UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Università degli Studi di Modena e Reggio Emilia

Dipartimento di Scienze Fisiche, Informatiche e Matematiche
Corso di Laurea Magistrale in Matematica

ANALISI COMPARATIVA DI METODI DI EXPLAINABLE AI PER L'IDENTIFICAZIONE DELLE VARIABILI BIOLOGICHE RILEVANTI NELLA RISPOSTA ALL'IMMUNOTERAPIA

Relatrice:
Prof.ssa Giorgia Franchini

Correlatori:
Dott. Giovanni Cappelletti
Dott. Matteo Lombardi

Tesi di Laurea di:
Beatrice Bonicelli

Anno Accademico 2024/2025

Abstract

Recenti studi hanno evidenziato come la composizione del microbiota intestinale possa influenzare la risposta terapeutica nei pazienti oncologici. L'identificazione di potenziali biomarcatori microbici predittivi di tale risposta rappresenta una sfida scientifica di grande interesse.

Il presente lavoro di tesi, che si pone come estensione e approfondimento di una pipeline preesistente finalizzata alla valutazione predittiva della risposta terapeutica, ha come obiettivo l'identificazione di possibili variabili biologiche maggiormente rilevanti nella risposta all'immunoterapia.

Il dataset analizzato comprende 569 campioni caratterizzati da 4630 specie microbiche, ognuna delle quali può essere ricondotta ad una catena tassonomica gerarchica del tipo: *f_Family* $\rightarrow$ *g_Genus* $\rightarrow$ *s_Species*.

La metodologia adottata si basa sul confronto di tre modelli di machine learning (Random Forest, Extra Trees e XGBoost) e sull'applicazione di tecniche di Explainable AI per interpretarne il comportamento.

In particolare, al fine di analizzare il contributo delle variabili biologiche e rendere più comprensibili i modelli considerati "black box", è stato adottato un framework di interpretabilità su tre livelli: SHAP (SHapley Additive exPlanations) per la stima del contributo delle feature, la Permutation Feature Importance per la valutazione della loro importanza e la Feature Ablation per l'analisi della rilevanza strutturale.

I risultati mostrano che alcuni taxa microbici emergono come potenziali biomarcatori della risposta all'immunoterapia, risultando consistenti tra diversi modelli e tecniche di interpretabilità. In particolare, l'analisi evidenzia come la risposta terapeutica sia associata a pattern complessi di abbondanza microbica piuttosto che alla presenza di singole specie isolate.

Indice

Introduzione

Negli ultimi anni, numerosi studi ([20], [27], [31]) hanno evidenziato come la composizione del microbiota intestinale possa influenzare la risposta dei pazienti oncologici trattati con immunoterapia. Questo effetto deriva dalla costante interazione tra microrganismi e barriera epiteliale intestinale, che avviene sia tramite contatti diretti, sia attraverso i numerosi metaboliti prodotti a partire dalla dieta. Tuttavia, non sono ancora del tutto noti quali specifici aspetti del microbiota siano correlati a una maggiore efficacia terapeutica, rendendo questo un campo di studio affascinante e in rapida evoluzione.

In questo contesto, il machine learning si è affermato come uno strumento chiave per l'analisi di grandi quantità di dati, offrendo la possibilità di individuare pattern complessi e relazioni non lineari difficilmente rilevabili con metodi statistici tradizionali. In ambito biomedico, tali tecniche risultano particolarmente adatte all'analisi di dataset ad alta dimensionalità, come quelli derivanti dagli studi metagenomici.

Tuttavia, molti modelli ad alte prestazioni, come le reti neurali profonde e gli ensemble methods, sono spesso considerati delle "black-box": pur fornendo predizioni accurate, risultano difficili da interpretare. In contesti sensibili come quello biomedico, questa mancanza di trasparenza rappresenta un limite rilevante, poiché la comprensione dei meccanismi decisionali è essenziale per garantire affidabilità, validazione scientifica e potenziale applicabilità clinica.

Per affrontare questa problematica, negli ultimi anni si è sviluppato il campo della Explainable Artificial Intelligence (XAI), che mira a rendere i modelli di machine learning più interpretabili. Tra i metodi più diffusi vi è SHAP (SHapley Additive exPlanations) [33], basato sulla teoria dei valori di Shapley, che consente di attribuire a ciascuna variabile un contributo quantitativo alla predizione del modello. Accanto a SHAP, tecniche come la Permutation Feature Importance e l'Ablation Feature Importance permettono di valutare l'impatto delle singole variabili sulle prestazioni complessive del modello, offrendo strumenti complementari per un'interpretazione più completa. L'utilizzo congiunto di queste metodologie consente non solo di ottenere modelli predittivi accurati, ma anche di comprenderne il comportamento, individuando le variabili maggiormente rilevanti. Questo aspetto è cruciale in ambito biomedico, dove l'interpretabilità rappresenta un requisito fondamentale per la validazione dei risultati.

Il presente lavoro di tesi si propone di applicare tecniche di Explainable Artificial Intelligence a dataset metagenomici, con l'obiettivo di identificare possibili biomarcatori predittivi della risposta all'immunoterapia, garantendo al contempo un adeguato livello di trasparenza ai fini di una potenziale validazione clinica.

Nel primo capitolo viene introdotto il machine learning, con particolare attenzione al problema delle "black-box" e alla necessità di sviluppare metodi in grado di superarne i limiti interpretativi. In tale contesto, vengono presentati SHAP e i principali algoritmi di spiegazione (KernelExplainer, TreeExplainer e DeepExplainer), insieme ai metodi di Permutation e Ablation Feature Importance.

Nel secondo capitolo viene descritto il dataset utilizzato e la metodologia adottata, includendo le fasi di preprocessing, preparazione dei dati e costruzione dei modelli.

Il terzo capitolo è dedicato all'analisi dei risultati e alla loro discussione, con particolare attenzione alle interpretazioni fornite dalle tecniche di XAI.

Infine, sono riportate le conclusioni del lavoro, evidenziando i principali risultati ottenuti e possibili sviluppi futuri.

Il Dataset Analizzato

Il dataset analizzato in questo progetto di tesi è stato fornito dal Laboratorio di Biotecnologie Microbiche dell'Università di Modena e Reggio Emilia e comprende 569 metagenomi del microbiota intestinale di pazienti oncologici trattati con immunoterapia.

Questi 569 campioni sono stati recuperati da diversi studi pubblicamente disponibili nel database NCBI SRA (Tabella 1). Tutti i campioni sono stati processati utilizzando i software Kraken 2 [55] e Bracken [30], supportati dal database Unified Human Gastrointestinal Genome v2.0.1 (UHGG [4]). Questi strumenti hanno consentito di ottenere, per ciascun metagenoma, un profilo di composizione microbica espresso in termini di abbondanza relativa percentuale.

Bioproject	**R**	**NR**	**Tumore**	**Paese**	**Referenza**
PRJEB43119	64	71	MM	Regno Unito, Olanda, Spagna	[26]
PRJNA866654	8	4	NSCLC	Nord America	[28]
PRJEB22863	26	56	RCC	Francia	[49]
PRJNA751792	70	113	NSCLC	Francia	[15]
PRJNA672867	33	14	MM	Nord America	[14]
PRJNA399742	15	12	MM	Nord America	[35]
PRJNA397906	19	15	MM	Nord America	[17]
PRJNA762360	15	7	MM	Nord America	[36]
PRJNA541981	12	15	MM	Nord America	[45]

Tabella 1: Elenco degli studi da cui sono stati recuperati i 569 metagenomi analizzati. Per ciascuno vengono riportati il codice identificativo (Bioproject), il numero di pazienti Responder (R) e Non-Responder (NR), il tipo di tumore, il Paese di provenienza dei campioni e la relativa pubblicazione.

Capitolo 1

L'Intelligenza Artificiale Spiegabile

L'Intelligenza Artificiale rappresenta oggi uno dei principali ambiti di innovazione tecnologica, con applicazioni sempre più diffuse in numerosi settori. L'evoluzione dei modelli computazionali e la crescente disponibilità di dati hanno contribuito allo sviluppo di sistemi sempre più sofisticati e performanti.

Tuttavia, l'aumento della complessità di tali modelli pone nuove sfide, in particolare per quanto riguarda l'interpretazione e la comprensione dei processi decisionali.

In questo capitolo verranno introdotti i concetti fondamentali alla base del Machine Learning e sarà approfondito il tema dell'Explainable Artificial Intelligence, con l'obiettivo di analizzare i principali metodi utilizzati per rendere i modelli più trasparenti e interpretabili.

1.1 Il Machine Learning e i suoi ambiti applicativi

Nel panorama attuale, caratterizzato da una crescita esponenziale del volume di dati provenienti da diversi settori, risulta fondamentale la conoscenza del **Machine Learning (ML)** (o *Apprendimento Automatico*).

Questa disciplina consente infatti di analizzare in modo intelligente grandi quantità di dati e di sviluppare applicazioni automatizzate sempre più sofisticate.

In questo contesto, il machine learning si pone come il paradigma di riferimento per lo sviluppo di sistemi predittivi e decisionali autonomi.

L'integrazione di algoritmi di apprendimento automatico non rappresenta più una semplice ottimizzazione tecnica, ma costituisce il fondamento logico per l'evoluzione dell'Intelligenza Artificiale nelle sue numerose applicazioni [51].

Esistono principalmente quattro categorie diverse di algoritmi di apprendimento automatico, tra cui l'apprendimento supervisionato, non supervisionato, semi-supervisionato e per rinforzo [9].

La Figura 1.1 ne fornisce una rappresentazione sintetica.

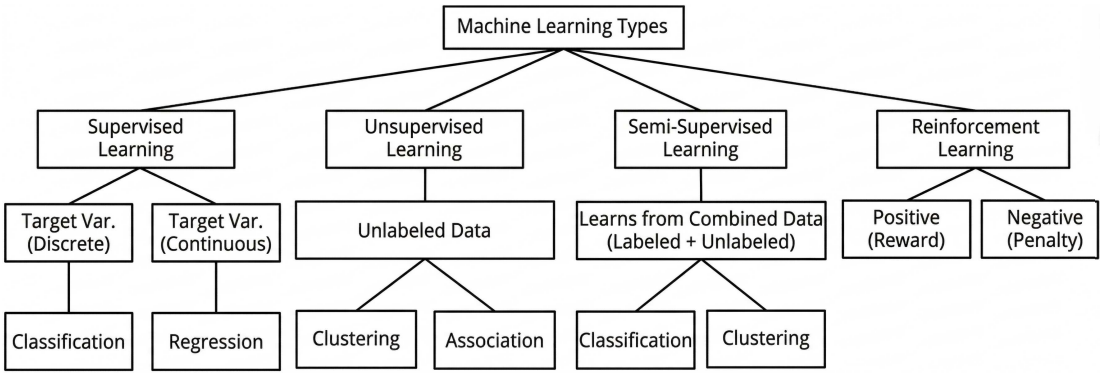

Figura 1.1: Schema riassuntivo dei vari tipi di Machine Learning.

