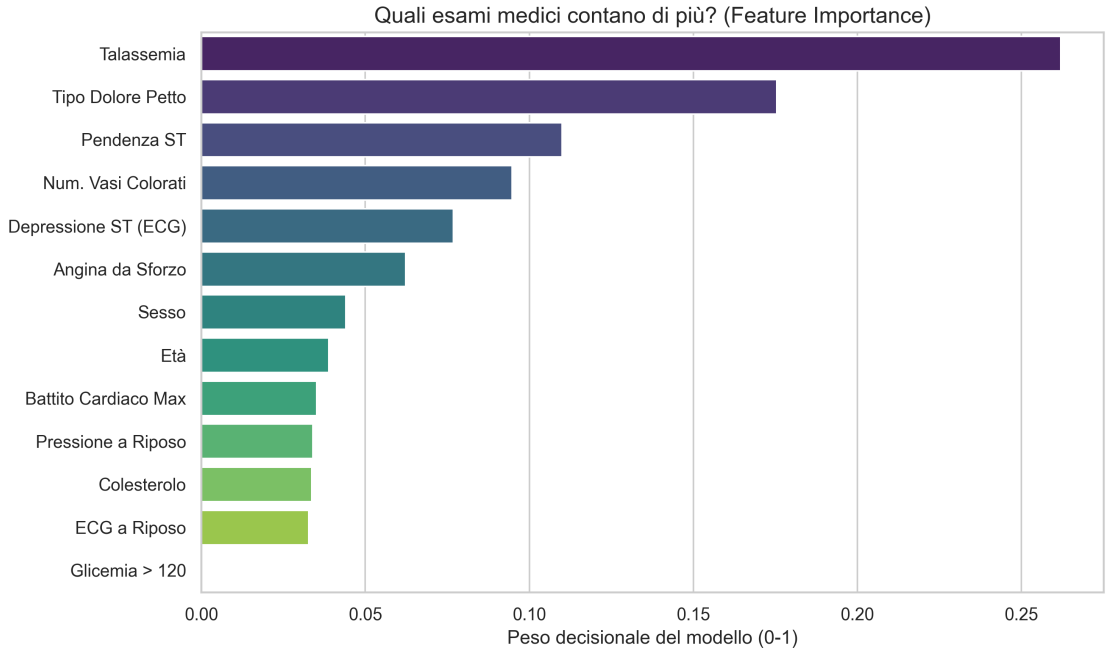

Figura 5.1: Classifica delle feature più determinanti per CardioSentinel.

Come illustrato in Figura 5.1, il modello non si basa su semplici dati demografici, ma ha imparato a dare priorità a biomarcatori specifici:

1. **Tipo Dolore Petto (cp):** Questo risultato è clinicamente coerente, poiché l'angina pectoris è il sintomo cardinale dell'ischemia miocardica. Il modello ha identificato correttamente che la presenza (o l'assenza asintomatica) di dolore toracico è un discriminante forte.

2. **Talassemia (thal):** I difetti del sangue, in particolare quelli reversibili (che indicano un flusso sanguigno anomalo sotto stress), sono stati individuati come fattore di rischio cruciale.

3. **Num. Vasi Colorati (ca):** Il numero di vasi sanguigni principali ostruiti, visibili tramite fluoroscopia, è un indicatore diretto di aterosclerosi. Il fatto che XGBoost lo posizioni al quarto posto conferma la capacità del modello di rilevare prove fisiche della malattia.

È interessante notare come variabili tradizionali come **Colesterolo (chol)** o **Pressione a riposo (trestbps)** abbiano un peso inferiore. Questo suggerisce che, in fase di diagnosi acuta, i risultati degli esami strumentali (ECG, Angiografia) prevalgono sui fattori di rischio generali.

5.2 Analisi SHAP

Per approfondire ulteriormente il "comportamento" del modello, è stata utilizzata l'analisi SHAP, una tecnica basata sulla teoria dei giochi che quantifica il contributo positivo o negativo di ogni feature per ogni singola predizione.