Nello specifico:

- L'**apprendimento supervisionato** utilizza dati di addestramento etichettati per apprendere una funzione che mappa gli input agli output desiderati. In questo caso, l'output è noto a priori e il modello viene addestrato minimizzando l'errore tra le predizioni e i valori reali.

  I principali problemi affrontati sono:

  - *Classificazione*: si vuole assegnare a ciascun dato in ingresso una categoria predefinita, scegliendo da un insieme finito di classi possibili. Può essere binaria (due classi) o multiclasse (tre o più classi).
  - *Regressione*: prevede un output continuo o un valore numerico basato sulle caratteristiche di input.

  Tra gli algoritmi di apprendimento supervisionato più comuni troviamo: Regressione Lineare, Regressione Polinomiale, K-Nearest Neighbors (KNN), Naive Bayes, Alberi Decisionali. Algoritmi più avanzati includono reti neurali, Support Vector Machines (SVM) e metodi di ensemble.

- L'**apprendimento non supervisionato**, invece, utilizza dati non etichettati per individuare pattern nascosti e apprendere rappresentazioni strutturate dei dati. A differenza del caso supervisionato, non esiste una "risposta corretta" predefinita.

  Tra gli algoritmi più diffusi troviamo: Fuzzy C-means, K-means, Clustering gerarchico, Minimi quadrati parziali (PLS). Questi metodi sono utilizzati, ad esempio, per rilevare anomalie o ridurre la dimensionalità dei dati.

- L'**apprendimento semi-supervisionato** rappresenta un approccio intermedio in cui solo una parte dei dati è etichettata. Il modello sfrutta sia le informazioni supervisionate sia la struttura dei dati non etichettati per migliorare le prestazioni.

- L'**apprendimento per rinforzo** si basa sull'interazione tra un *agente* e un *ambiente*. L'agente apprende attraverso un processo iterativo di tentativi ed errori, ricevendo ricompense o penalità in base alle azioni intraprese. L'obiettivo è massimizzare la ricompensa cumulativa nel tempo.

Il processo tipico di machine learning consiste nello sviluppo di modelli statistici capaci di apprendere dai dati per effettuare predizioni o supportare decisioni. Si inizia con la raccolta e la valutazione di dati di qualità, eventualmente etichettati, seguita dalla selezione dell'algoritmo più adatto, che può essere supervisionato, non supervisionato o semi-supervisionato, in base al tipo di apprendimento desiderato. Successivamente, i dati vengono preparati tramite pulizia, trasformazioni e gestione di valori mancanti o anomali, in modo da garantire l'idoneità all'addestramento.

Il modello viene quindi addestrato attraverso iterazioni sui dati, ottimizzandone le prestazioni, e valutato su dati non utilizzati in precedenza per misurarne accuratezza e affidabilità.

Infine, il modello viene distribuito in produzione, con monitoraggio continuo delle prestazioni e interventi di ottimizzazione per correggere eventuali errori o bias, garantendo così il suo funzionamento efficace nel tempo [11].

Il principio centrale del machine learning è che, ottimizzando le prestazioni di un modello su un insieme di dati rappresentativi del problema reale, esso sia in grado di generalizzare e fornire previsioni accurate su dati mai osservati in precedenza.

Il processo di addestramento rappresenta quindi un mezzo per raggiungere l'obiettivo fondamentale del machine learning: la *generalizzazione*. In particolare, un modello ben addestrato è in grado di applicare le conoscenze apprese per produrre output corretti in contesti reali; questa fase operativa è comunemente definita *inferenza*.

Il funzionamento generale di un modello di apprendimento automatico può essere schematizzato come segue:

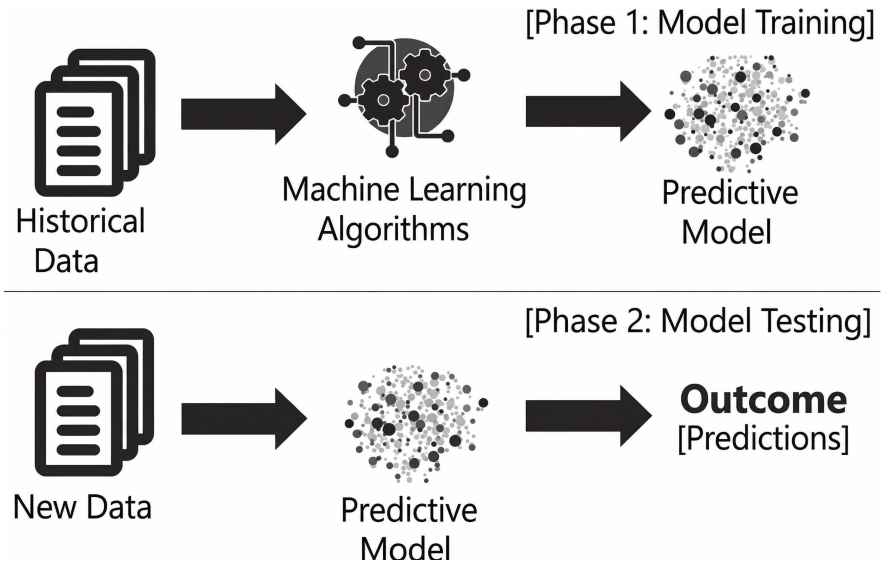

Figura 1.2: Struttura generale di un modello predittivo basato sul Machine Learning considerando sia la fase di addestramento che quella di test.

Il machine learning trova applicazione in numerosi settori. In particolare, l'analisi predittiva consente di sfruttare le relazioni tra variabili osservate per prevedere eventi futuri, supportando il processo decisionale. Questo approccio è ampiamente utilizzato, ad esempio, nell'ottimizzazione della logistica e del magazzinaggio.

Più in generale, le tecniche di apprendimento automatico sono impiegate in ambiti quali commercio elettronico, telecomunicazioni, servizi bancari e finanziari, sanità, marketing, trasporti e social network.

Ulteriori applicazioni includono la bioinformatica, la chemoinformatica, l'analisi delle reti, la classificazione di sequenze biologiche, la robotica e l'ingegneria avanzata.

In particolare nel settore biomedico, il machine learning sta aprendo nuove prospettive per la diagnosi, il trattamento e la prevenzione delle malattie. Questa tecnologia viene utilizzata in particolare per l'analisi di immagini mediche (radiografie, TAC, risonanze magnetiche), per l'interpretazione di dati genomici complessi e per la gestione di grandi database clinici.

Nel complesso, il machine learning rappresenta oggi uno strumento fondamentale per l'estrazione di conoscenza dai dati e per il supporto a processi decisionali complessi in un'ampia varietà di contesti applicativi.

Tuttavia, l'elevata complessità dei modelli di machine learning più avanzati rende spesso difficile comprendere il processo decisionale sottostante, sollevando importanti questioni legate alla trasparenza e all'interpretabilità, che ad oggi identifica una delle sfide principali nell'ambito dell'intelligenza artificiale.

1.2 L'intelligenza Artificiale spiegabile

Con l'aumento dell'impiego di modelli di Intelligenza Artificiale, l'**Explainable Artificial Intelligence (XAI)** è diventata fondamentale per garantire trasparenza e aumentare l'affidabilità nei modelli di machine learning.

La XAI comprende un insieme di metodi e processi, utili a rendere comprensibili le decisioni prese dai modelli di machine learning, contribuendo a garantire fiducia, sicurezza, trasparenza e affidabilità nella loro messa in produzione [3][6][25][29].

Nell'ambito dell'Intelligenza Artificiale, i modelli di machine learning si collocano in uno spettro in cui quelli basati su metodologie più semplici sono spesso più interpretabili ma meno accurati, mentre quelli più complessi risultano generalmente più precisi ma più difficili da interpretare. Questa accuratezza nei modelli più complessi è dovuta alla loro migliore capacità di rappresentare pattern complessi e relazioni non lineari, a discapito di una maggiore trasparenza nel modello.

Perciò, con l'aumentare della complessità dei modelli di machine learning, diventa sempre più difficile comprendere come un determinato algoritmo giunga ad un determinato risultato.

In questi casi, il modello può essere considerato una cosiddetta *black-box*: un sistema in cui, pur essendo noti l'input e l'output, risulta estremamente complesso ricostruire la logica interna che ha portato a una determinata previsione [7].

La spiegabilità di un modello non riguarda esclusivamente la comprensione del suo funzionamento interno, ma anche la capacità di interpretare e giustificare

i risultati prodotti. In questo contesto, è possibile distinguere due principali tipologie di spiegabilità: **locale** e **globale**.

- **Spiegabilità locale**: riguarda la comprensione della previsione di un modello per una specifica istanza o per un piccolo sottoinsieme di istanze. L'obiettivo è capire perché il modello abbia prodotto una determinata previsione per uno specifico input e come eventuali modifiche a tale input possano influenzarne la decisione. La spiegabilità locale è utile per analizzare le motivazioni alla base di una singola previsione e per studiare il comportamento del modello su casi specifici.
- **Spiegabilità globale**: mira a fornire una comprensione complessiva del processo decisionale di un modello sull'intero dataset. L'obiettivo è capire come il modello utilizzi le variabili di input per generare le proprie decisioni e quale sia il contributo relativo di ciascuna caratteristica alle previsioni complessive. La spiegabilità globale è utile per comprendere il funzionamento generale del modello e per valutare la coerenza delle sue decisioni su larga scala.

Nel seguito saranno presentati tre metodi di Explainable AI: SHAP (SHapley Additive exPlanations), Permutation Feature Importance e Ablation Feature Importance.

1.3 SHAP: SHapley Additive exPlanations

Il metodo **SHAP** (**SHapley Additive exPlanations**) è stato introdotto nel 2017 da *Scott Lundberg* e *Su-In Lee* [33] con l'obiettivo di fornire spiegazioni interpretabili delle previsioni dei modelli di machine learning basate su solidi fondamenti teorici.

L'approccio si basa sui valori di Shapley, derivati dalla teoria dei giochi cooperativi, e consente di attribuire in modo equo il contributo di ciascuna caratteristica (feature) alla previsione finale del modello.

Lo sviluppo di SHAP nasce dall'esigenza di superare la frammentazione preesistente nel campo della XAI, proponendo un approccio unificato capace di integrare e generalizzare diversi metodi precedentemente introdotti.

Questa metodologia si distingue nel panorama della Explainable AI proprio per la sua capacità di garantire una coerenza matematica che altri approcci non riescono ad offrire.

SHAP non si limita ad indicare quali caratteristiche siano rilevanti, ma quantifica l'impatto di ognuna di esse rispetto alla previsione media del modello, permettendo una scomposizione additiva dell'output.

1.3.1 Valori SHAP

Dal punto di vista matematico, i *valori SHAP* (o *SHAP values*) costituiscono un metodo utile per attribuire in modo equo il contributo di ciascun "giocatore"

al risultato finale di una coalizione, calcolando la media dei contributi marginali su tutte le possibili combinazioni di partecipanti [52].

Nell'applicazione al machine learning, le caratteristiche di input vengono interpretate come giocatori cooperanti, mentre la previsione del modello rappresenta il guadagno complessivo da distribuire tra essi.

L'obiettivo dei valori SHAP è quello di assegnare a ciascuna feature un peso numerico che rappresenti l'impatto marginale medio della stessa sulla previsione del modello [48].

Grazie a questa formulazione, i valori SHAP garantiscono il rispetto di quattro assiomi principali:

1. **Efficienza**: garantisce che il valore totale sia distribuito tra tutti i giocatori;
2. **Simmetria**: caratteristiche con contributi identici ricevono lo stesso valore;
3. **Giocatore fittizio**: un giocatore che non contribuisce in alcun modo a nessuna coalizione riceve un valore pari a zero;
4. **Additività**: consente di applicare il valore di Shapley a più giochi sommando i contributi individuali.

Queste proprietà garantiscono che il contributo di ciascun giocatore venga valutato in modo coerente all'interno di un contesto cooperativo.

I valori SHAP possono essere utilizzati sia in una prospettiva di analisi locale sia globale:

- a livello locale permettono di interpretare le motivazioni alla base di una singola previsione, isolando l'impatto di ogni variabile per quel caso specifico;
- a livello globale consentono di analizzare l'importanza complessiva delle feature e il comportamento generale del modello sull'intero dataset.

Formalmente, il valore SHAP ϕ_i associato alla caratteristica i-esima è definito come segue [38]:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left[f(S \cup \{i\}) - f(S) \right] \tag{1.1}$$

dove x indica l'istanza specifica che stiamo considerando, $N = \{1, 2, \ldots, n\}$ indica l'insieme delle caratteristiche, S rappresenta un qualsiasi sottoinsieme di caratteristiche che non include la caratteristica i-esima e $|S|$ ne indica la cardinalità.

La funzione $f(S)$ rappresenta la funzione di predizione del modello calcolata utilizzando solo le feature presenti nel sottoinsieme S.

L'espressione $[f(S \cup \{i\}) - f(S)]$ rappresenta il contributo marginale della feature i-esima, ovvero la differenza nella previsione quando tale feature viene aggiunta alla coalizione.

Infine, il termine $\frac{|S|!(n-|S|-1)!}{n!}$ rappresenta il peso assegnato a tale contributo marginale, calcolato in base alle permutazioni possibili con cui le feature possono essere introdotte nel modello.

Il calcolo esatto dei valori SHAP richiede di considerare tutte le possibili combinazioni di caratteristiche, comportando un costo computazionale esponenziale rispetto al numero di feature. Per questo motivo sono stati sviluppati diversi algoritmi di approssimazione che permettono di calcolare tali valori in modo efficiente per differenti modelli di machine learning.

1.3.2 Metodo SHAP

Il metodo SHAP si basa sui valori di Shapley derivati dalla teoria dei giochi cooperativi. Sebbene tali valori fossero già stati applicati all'interpretabilità dei modelli di machine learning, SHAP introduce algoritmi efficienti per la loro stima e unifica diversi approcci precedentemente proposti (come LIME [50], DeepLIFT, Layer-Wise Relevance Propagation, Shapley Regression/Sampling values, Quantitative Input Influence [33]).

L'obiettivo fondamentale di SHAP è quello di fornire una spiegazione alla previsione di una specifica istanza x, quantificando il contributo individuale di ogni caratteristica.

In questo contesto, i valori delle caratteristiche dell'istanza vengono interpretati come i "giocatori" di una coalizione che collaborano con l'obiettivo di determinare il risultato finale del modello.

Il valore di Shapley determina come distribuire equamente la differenza tra la previsione del modello per l'istanza considerata e il valore medio delle predizioni sull'intero dataset.

Ogni caratteristica riceve quindi un punteggio che riflette il suo contributo medio alla previsione del modello.

Questo approccio garantisce che l'attribuzione dell'importanza sia basata su rigorose proprietà di equità derivanti proprio dalla teoria dei giochi cooperativi.

Il metodo SHAP introduce una vera innovazione, che consiste nel definire la spiegazione dei valori di Shapley come un modello lineare di attribuzione additiva.

Se il modello originale $f(x)$ è una "black box" complessa e non lineare, SHAP definisce un modello esplicativo $g(z')$ che risulta lineare rispetto a variabili binarie $z' \in \{0,1\}^N$, dove z' rappresenta il vettore di coalizione:

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i z'_i. \tag{1.2}$$

In tale formulazione, ϕ_0 rappresenta il valore atteso della predizione, N il numero massimo di caratteristiche considerate e ϕ_i il valore SHAP associato alla i-esima caratteristica.

Il vettore di coalizione z' viene talvolta indicato anche come *simplified features*, in quanto le caratteristiche utilizzate nel modello esplicativo possono rappresentare una versione semplificata delle caratteristiche originali del modello.

Nel vettore di coalizione, ciascun elemento assume valore pari a 1 se la caratteristica corrispondente è considerata "presente" e valore pari a 0 se è "assente".

È importante notare che le caratteristiche presenti nel vettore di coalizione non coincidono necessariamente con quelle utilizzate direttamente dal modello originale; l'aspetto rilevante è l'esistenza di una mappatura coerente tra le due rappresentazioni.

Una volta calcolati i valori SHAP per ciascuna istanza del dataset, è possibile analizzarli attraverso specifiche tecniche di visualizzazione che permettono di sintetizzare e interpretare il contributo delle caratteristiche al comportamento del modello.

Visualizzazione delle Spiegazioni: Bar Plot e Beeswarm Plot

Il framework SHAP fornisce diversi strumenti grafici che consentono di passare dall'interpretazione locale delle singole predizioni a una comprensione più globale del modello, in particolare due tipologie di visualizzazione principali [46]:

1. **SHAP Bar Plot**. Il Bar plot costituisce la forma più immediata per rappresentare l'importanza globale delle caratteristiche. Per ogni feature i, viene calcolata la media dei valori SHAP assoluti su tutte le n istanze del dataset:

$$I_i = \frac{1}{n} \sum_{j=1}^{n} |\phi_i^{(j)}| \tag{1.3}$$

   Questo grafico permette di identificare rapidamente quali feature abbiano l'impatto medio maggiore sulla previsione, pur non fornendo informazioni sulla direzione dell'effetto, ovvero se l'impatto sia positivo o negativo rispetto all'output.

   La figura seguente riporta un esempio di bar plot, utile per comprenderne il funzionamento.

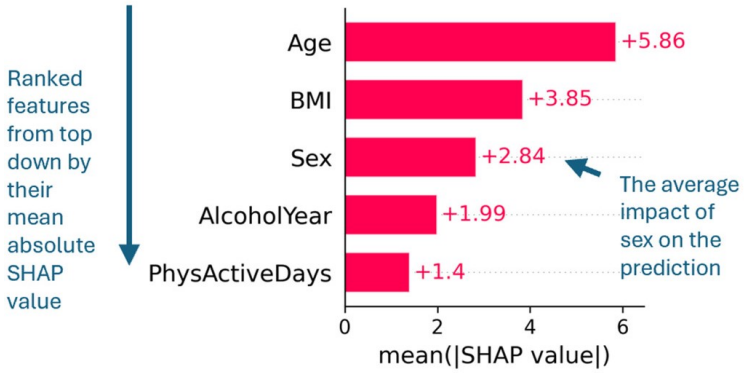

Figura 1.3: Esempio di Bar Plot. Immagine tratta da: *Ponce-Bobadilla, A.et al.*, 2024 [46].

2. **SHAP Beeswarm Plot**. Il Beeswarm plot è una tecnica di visualizzazione avanzata, che combina l'ordine di importanza delle caratteristiche con la distribuzione dei loro effetti. In questo grafico, ogni riga rappresenta una feature e ogni punto corrisponde a una singola istanza del dataset. Le sue componenti informative sono:

   - **Posizione sull'asse delle ascisse**: indica il valore SHAP. I punti a destra dello zero rappresentano un impatto positivo sulla previsione, mentre quelli a sinistra un impatto negativo.
   - **Colore**: rappresenta il valore effettivo assunto dalla caratteristica nell'istanza. Generalmente il blu viene utilizzato per valori bassi, mentre il rosso per valori alti.
   - **Densità**: l'accumulo dei punti nelle diverse aree del grafico permette di visualizzare la distribuzione della popolazione e la frequenza di determinati effetti.

   L'analisi del Beeswarm plot consente di identificare immediatamente relazioni complesse e non lineari. Ad esempio, la concentrazione di punti rossi nella regione positiva dell'asse x suggerisce una relazione positiva tra la variabile e l'output del modello. Inoltre, la presenza di code lunghe o asimmetrie rivela l'impatto di possibili outlier o effetti marginali che un semplice Bar plot non sarebbe in grado di evidenziare.

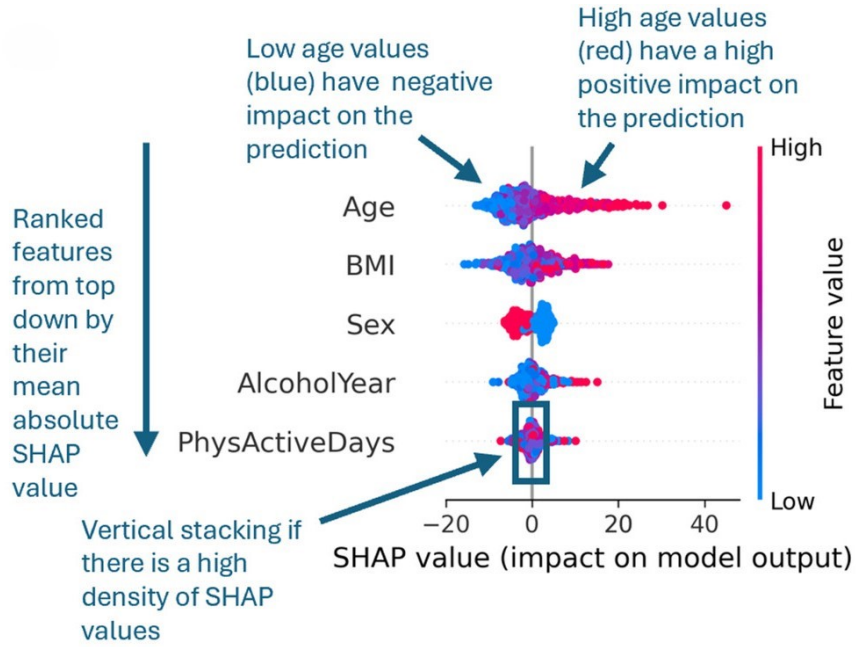

Figura 1.4: Esempio di Beeswarm Plot. Immagine tratta da: *Ponce-Bobadilla, A. et al.*, 2024 [46].

1.3.3 Varianti dell'explainer

SHAP offre diversi algoritmi di spiegazione, ottimizzati per differenti tipologie di modelli. In questo paragrafo analizzeremo le seguenti tre varianti principali: KernelSHAP, TreeSHAP e DeepSHAP.

KernelSHAP

Il calcolo esatto dei valori di Shapley teorici richiede la valutazione del modello su tutti i possibili sottoinsiemi di caratteristiche, ovvero circa 2^N coalizioni per N feature, rendendolo computazionalmente proibitivo nella maggior parte dei casi.