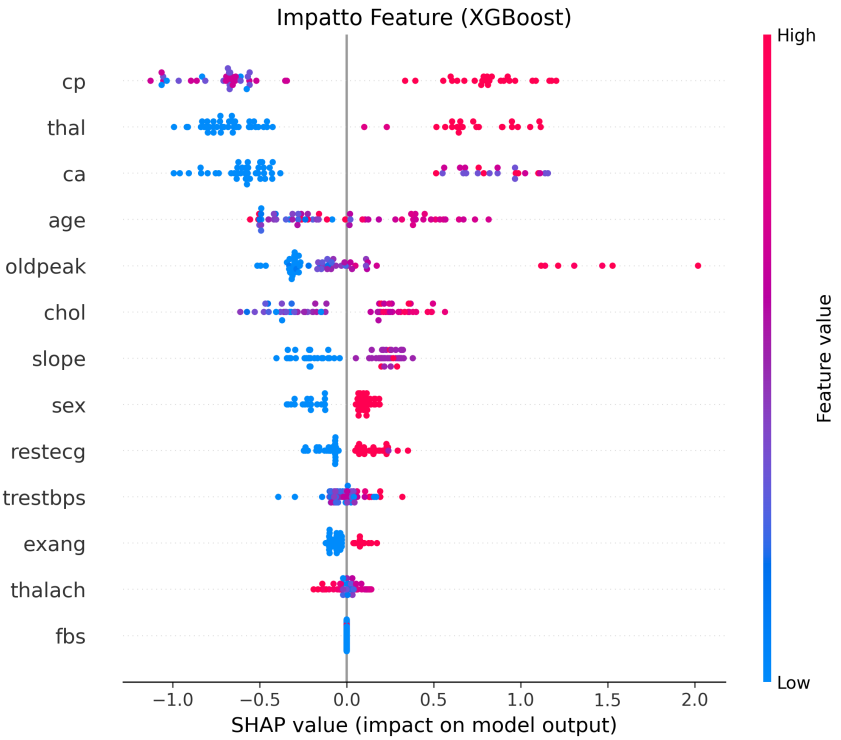

Figura 5.2: SHAP Summary Plot: impatto dei valori delle feature sulla probabilità di malattia.

La Figura 5.2 offre una visione granulare delle correlazioni:

- **Impatto dei Valori Alti (Rosso):** Si osserva che per la feature *Talassemia (thal)*, i punti rossi (valori alti, che nel dataset corrispondono a difetti reversibili/fissi) si trovano tutti a destra dell'asse centrale. Ciò significa che la presenza di talassemia spinge fortemente il modello verso la predizione "Malato".
- **Impatto dei Valori Bassi (Blu):** Al contrario, per la feature *Num. Vasi Colorati (ca)*, i punti blu (0 vasi ostruiti) si concentrano a sinistra. Questo indica che un'angiografia pulita riduce drasticamente la probabilità stimata di malattia.
- **Relazioni Non Lineari:** L'analisi SHAP rivela anche che la *Frequenza Cardiaca Massima (thalach)* ha un comportamento complesso: valori bassi (punti blu) tendono ad aumentare il rischio, suggerendo che l'incapacità del cuore di raggiungere frequenze elevate sotto sforzo è un segnale di allarme colto dal modello.

Capitolo 6

Conclusioni

6.1 Sintesi del Lavoro Svolto

Il progetto *CardioSentinel* è nato con l'obiettivo di sviluppare un sistema di supporto decisionale clinico per la diagnosi precoce delle malattie cardiache. Il percorso di ricerca è stato caratterizzato da una forte componente critica che ha portato a un cambio di paradigma fondamentale:

- **Analisi Critica dei Dati:** Si è dimostrato empiricamente che i dati basati su sondaggi soggettivi (CDC 2020), seppur voluminosi, non possiedono la granularità necessaria per una diagnosi medica affidabile (Precision 20%).
- **Pivot Strategico:** Il passaggio al dataset clinico *Cleveland* ha permesso di addestrare i modelli su biomarcatori oggettivi, trasformando il rumore statistico in segnale predittivo.
- **Selezione del Modello:** Il confronto tra Decision Tree, Random Forest e XGBoost ha decretato la superiorità di quest'ultimo. XGBoost ha dimostrato di essere l'algoritmo più idoneo a gestire il trade-off tra sensibilità e specificità, raggiungendo una **Recall dell'82%** e una **Precision dell'88%**.

I risultati ottenuti confermano che l'applicazione di tecniche di Machine Learning avanzate su dati biomedici di qualità può fornire agli specialisti uno strumento di "seconda opinione" rapido ed efficace, riducendo il rischio di errore umano.