Esistono quindi metodi di approssimazione che affrontano questa problematica. Tra questi il metodo **KernelSHAP** [33].

Questo approccio spiega le singole previsioni stimando un modello surrogato locale su istanze perturbate generate a partire da diverse combinazioni (coalizioni) di variabili.

Invece di affidarsi a perturbazioni arbitrarie, il metodo utilizza uno schema di ponderazione basato su un kernel derivato dalla teoria dei valori di Shapley. I coefficienti del modello lineare risultante rappresentano il contributo delle singole caratteristiche nel determinare la specifica previsione del modello.

Per simulare l'assenza di una feature, KernelSHAP introduce un dataset di *background*.

La generazione di un'istanza perturbata, corrispondente a una specifica coalizione, avviene mantenendo i valori originali dell'istanza x per le variabili incluse nella coalizione stessa. Al contrario, alle caratteristiche non presenti nella coalizione vengono assegnati valori campionati dal dataset di background.

Tale procedura consente di approssimare l'effetto della rimozione di una caratteristica integrando la funzione di predizione sulla distribuzione marginale degli input assenti, sostituendoli con valori rappresentativi della popolazione di riferimento.

Una volta generati i campioni perturbati, il modello "black-box" originale viene utilizzato per ottenere le relative previsioni.

KernelSHAP procede quindi all'addestramento di un modello di regressione lineare ponderata che funge da surrogato locale. Le variabili indipendenti di tale modello sono vettori binari $z' \in \{0,1\}^N$, dove ogni componente z'_i funge da indicatore della presenza ($z'_i = 1$) o dell'assenza ($z'_i = 0$) della i-esima caratteristica nella coalizione.

Il modello surrogato assume la stessa forma additiva introdotta nella formulazione generale di SHAP:

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i z'_i \tag{1.4}$$

dove i coefficienti ϕ_i rappresentano le stime dei valori di Shapley per ciascuna caratteristica.

Il cuore del modello risiede nello schema di ponderazione adottato nella regressione. I campioni perturbati non contribuiscono equamente alla stima, ma ad

ogni coalizione viene assegnato un peso specifico, definito *kernel SHAP*, derivato analiticamente dalla teoria dei giochi cooperativi.

Questo schema attribuisce un'importanza maggiore alle coalizioni caratterizzate da una cardinalità molto bassa o molto elevata.

Matematicamente, tale schema garantisce che i coefficienti ϕ_i risultanti soddisfino i principali assiomi teorici dei valori di Shapley, in particolare:

- **Efficienza**: la somma dei contributi delle caratteristiche approssima la differenza tra la previsione dell'istanza e il valore atteso medio del modello.
- **Consistenza**: se il contributo marginale di una caratteristica aumenta in un modello rispetto a un altro, il corrispondente valore di Shapley non deve diminuire.

Attraverso questo processo di stima pesata, KernelSHAP trasforma il problema della stima dei valori di Shapley in un problema di regressione lineare ponderata, garantendo una soluzione coerente con i vincoli richiesti dall'interpretabilità dei modelli.

L'estrema versatilità di KernelSHAP risiede nella sua natura *model-agnostic*, che ne consente l'applicazione a qualsiasi architettura di machine learning, indipendentemente dalla complessità della struttura interna.

Tuttavia, l'efficacia del metodo è strettamente legata alla configurazione del dataset di background, il quale definisce il valore di riferimento per le feature omesse. Una scelta metodologica comune, volta a bilanciare rappresentatività e costi computazionali, consiste nell'utilizzare un sottoinsieme del set di addestramento o una sua sintesi statistica (come ad esempio i centroidi ottenuti tramite l'algoritmo k-means).

Dal punto di vista computazionale, il metodo può risultare particolarmente oneroso. La generazione delle coalizioni, l'inferenza del modello originale e l'adattamento della regressione ponderata comportano infatti un costo computazionale elevato, soprattutto nel caso di modelli con dataset ad alta dimensionalità o set di background molto ampi.

Poiché KernelSHAP fornisce stime stocastiche dei valori di Shapley reali, l'accuratezza delle stime dipende direttamente dal numero di campioni generati, il quale deve essere proporzionale al numero di feature (N) per garantire la convergenza delle stime.

In sintesi, KernelSHAP rappresenta uno strumento fondamentale e teoricamente rigoroso per estendere i principi della teoria dei giochi cooperativi a qualsiasi sistema "black-box".

Sebbene costituisca il pilastro del framework SHAP per l'interpretazione di modelli complessi, le sue esigenze computazionali suggeriscono, dove possibile, l'adozione di varianti specializzate e più efficienti come TreeSHAP o DeepSHAP [1].

TreeSHAP

TreeSHAP [33] è un algoritmo di interpretabilità progettato specificamente per modelli basati su alberi, come alberi decisionali, Random Forest e Gradient Boosted Decision Trees.

A differenza di KernelSHAP, che adotta un approccio *model-agnostic* basato su campionamenti, TreeSHAP è un metodo *model-specific*: esso sfrutta la struttura interna del modello per calcolare i valori di Shapley in modo esatto ed efficiente [34].

Il nucleo dell'algoritmo consiste nel calcolare i contributi delle caratteristiche attraverso una procedura ricorsiva che percorre i nodi dell'albero decisionale. Invece di valutare 2^N possibili coalizioni, TreeSHAP utilizza un approccio di programmazione dinamica per considerare simultaneamente tutti i possibili percorsi decisionali.

Questo permette di aggregare i pesi delle coalizioni e i valori delle foglie in un unico passaggio, riducendo drasticamente la complessità computazionale da esponenziale a polinomiale.

Attribuendo i contributi marginali man mano che si discende lungo i rami dell'albero, TreeSHAP garantisce la stima dei valori di Shapley esatti, eliminando la varianza tipica dei metodi basati su campionamento e rendendo possibile l'interpretazione di modelli complessi su dataset di grandi dimensioni.

Grazie a questa proprietà, TreeSHAP rappresenta la soluzione preferibile quando il modello da interpretare è basato su alberi, poiché consente di ottenere valori di Shapley esatti con un costo computazionale significativamente inferiore rispetto ai metodi *model-agnostic* come KernelSHAP.

TreeSHAP utilizza un algoritmo specializzato basato sul calcolo delle aspettative condizionali del modello. A differenza di KernelSHAP, che si affida a perturbazioni stocastiche, TreeSHAP determina l'aspettativa condizionale esatta:

$$\mathbb{E}[f(X) \mid X_S = x_S] \tag{1.5}$$

che rappresenta l'output atteso del modello condizionato nel caso in cui siano noti solo i valori delle caratteristiche appartenenti al sottoinsieme di coalizione S.

L'algoritmo calcola simultaneamente tali aspettative per tutti i possibili sottoinsiemi S, propagando i contributi lungo i rami dell'albero decisionale.

Operativamente, l'algoritmo analizza i percorsi dell'albero secondo una logica ricorsiva. Quando incontra un nodo di suddivisione (split) basato su una feature appartenente a S, TreeSHAP segue esclusivamente il ramo coerente con il valore effettivo dell'istanza x. Al contrario, se la suddivisione avviene su una feature non inclusa in S, l'algoritmo percorre entrambi i rami disponibili. In quest'ultimo caso, viene calcolata una media ponderata dei risultati, dove i pesi riflettono la proporzione di campioni del set di addestramento che sono transitati per quel nodo.

Questo meccanismo di "integrazione sui pesi dei nodi" permette di marginalizzare l'effetto delle caratteristiche assenti in modo efficiente, senza la necessità di utilizzare un dataset di background esterno, poiché la distribuzione dei dati è implicitamente codificata nella struttura dell'albero.

I vantaggi di TreeSHAP sono molteplici:

- **Efficienza Computazionale**: TreeSHAP è significativamente più veloce di KernelSHAP per i modelli tree-based. La sua complessità computazionale è approssimativamente
$$\mathcal{O}(TLD^2),$$
dove T rappresenta il numero di alberi, L il numero massimo di foglie e D la profondità massima dell'albero. Questa complessità risulta nettamente inferiore rispetto alla complessità esponenziale richiesta dal calcolo esatto dei valori di Shapley o al costo computazionale basato sul campionamento di KernelSHAP.

- **Esattezza matematica**: a differenza di KernelSHAP, che fornisce stime approssimate, TreeSHAP calcola i valori di Shapley teoricamente esatti per il modello considerato. Ciò elimina l'incertezza legata alla convergenza delle stime.

- **Analisi delle Interazioni**: TreeSHAP può essere esteso per calcolare in modo efficiente i valori di interazione SHAP, consentendo di analizzare come coppie di caratteristiche interagiscano tra loro nel determinare la previsione del modello.

Il limite principale di TreeSHAP risiede nella sua specificità. Essendo un metodo *model-specific*, la sua applicabilità è limitata esclusivamente ai modelli basati su alberi. Qualora il problema richiedesse l'impiego di modelli lineari, SVM o reti neurali, è necessario ricorrere ad approcci alternativi. Tra questi, KernelSHAP si conferma la scelta principale, per la sua natura *model-agnostic*, mentre DeepSHAP rappresenta l'estensione specifica e ottimizzata per le architetture di deep learning.

Tuttavia, data la prevalenza e le elevate prestazioni di modelli tree-based nelle applicazioni basate su dati tabulari, TreeSHAP risulta essere uno strumento particolarmente efficace e ampiamente utilizzato. Esso permette di ottenere spiegazioni esatte e computazionalmente efficienti, rendendo l'interpretabilità un processo scalabile anche per sistemi di grande complessità.

DeepSHAP

DeepSHAP [33] è un'estensione del framework SHAP specificamente progettata per l'interpretabilità di modelli di deep learning.

Questo approccio integra i principi della metodologia SHAP con DeepLIFT (*Deep Learning Important FeaTures* [54]), un algoritmo che confronta l'attivazione di ciascun neurone con la sua "attivazione di riferimento" ed assegna punteggi di contributo in base alla differenza.

DeepSHAP riduce efficacemente il divario di interpretabilità tra i modelli convenzionali e le complesse architetture neurali, fornendo un indicatore coerente e unificato dell'importanza delle feature.

Come dimostrato da *Lundberg* e *Lee*, le regole di attribuzione utilizzate in DeepLIFT possono essere adattate per approssimare i valori di Shapley in modo computazionalmente efficiente.

Il metodo si basa sulla propagazione all'indietro dei contributi attraverso la rete neurale. Considerando molteplici campioni di *background*, DeepExplainer stima i valori SHAP in modo da soddisfare la proprietà di additività.

La somma delle attribuzioni deve corrispondere alla differenza tra l'output attuale del modello $f(x)$ e il valore atteso della predizione sulla distribuzione di riferimento $\mathbb{E}[f(X)]$.

Formalmente, la proprietà di additività dei valori SHAP può essere espressa come:

$$\sum_{i=1}^{N} \phi_i = f(x) - \mathbb{E}[f(X)] \tag{1.6}$$

Questo approccio permette di gestire la non-linearità tipica delle reti profonde, fornendo spiegazioni locali coerenti e interpretabili anche per modelli con milioni di parametri.

L'adozione di DeepSHAP nel contesto delle reti neurali profonde offre molteplici benefici:

- **Trasparenza**: illustrando l'influenza di ogni feature di input sull'output, rende interpretabili modelli di deep learning tradizionalmente considerati sistemi "black-box".
- **Interpretabilità e affidabilità**: consente di quantificare il contributo individuale delle caratteristiche, migliorando la comprensione delle decisioni del modello e aumentando la fiducia nei sistemi di deep learning.
- **Debug e ottimizzazione del modello**: fornisce una prospettiva critica sul comportamento del modello, facilitando l'individuazione di debolezze strutturali, comportamenti anomali o overfitting.
- **Mitigazione dei bias ed equità**: permette di identificare quali caratteristiche influenzano maggiormente le previsioni, facilitando il rilevamento di eventuali distorsioni nei dati o nel modello.

Nonostante l'efficacia di DeepSHAP nell'interpretazione di reti neurali complesse, il metodo presenta alcuni limiti operativi.

In primo luogo, la complessità computazionale rappresenta un vincolo significativo: la necessità di integrare le predizioni su ampi dataset di background può limitare l'uso del metodo in contesti real-time. Inoltre, è opportuno considerare il rischio di propagazione delle distorsioni: DeepSHAP spiega fedelmente il modello, il che significa che rifletterà eventuali bias o errori sistematici presenti nei dati di addestramento.

In conclusione, DeepSHAP rappresenta uno dei metodi più efficaci per interpretare modelli di deep learning, combinando il rigore teorico dei valori di Shapley con l'efficienza computazionale di DeepLIFT.

Pur presentando alcuni limiti operativi, come il costo computazionale e la dipendenza dal dataset di background, il metodo fornisce uno strumento potente per analizzare e comprendere il comportamento delle reti neurali profonde.

I metodi presentati in questa sezione costituiscono il nucleo del framework SHAP e permettono di applicare i principi della teoria dei giochi cooperativi a diverse classi di modelli di machine learning. In funzione della struttura del modello da interpretare, è quindi possibile scegliere tra approcci *model-agnostic* o algoritmi specializzati che consentono di ottenere spiegazioni più efficienti e accurate.

1.4 Permutation Feature Importance

La **Permutation Feature Importance** (PFI) è una tecnica che fornisce una misura dell'importanza di ogni caratteristica basata sull'effettivo degrado delle prestazioni del modello, risultando particolarmente efficace per l'analisi di stimatori non lineari o complessi [44].

L'idea è misurare quanto peggiorano le prestazioni del modello quando i valori di una particolare caratteristica vengono permutati in modo casuale, lasciando inalterate le altre variabili, interrompendo così la relazione tra la caratteristica e la variabile target [18].

Questo permette di confrontare le prestazioni del modello prima e dopo la permutazione della feature, ottenendo così un valore di importanza che quantifica la diminuzione delle prestazioni del modello attribuibile alla feature modificata.

Il procedimento consiste innanzitutto nell'addestrare un modello di machine learning utilizzando le feature originali del dataset di training. Successivamente si calcola una metrica di valutazione delle prestazioni del modello (ad esempio l'accuratezza).

Per stimare l'importanza di una specifica feature, i valori di tale feature vengono permutati casualmente nel dataset, interrompendo la relazione tra quella variabile e la variabile target.

Il modello addestrato viene quindi applicato al dataset perturbato e le prestazioni vengono nuovamente calcolate.

L'importanza della feature è determinata confrontando le prestazioni del modello ottenute con i dati originali e quelle ottenute dopo la permutazione. Se la permutazione di una feature provoca una riduzione significativa delle prestazioni del modello, ciò indica che quella feature è rilevante per il processo di predizione.

Il metodo può essere ripetuto più volte utilizzando diverse permutazioni casuali della stessa caratteristica, ottenendo una distribuzione dei punteggi di importanza. Questo processo permette di calcolare non solo una stima puntuale, ma anche una misura della varianza e degli intervalli di confidenza per ogni caratteristica.

Tale robustezza statistica è fondamentale per distinguere tra segnali di importanza reali e fluttuazioni casuali dovute alla specifica permutazione o al rumore presente nei dati.

Per ogni caratteristica si ottiene un valore di importanza che indica se il modello si affida significativamente ad una specifica caratteristica per le proprie previsioni. In quel caso la rottura della sua relazione con il target, causerà un netto degrado del punteggio. Viceversa, la permutazione di una caratteristica non predittiva o ridondante, lascerà le prestazioni del modello sostanzialmente invariate.

Il calcolo dell'importanza viene quindi definito come la differenza tra il punteggio ottenuto sul dataset originale e quello ottenuto sul dataset perturbato. Questo approccio offre il vantaggio di non richiedere il riaddestramento del modello per ogni caratteristica.

Per comprendere il funzionamento della Permutation Feature Importance, è utile formalizzare i passaggi computazionali necessari per la sua esecuzione.

Algorithm 1 Permutation Feature Importance

Require: Modello addestrato m, dataset D, metrica di punteggio s, numero di ripetizioni k;
Ensure: Importanza $\hat{i}_j$ per ogni caratteristica f_j;
1: Calcola il punteggio di riferimento $s_{base} = s(m, D)$
2: **for** ogni caratteristica $j \in \{1, \ldots, N\}$ **do**
3: **for** $k = 1$ **to** K **do**
4: Genera $\tilde{D}_{k,j}$ permutando casualmente la colonna j di D
5: Calcola il punteggio sulla versione perturbata: $s_{k,j} = s(m, \tilde{D}_{k,j})$
6: **end for**
7: $\hat{i}_j = s_{base} - \frac{1}{K} \sum_{k=1}^{K} s_{k,j}$ (variazione media del punteggio)
8: **end for**
9: **return** $\{\hat{i}_j\}_{j=1}^{N}$

Il valore di importanza $\hat{i}_j$ rappresenta il calo medio delle prestazioni del modello quando l'informazione contenuta nella caratteristica j viene eliminata.

Importante sottolineare due aspetti fondamentali per il modello:

- La scelta del dataset D: è essenziale calcolare la Permutation Feature Importance utilizzando il test set o un campione trattenuto.
  Per l'algoritmo eseguito sul test set l'importanza indicherà quanto ogni caratteristica contribuisce alla capacità di generalizzazione del modello su dati non visti.

- La metrica di punteggio s: la natura dell'importanza di una caratteristica può cambiare drasticamente a seconda della metrica scelta.

Un vantaggio della Permutation Feature Importance risiede nella sua natura *model-agnostic* [39]. Non dipendendo da parametri interni o da specifiche architetture, risulta essere applicabile a qualunque tipologia di predittore. Ciò consente di confrontare direttamente l'importanza delle caratteristiche tra modelli di natura diversa, utilizzando una metrica comune basata sulle prestazioni.

Sebbene la Permutation Feature Importance sia estremamente versatile, essa può fornire risultati fuorvianti in presenza di variabili fortemente correlate: in tali scenari, il modello potrebbe compensare la perdita di una variabile permutata utilizzando le informazioni presenti nelle sue correlate, sottostimando così l'importanza reale di entrambe.

Il costo computazionale della Permutation Feature Importance cresce linearmente con il numero di caratteristiche e con il numero di ripetizioni della permutazione.

Per dataset ad alta dimensionalità o modelli particolarmente costosi, il calcolo dell'importanza può risultare oneroso.

1.5 Ablation Feature Importance

L'**Ablation Feature Importance** (AFI) è una tecnica utile per determinare l'importanza di singole caratteristiche rimuovendole sistematicamente ("ablating") e misurando l'impatto risultante sulle prestazioni del modello [32].

Il principio è intuitivo: se una caratteristica possiede un'elevata importanza, la sua rimozione dal dataset comporterà un deterioramento significativo delle prestazioni del modello; viceversa, l'eliminazione di variabili ridondanti o non informative produrrà variazioni trascurabili nel punteggio finale.

Si considera un dataset su cui si addestra un modello di machine learning, e si calcola una metrica di valutazione delle prestazioni per il modello.

Successivamente si rimuove una caratteristica, eliminandola temporaneamente dai dati di input per osservare i cambiamenti nelle prestazioni del modello.

Si misura quindi l'impatto della caratteristica valutando le prestazioni del modello con la feature rimossa. Questo processo si ripete per ogni caratteristica per determinare l'importanza relativa di tutte le caratteristiche.

A differenza delle tecniche di permutazione, l'Ablation Feature Importance fornisce una misura dell'importanza che tiene conto della capacità del modello di adattarsi alla mancanza di una specifica informazione fin dalla fase di addestramento.

La procedura formale è descritta dal seguente algoritmo:

Algorithm 2 Feature Ablation Importance

Require: Dataset di training D_{train}, dataset di test D_{test}, modello m, metrica di punteggio s

Ensure: Importanza a_j per ogni caratteristica f_j

1: Addestra il modello m su D_{train} utilizzando l'intero set di caratteristiche
2: Calcola il punteggio di riferimento: $s_{base} = s(m, D_{\text{test}})$
3: **for** ogni caratteristica $j = 1, \ldots, N$ **do**
4: Rimuovi la caratteristica f_j da D_{train} e D_{test}
5: **Riaddestra** il modello m sul dataset ridotto
6: Calcola il nuovo punteggio: $s_j = s(m, D_{\text{test}})$
7: $a_j = s_{base} - s_j$ (decremento di performance)
8: **end for**
9: Ordina le caratteristiche in ordine decrescente rispetto ad a_j
10: **return** $\{a_j\}_{j=1}^{N}$

Analogamente alla Permutation Feature Importance, anche questo metodo è *model-agnostic* e può quindi essere applicato a qualsiasi modello di machine learning [22].

L'elemento distintivo dell'Ablation Feature Importance risiede nella fase di riaddestramento del modello. Mentre altre tecniche, come la Permutation Im-

portance, si limitano a perturbare i dati di test su un modello già addestrato, l'Ablation valuta l'importanza di una caratteristica osservando come l'algoritmo di apprendimento reagisce alla sua totale assenza fin dalla fase di training.

Tuttavia, l'adozione di questo metodo comporta alcune considerazioni critiche. Poiché il modello deve essere riaddestrato per ogni singola caratteristica, il costo computazionale scala linearmente con il numero di feature (N).

Se l'addestramento del modello richiede tempi significativi, il calcolo dell'Ablation Importance può diventare computazionalmente oneroso. Inoltre, in presenza di variabili fortemente correlate, l'Ablation tende a mostrare un'importanza inferiore per ciascuna di esse. Se due variabili contengono la stessa informazione, rimuovendone una il modello può "recuperare" l'accuratezza sfruttando l'altra durante il riaddestramento, indicando che nessuna delle due è strettamente indispensabile se l'altra è presente.

Questa tecnica fornisce spesso una stima più fedele di quanto una variabile contribuisca alla capacità di generalizzazione del modello, poiché evita la creazione di combinazioni di dati potenzialmente irrealistiche che possono emergere nei metodi basati sulla permutazione.

Dal punto di vista concettuale, l'Ablation Feature Importance può essere interpretata come una variante più rigorosa della Permutation Feature Importance. Mentre la permutazione altera artificialmente la distribuzione dei dati di test, l'ablation valuta l'importanza delle caratteristiche osservando come il modello si riadatta alla loro completa assenza durante il processo di apprendimento.

1.6 Sintesi dei metodi considerati

Nel presente capitolo sono stati introdotti diversi metodi di Explainable Artificial Intelligence (XAI) utili all'interpretazione delle previsioni dei modelli di machine learning.

In particolare, sono stati analizzati gli approcci basati sui valori di Shapley, tra cui KernelSHAP, TreeSHAP e DeepSHAP, e tecniche di analisi dell'importanza delle caratteristiche come Permutation Feature Importance e Ablation Feature Importance.

Nel presente lavoro di tesi saranno confrontate le seguenti tre tecniche: TreeSHAP, Permutation Feature Importance (PFI) e Ablation Feature Importance (AFI), poiché tali metodi consentono di analizzare in modo complementare il contributo delle singole caratteristiche alle previsioni dei modelli considerati.

TreeSHAP permette di ottenere spiegazioni precise per modelli basati su alberi, mentre Permutation e Ablation forniscono una misura diretta dell'importanza delle variabili attraverso la perturbazione o la rimozione delle caratteristiche.

Metodo	**Tipo di spiegazione**	**Costo computazionale**
TreeSHAP	Locale/Globale	Basso
Permutation Importance	Globale	Medio
Ablation Importance	Globale	Elevato

Tabella 1.1: Confronto tra i tre metodi di interpretabilità utilizzati.

Capitolo 2

Dataset e Metodologia

In questo capitolo vengono descritti il dataset utilizzato e la metodologia adottata per l'analisi.

L'intera pipeline è stata implementata in Python mediante l'utilizzo di diverse librerie di machine learning.

Nelle sezioni seguenti vengono presentate le principali caratteristiche del dataset, le procedure di preparazione e organizzazione dei dati e l'approccio modellistico utilizzato.

Infine, vengono illustrate le tecniche di interpretabilità impiegate per analizzare il contributo delle variabili biologiche alle predizioni dei modelli.

2.1 Descrizione del Dataset

Il dataset oggetto della presente sperimentazione raccoglie i profili di 569 pazienti provenienti da cinque nazioni diverse: Regno Unito, Olanda, Spagna, Francia e Nord America.

Ognuno di questi campioni presenti nel dataset proviene da pazienti oncologici affetti da melanoma (MM), carcinoma renale (RCC) e carcinoma polmonare non a piccole cellule (NSCLC), tutti sottoposti a trattamento con immunoterapia.

Inoltre, ogni campione contiene informazioni di base sul paziente, quali dati anagrafici e clinici (riportati in Appendice A, Tabella A.1), di seguito indicati come *metadati*.

Tra i *metadati* troviamo anche la variabile target (`responder`) che è binaria ed è definita come segue:

- **Classe 0 (Responder)**: 262 pazienti (circa il 46%);
- **Classe 1 (Non-Responder)**: 307 pazienti (circa il 54%).

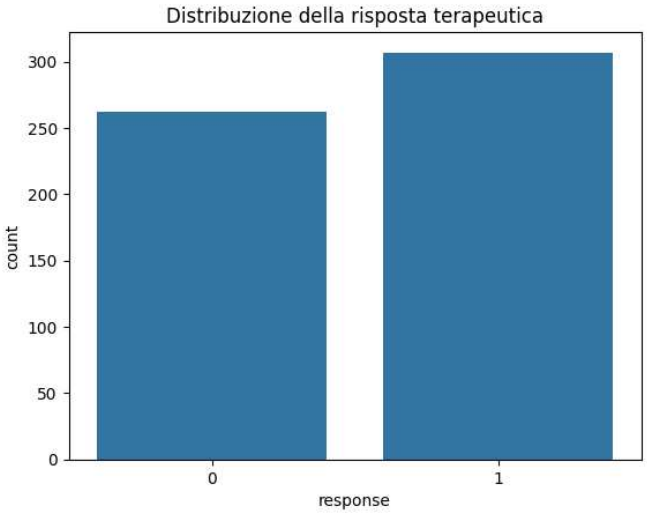

Figura 2.1: Distribuzione della variabile target.

Come evidenziato nella Figura 2.1, la classe maggioritaria è la Classe 1, composta dai pazienti che non hanno risposto alla terapia. Tuttavia, lo scarto rispetto alla Classe 0 (pazienti Responder) risulta contenuto, indicando un dataset complessivamente bilanciato.

Oltre ai *metadati*, ogni record include 4630 caratteristiche numeriche che quantificano l'abbondanza relativa delle diverse specie microbiche presenti nel microbiota intestinale.

In linea con gli obiettivi di questa tesi, focalizzata sull'identificazione di biomarcatori batterici tramite tecniche di Explainable AI, l'analisi esclude i *metadati* per concentrarsi esclusivamente sulle feature microbiotiche.

È importante sottolineare che l'analisi esplorativa dei dati (EDA), lo studio delle distribuzioni e delle correlazioni tra le variabili sono stati oggetto di analisi preliminari esterne alla presente trattazione.

Tali analisi hanno evidenziato un'elevata sparsità dei dati e una complessa struttura di interdipendenza tra i taxa, suggerendo l'utilizzo di modelli non lineari e tecniche avanzate di interpretabilità per analizzare correttamente il contributo delle singole specie microbiche nella predizione della risposta terapeutica.

Il dataset è stato suddiviso nel seguente modo:

- **Train set**: 80% dei dati, pari a 455 osservazioni;
- **Test set**: 20% dei dati, pari a 114 osservazioni.

2.2 Selezione delle Feature e Preprocessing

Come discusso in precedenza, il dataset originale include numerosi *metadati* non strettamente rilevanti per l'identificazione di biomarcatori batterici. Pertanto, è stata effettuata una fase preliminare di selezione e pulizia delle variabili.

In primo luogo è stata isolata la variabile target `response`. Eventuali osservazioni prive di tale etichetta sono state rimosse al fine di garantire la coerenza nella fase di addestramento dei modelli.

Infine, sono stati rimossi gli indici tecnici non informativi.

Successivamente è stata effettuata una selezione delle caratteristiche volta a includere esclusivamente le 4630 feature microbiotiche.

Le colonne del dataset sono state quindi filtrate sulla base dei prefissi tassonomici standard.

Le variabili considerate rappresentano diversi livelli della gerarchia tassonomica, includendo gradi differenti di risoluzione biologica: dai livelli più generali, come il *Phylum* (*p_*), fino ai livelli più specifici come il *Genus* (*g_*) e la *Species* (*s_*).

Ogni osservazione può quindi essere ricondotta a una catena tassonomica gerarchica del tipo:

$$p\_Phylum \rightarrow f\_Family \rightarrow g\_Genus \rightarrow s\_Species$$

Questa procedura ha permesso di isolare le 4630 variabili numeriche che rappresentano l'abbondanza relativa dei diversi taxa presenti nel microbiota intestinale.

Infine, per rendere il dataset idoneo alla computazione algoritmica, è stata effettuata una fase di conversione numerica delle variabili, gestendo correttamente i separatori decimali e verificando l'assenza di valori mancanti (*NaN*).

Trasformazione CLR - Centered Log-Ratio

La trasformazione CLR (Centered Log-Ratio) è una tecnica statistica utilizzata per analizzare i dati composizionali, ovvero dati che rappresentano proporzioni o percentuali di un intero.

Nello specifico, trasforma i dati dal simplesso[1], allo spazio euclideo reale, permettendo l'uso di tecniche statistiche standard [2].

Un aspetto critico dei dati del microbiota intestinale è proprio la loro natura composizionale, infatti, tali dati sommano a 100 [21]. Quindi, se un batterio aumenta, gli altri devono necessariamente diminuire, creando false correlazioni.

La trasformazione CLR permette all'algoritmo di capire i rapporti reali tra le varie specie. Nello specifico, questa trasformazione, calcolata come il logaritmo del rapporto tra l'abbondanza di ogni taxa e la media geometrica di tutte le caratteristiche del campione[2], permette di [23]:

1. Rimuovere il vincolo della somma costante, normalizzando i dati;
2. Ridurre lo sbilanciamento causato dalle diverse scale di abbondanza tra taxa dominanti e rari;
3. Migliorare la stabilità e l'interpretabilità dei calcoli successivi.

Per gestire la presenza di valori pari a zero, che renderebbero nullo l'argomento del logaritmo, è stato introdotto un `pseudocount` pari a $1e-6$ prima della trasformazione.

[1] Il simplesso è uno spazio limitato a somma costante.

[2] Dato $x = (x_1, ..., x_D)$, vettore composizionale, la trasformazione CLR si ottiene come: $\text{clr}(x) = \left[\ln\left(\frac{x_1}{g(x)}\right), \ldots, \ln\left(\frac{x_D}{g(x)}\right)\right]$ dove $g(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$ è la media geometrica delle componenti.

2.3 Descrizione, Addestramento e Validazione dei Modelli

Dopo aver trasformato i dati, per renderli compatibili all'analisi, siamo passati alla classificazione della risposta terapeutica. Nello specifico sono stati selezionati tre diversi algoritmi basati su insiemi di alberi decisionali (*Ensemble Methods*).

Tale scelta è motivata dalla loro efficacia nel gestire dataset ad alta dimensionalità ($P \gg N$) e dalla capacità di catturare relazioni non lineari tra le variabili microbiotiche [41].

I modelli implementati sono:

1. **Random Forest (RF)**: un algoritmo di apprendimento supervisionato, il cui principio fondamentale è combinare le previsioni di molti alberi decisionali, addestrati su sottoinsiemi casuali dei dati (*bootstrapping*), al fine di ottenere una previsione finale più robusta e meno incline all'overfitting rispetto ad un singolo albero.

   Ogni albero considera, in ogni split, un sottoinsieme casuale di feature, invece di tutte le caratteristiche disponibili. Questo garantisce diversità tra gli alberi, riducendo la correlazione e aumentando la capacità predittiva complessiva.

   Una volta costruiti gli alberi, le previsioni vengono aggregate: nel caso della classificazione si adotta il voto di maggioranza (*majority voting)* tra gli alberi, mentre nella regressione si calcola la media delle predizioni.

   La Random Forest presenta numerosi vantaggi. Innanzitutto riduce il rischio di overfitting grazie alla combinazione di alberi differenti. È adatta a gestire dataset complessi, caratterizzati da molte variabili o dati mancanti, e fornisce una misura dell'importanza delle feature per la previsione. Inoltre, poiché ogni albero può essere addestrato indipendentemente, l'algoritmo si presta bene alla parallelizzazione, rendendo l'addestramento efficiente anche su grandi quantità di dati [8].

2. **Extra Trees (ET)**: un algoritmo di apprendimento supervisionato, che migliora il concetto di Random Forest introducendo un ulteriore livello di casualità nella costruzione degli alberi.

   A differenza del Random Forest, in cui per ogni nodo si cerca la soglia di split ottimale tra le feature selezionate, Extra Trees è caratterizzato da una maggiore casualità nella scelta delle soglie di split, mantenendo comunque l'apprendimento supervisionato. Questa maggiore casualità rende la costruzione degli alberi molto più veloce ed efficiente, poichè evita i calcoli intensivi necessari per individuare lo split ottimale.

   Un'altra differenza significativa riguarda l'utilizzo dei dati: mentre la Random Forest impiega il bootstrapping, campionando i dati con rimpiazzo per ciascun albero, le Extra Trees addestrano ogni albero sull'intero dataset di training, affidando la diversità dei modelli esclusivamente alla casualità delle soglie di split. Grazie a questo approccio, le Extra Trees tendono a

ridurre la varianza del modello e a mitigare il rischio di overfitting, risultando particolarmente efficaci in presenza di dataset complessi o rumorosi.

In sintesi, le Extra Trees rappresentano una variante della Random Forest ottimizzata per ridurre i tempi di calcolo e migliorare la robustezza predittiva del modello, combinando la semplicità del calcolo con una maggiore casualità nella scelta delle soglie di split [19].

3. **Extreme Gradient Boosting (XGBoost)**: algoritmo di *boosting* sequenziale altamente ottimizzato, ampiamente utilizzato per problemi di classificazione e regressione grazie alla sua velocità, precisione e capacità di gestire grandi dataset. L'idea alla base del boosting è costruire il modello in modo sequenziale, aggiungendo un nuovo albero alla volta per correggere gli errori prodotti dagli alberi precedenti. In XGBoost, questa procedura viene ulteriormente perfezionata grazie all'integrazione della regolarizzazione direttamente nella funzione obiettivo, che combina la misura dell'errore (loss function, come MSE o LogLoss) con un termine che penalizza la complessità delle foglie, limitando il rischio di overfitting e migliorando la generalizzazione del modello.

   Un aspetto distintivo di XGBoost è l'uso sia della derivata prima (gradiente) sia della seconda (hessiano) della funzione di perdita per approssimare il miglioramento apportato da ciascun nuovo albero. Questo consente una convergenza più rapida e stabile rispetto ad altri algoritmi di boosting. Per gestire dataset molto grandi, XGBoost introduce inoltre un algoritmo di *split finding* approssimato, basato sulla distribuzione dei gradienti, riducendo i tempi di calcolo senza compromettere l'accuratezza. L'algoritmo è anche in grado di gestire automaticamente i valori mancanti, imparando la direzione migliore da prendere quando incontra dati assenti.

   Nonostante il boosting sia un processo sequenziale, XGBoost ottimizza l'efficienza computazionale attraverso la parallelizzazione della ricerca dello split all'interno di ciascun albero.

   In sintesi, XGBoost unisce una funzione obiettivo regolarizzata, che limita l'overfitting, a una serie di ottimizzazioni che ne aumentano la velocità e la scalabilità, rendendolo uno degli algoritmi più efficaci e diffusi per problemi di classificazione e regressione su dataset complessi e di grandi dimensioni [12].

2.3.1 Addestramento dei Modelli

In seguito viene descritta la metodologia di addestramento dei modelli per il problema specifico che stiamo considerando. Ogni modello è stato addestrato sui dati trasformati tramite CLR e comprendenti tutte le 4630 feature microbiotiche.

Per tutti i modelli, quando applicabile, sono stati fissati `random_state = 42` e `n_jobs`$=-1$ per garantire riproducibilità e velocità nella fase di training.

Random Forest (RF)

Il modello Random Forest è stato configurato con 500 stimatori per garantire la stabilità della varianza delle predizioni.

Inoltre, per gestire il lieve sbilanciamento tra Responder e Non-Responder, è stato utilizzato il parametro `class_weight="balanced"` che, invece di dare la stessa importanza a ogni campione, assegna automaticamente pesi inversamente proporzionali alle frequenze delle classi nel dataset di input.

Extra Trees (ET)

L'algoritmo Extra Trees è stato implementato in modo analogo al Random Forest con 500 stimatori e impostando il parametro `class_weight="balanced"`.

Come descritto in precedenza, questo modello seleziona le soglie di split in modo totalmente casuale per ciascuna caratteristica, riducendo ulteriormente il rischio di overfitting su dati microbiomici rumorosi.

XGBoost (XGB)

Per il modello XGBoost, data la sua sensibilità agli iperparametri, è stata scelta una configurazione ottimizzata per massimizzare la capacità di generalizzazione e contrastare l'overfitting.

Sono stati utilizzati 400 stimatori con `learning_rate`=0.05. È stato scelto un learning rate basso (0.05) per garantire aggiornamenti graduali e stabili, favorendo la cattura di pattern biologici deboli e prevenendo l'overfitting. Il numero di stimatori è stato adeguato a questo learning rate per consentire una convergenza efficace.

La profondità massima degli alberi (`max_depth`=4) è stata limitata per agire come forma di regolarizzazione strutturale, evitando che il modello impari rumore specifico del campione in un contesto ad alta dimensionalità, tipico dei dataset microbiomici, e favorendo la generalizzazione su pattern biologici più ampi.

Per gestire l'elevata sparsità dei dati, sono stati fissati `subsample`=0.8 e `colsample_bytree`=0.8, in modo che ogni albero venga addestrato solo su una porzione (80%) dei campioni e delle feature, prevenendo la memorizzazione di pattern specifici del rumore.

Infine, per la classificazione binaria, è stata utilizzata la funzione obiettivo `binary:logistic`, monitorando le performance tramite la metrica `logloss`.

2.3.2 Analisi dei Risultati

Le performance di ciascun modello sono state valutate attraverso una *Stratified k-Fold Cross-Validation*, con $k = 5$.

A differenza del *k-fold standard*, che suddivide i dati in modo casuale, la versione stratificata assicura che ogni fold mantenga circa la stessa proporzione di campioni per ciascuna classe presente nel dataset originale [13].

Ad ogni iterazione, il modello viene addestrato su $k - 1$ fold, mentre il fold rimanente viene usato nella fase di test.

Nonostante il dataset sia quasi bilanciato, la stratificazione è considerata una procedura standard per assicurare che il modello venga valutato in condizioni coerenti durante tutto il processo di cross-validation.

Questa procedura garantisce che ogni partizione del dataset mantenga la proporzione originale tra Responder e Non-Responder, fornendo stime medie robuste.

In questo lavoro si assume come classe positiva la Classe 1, corrispondente ai Non-Responder.

Le metriche considerate sono:

- **Accuracy**: misura la proporzione di previsioni corrette sul totale delle osservazioni.
- **ROC-AUC**: Area Under the Receiver Operating Characteristic Curve. Misura la capacità del modello di distinguere tra le due classi considerando tutte le possibili soglie decisionali. Un valore pari a 0.5 indica prestazioni casuali, mentre valori prossimi a 1 indicano un'elevata capacità discriminativa.
- **Precision**: misura la proporzione di previsioni positive corrette rispetto al totale delle previsioni positive effettuate dal modello.
- **Recall**: misura la capacità del modello di identificare tutti i casi positivi reali, riducendo i falsi negativi.
- **F1-score**: rappresenta la media armonica tra Precision e Recall, fornendo una misura bilanciata tra le due.

I risultati medi e le relative deviazioni standard delle metriche di accuratezza, ROC-AUC, precision, recall e F1-score sono riassunti nella seguente tabella:

Modello	**Accuracy**	**ROC-AUC**	**Precision**	**Recall**	**F1-score**
RF	0.554 ± 0.034	0.585 ± 0.030	0.569 ± 0.024	0.707 ± 0.055	0.630 ± 0.033
ET	0.585 ± 0.025	0.596 ± 0.034	0.589 ± 0.016	0.759 ± 0.054	0.663 ± 0.030
XGB	0.594 ± 0.009	0.631 ± 0.032	0.61 ± 0.013	0.687 ± 0.024	0.646 ± 0.010

Tabella 2.1: Confronto delle prestazioni medie ottenute per i tre modelli testati per tutte le metriche (Precision, Recall e F1-score si riferiscono alla Classe 1).

Come mostrato in Tabella 2.1, il modello XGBoost ha dimostrato la capacità di discriminazione più elevata, raggiungendo una ROC-AUC di 0.631. Questo valore, sebbene possa apparire contenuto rispetto ad altri ambiti del machine learning, è considerato un risultato solido e realistico nel campo del microbiota intestinale. La complessità dei dati metagenomici, caratterizzati da una forte sparsità, impone intrinsecamente un limite superiore alla capacità predittiva dei modelli [43].

Un dato particolarmente rilevante è la ridotta deviazione standard dell'accuratezza per XGBoost (± 0.009). Questo indica stabilità del modello e un

apprendimento coerente dei pattern batterici, indipendentemente dalla specifica suddivisione dei dati.

L'F1-score della Classe 1 (Non-Responder), che si attesta mediamente sopra lo 0.63 per tutti i modelli, conferma che la strategia di bilanciamento adottata (parametri di regolarizzazione e pesi delle classi) ha funzionato correttamente. Il modello ha mostrato una buona capacità di identificare correttamente i pazienti che non rispondono alla terapia.

In conclusione, i risultati indicano che il profilo tassonomico dei pazienti contiene un segnale predittivo reale, ma debole, evidenziando come la complessità del fenomeno richieda modelli più ricchi o dati aggiuntivi per ottenere prestazioni più elevate.

2.4 Metodologie di Spiegabilità (XAI)

L'obiettivo centrale di questo lavoro di tesi risiede nella determinazione di quali taxa batterici influenzino maggiormente la risposta terapeutica.

È doveroso precisare che, data la natura complessa del microbiota intestinale, i risultati ottenuti non intendono stabilire una verità biologica assoluta o definitiva. Tuttavia, attraverso l'impiego di diverse tecniche di Explainable AI (XAI), ci si prefigge di valutare in modo consistente e incrociato i pattern identificati dai modelli.

La convergenza di algoritmi differenti (Random Forest, Extra Trees, XGBoost) e di diverse metriche di importanza (SHAP, Permutation Importance, Ablation Importance) permette di filtrare il rumore di fondo e identificare quei biomarcatori che mostrano una rilevanza statistica e biologica ripetibile, fornendo così una base solida per future ipotesi cliniche.

Ognuno dei tre algoritmi di Explainability, descritti nel Capitolo 1, è stato applicato ai tre modelli: Random Forest, Extra Trees e XGBoost.

I risultati ottenuti e i relativi commenti saranno descritti nel Capitolo 3.

2.4.1 Analisi tramite SHAP

In seguito alla scelta e all'uso di modelli tree based per la classificazione, è stato implementato il framework SHAP utilizzando l'algoritmo specifico `TreeExplainer`.

Questa scelta permette di ottenere una stima precisa del contributo di ciascuna feature, sfruttando la struttura ad albero dei modelli utilizzati.

La configurazione sperimentale ha previsto i seguenti passaggi tecnici:

1. **Target di analisi**: i valori SHAP sono stati calcolati per la Classe 0 (Responder). In questo contesto binario, un valore SHAP positivo indica un incremento della probabilità di successo terapeutico.

2. **Dataset**: il calcolo dei valori SHAP è stato eseguito sul test set (`X_test`), ovvero su dati non visti dai modelli durante il training. Questa scelta è fondamentale per valutare la capacità dei modelli di identificare biomarcatori generalizzabili e non legati a pattern specifici del dataset di addestramento.

3. **Visualizzazione**: per ogni modello sono state generate due tipologie di rappresentazioni:

    - **Bar Plot**: utilizzato per valutare l'importanza globale delle feature, basata sulla media dei valori assoluti dei contributi SHAP, evidenziando i 20 taxa con il maggiore impatto complessivo sulla predizione del modello.
    - **Beeswarm Plot**: utilizzato per analizzare l'impatto locale di ciascuna feature, visualizzando la relazione tra abbondanza relativa dei taxa e il segno del contributo alla predizione.

Questa analisi consente di interpretare in maniera dettagliata le decisioni del modello, identificando i taxa più influenti e comprendendo come la loro abbondanza relativa contribuisca al successo terapeutico.

2.4.2 Identificazione delle feature rilevanti

Data l'elevata dimensionalità del dataset (4630 feature per 569 osservazioni), è stata adottata una strategia di identificazione delle variabili più rilevanti basata sui valori SHAP.

Nella fase iniziale, i modelli Random Forest, Extra Trees e XGBoost sono stati addestrati utilizzando l'intero set di feature microbiotiche. Successivamente, sono stati calcolati i valori SHAP per ciascun modello, al fine di stimare il contributo di ogni singola feature alla predizione.

Sulla base dei valori medi assoluti di SHAP, è stato possibile individuare le feature con il maggiore impatto sul comportamento dei modelli.

Per ciascun algoritmo, sono state selezionate le prime 50 feature con valore medio assoluto più elevato. Questo sottoinsieme è stato poi impiegato nelle analisi di interpretabilità mediante Permutation Importance, mentre le prime 20 feature più rilevanti secondo SHAP sono state utilizzate per l'Ablation Importance.

Tali scelte hanno permesso di ridurre la dimensionalità del problema e di limitare il costo computazionale delle analisi.

2.4.3 Analisi tramite Permutation Feature Importance

I risultati ottenuti tramite SHAP sono stati sottoposti a una procedura di validazione statistica mediante Permutation Feature Importance (PFI). Questa fase è cruciale per confermare che i taxa identificati non siano solo correlati al risultato, ma siano effettivamente necessari per la performance del modello.

La configurazione tecnica della Permutation Importance ha previsto quanto segue:

1. **Selezione delle Top-50 SHAP**: per ottimizzare il costo computazionale e ridurre il rumore, la PFI è stata eseguita sulle prime 50 feature identificate come più rilevanti tramite SHAP per ciascun modello.

2. **Metrica di scoring**: come criterio di valutazione è stata utilizzata la metrica ROC-AUC. Questa metrica è particolarmente adatta per dataset sbilanciati, e rispetto ad altre metriche come accuracy o F1-score, fornisce una misura più robusta di quanto una feature contribuisca effettivamente alla separazione tra le due classi. In questo caso, il punteggio risultante misura quanto la capacità discriminante del modello diminuisce quando i valori di una specifica feature vengono casualmente rimescolati.

3. **Robustezza statistica**: l'analisi è stata ripetuta per 20 iterazioni (`n_repeats=20`) per ciascuna feature, al fine di garantire che l'importanza stimata sia stabile e non dovuta a fluttuazioni casuali.

Per ciascun algoritmo è stato infine generato un grafico di Permutation Feature Importance, che mostra l'impatto della permutazione delle singole feature sulla metrica ROC-AUC.

I valori della PFI rappresentano la diminuzione media della ROC-AUC quando i valori di una feature vengono casualmente permutati. Ognuno di questi valori va interpretato come misura di quanto la performance complessiva del modello dipenda da quella feature. Più alto è il valore, maggiore è l'importanza della feature, valori prossimi a zero, invece, indicano feature poco rilevanti.

2.4.4 Analisi tramite Ablation Feature Importance

L'ultima fase del framework di spiegabilità consiste nel calcolo della Ablation Feature Importance (AFI), che prevede la rimozione sistematica di ciascuna feature dal dataset, seguita da un riaddestramento completo dell'algoritmo.

La configurazione tecnica dell'Ablation ha previsto quanto segue:

1. **Selezione delle Top-20 SHAP**: per limitare il costo computazionale dell'analisi, l'Ablation è stata eseguita sulle prime 20 feature identificate come più rilevanti tramite SHAP per ciascun modello.

2. **Metrica di scoring**: come criterio di valutazione, analogamente alla Permutation Importance, è stata utilizzata la metrica ROC-AUC.

3. **Baseline di confronto**: inizialmente viene addestrata un'istanza del modello completo (`rf_full`, `et_full`, `xgb_full`) utilizzando l'intero set di feature, al fine di stabilire il valore di riferimento della ROC-AUC sul test set.

4. **Rimozione sistematica**: per ciascuno dei 20 taxa identificati come più influenti dall'analisi SHAP viene generato un nuovo dataset privato esclusivamente di quella specifica caratteristica.

5. **Refitting e scoring**: per ciascun dataset "ablato" viene addestrata una nuova istanza del modello (`rf_temp`, `et_temp`, `xgb_temp`) utilizzando la stessa configurazione iperparametrica del modello originale, calcolando successivamente la nuova performance sul test set.

6. **Quantificazione dell'importanza**: l'Ablation Feature Importance viene calcolata come variazione della performance rispetto alla baseline:

$$\Delta AUC = AUC_{\text{baseline}} - AUC_{\text{ablation}}$$

Questa procedura permette di distinguere i biomarcatori critici da quelli ridondanti.

Un valore elevato di ΔAUC indica che la rimozione della feature provoca una significativa riduzione della performance del modello, suggerendo che l'informazione contenuta in quel batterio è informativa e rilevante per la predizione della risposta terapeutica.

Al contrario, valori prossimi allo zero indicano che il modello è in grado di compensare la perdita della feature utilizzando informazioni simili provenienti da altri taxa presenti nel microbiota.

Infine, valori negativi possono indicare che la rimozione della feature non penalizza il modello o può addirittura migliorarne leggermente la performance, suggerendo una possibile ridondanza o presenza di rumore informativo.

I risultati dell'analisi sono riportati in un grafico che indica l'incremento o il decremento della performance.

2.5 Identificazione delle variabili biologiche rilevanti

Al termine della fase di interpretazione dei modelli, è stata condotta un'analisi conclusiva volta a identificare le variabili biologiche più associate alla risposta all'immunoterapia, con l'obiettivo di selezionare le feature potenzialmente rilevanti comuni ai tre modelli.

In una prima fase sono state selezionate le 50 feature più rilevanti identificate tramite SHAP per ciascun modello (Random Forest, Extra Trees e XGBoost).

Successivamente, è stata calcolata l'intersezione tra tali insiemi di feature, al fine di individuare i taxa che risultano sistematicamente rilevanti indipendentemente dall'algoritmo utilizzato. Questa procedura consente di identificare segnali più robusti, riducendo il rischio di individuare pattern specifici di un singolo modello.

Per i taxa appartenenti a questa intersezione è stata successivamente analizzata la distribuzione delle abbondanze nel dataset originale.

In particolare, per ciascun taxon è stata rappresentata graficamente la distribuzione dei valori nelle due classi target (Responder e Non-Responder) mediante boxplot affiancati, accompagnati dalla visualizzazione dei singoli campioni tramite scatter plot. Considerata l'elevata variabilità tipica dei dati microbiotici, le distribuzioni sono state rappresentate utilizzando una scala logaritmica sull'asse delle ordinate.

Successivamente è stata applicata un'analisi di Ablation Feature Importance sul sottoinsieme delle feature rilevanti identificate nella fase di intersezione. Per ciascun modello è stata eseguita la rimozione iterativa di ogni feature selezionata, seguita dal riaddestramento del modello, consentendo di valutare l'impatto della sua assenza sulla performance predittiva.

I risultati sono stati rappresentati graficamente tramite un barplot comparativo che mostra, per ciascun taxon, il ΔAUC sui tre modelli, calcolato come la differenza tra $AUC_{baseline} - AUC_{senza\_feature}$.

Questa procedura permette di distinguere le variabili realmente indispensabili da quelle potenzialmente ridondanti e di visualizzare immediatamente eventuali differenze tra modelli.

Parallelamente è stata effettuata un'analisi descrittiva delle abbondanze medie per tutte le 50 feature selezionate tramite SHAP. In particolare, per ciascun taxon è stata calcolata la differenza tra l'abbondanza media osservata nei pazienti Responder e quella nei pazienti Non-Responder, al fine di ottenere un'indicazione preliminare della possibile associazione biologica con la risposta terapeutica.

Infine, per verificare la capacità predittiva dell'insieme di feature selezionate, è stata effettuata un'ulteriore analisi di *Extreme Feature Selection*.

Per ciascun modello è stato costruito un dataset ridotto contenente esclusivamente le 50 feature più rilevanti identificate tramite SHAP per quello specifico algoritmo.

I modelli sono stati quindi riaddestrati utilizzando tali dataset ridotti e valutati mediante una suddivisione train/test stratificata (80/20). Le prestazioni sono state misurate tramite le metriche ROC-AUC e accuracy, al fine di verificare se il sottoinsieme di variabili selezionate fosse sufficiente a preservare la capacità discriminante dei modelli.

2.6 Riassunto finale della pipeline

In questo lavoro è stata sviluppata una pipeline di analisi finalizzata all'identificazione di variabili biologiche rilevanti associate alla risposta all'immunoterapia a partire da dati di microbiota intestinale.

In una prima fase sono stati addestrati tre modelli di apprendimento automatico basati su alberi decisionali (Random Forest, Extra Trees e XGBoost), utilizzando l'intero set di feature microbiotiche disponibili.

Successivamente sono state applicate tecniche di Explainable Artificial Intelligence per interpretare il comportamento dei modelli e individuare le variabili più rilevanti. In particolare, i valori SHAP sono stati utilizzati per ottenere un ranking globale delle feature, dal quale sono state selezionate le 50 variabili con maggiore contributo medio alla predizione per ciascun modello.

Le feature individuate tramite SHAP sono state quindi sottoposte a ulteriori procedure di validazione. In primo luogo è stata applicata la Permutation Feature Importance, che consente di valutare la diminuzione delle prestazioni del modello quando l'informazione di una determinata variabile viene distrutta tramite permutazione casuale.

Successivamente è stata applicata l'Ablation Feature Importance, che prevede la rimozione sistematica di una feature dal dataset seguita dal riaddestramento del modello, al fine di valutare l'impatto della sua assenza sulla capacità predittiva.

Per aumentare la robustezza dei risultati, è stata inoltre calcolata l'intersezione tra le 50 feature più rilevanti identificate dai tre modelli. I taxa appartenenti a

questa intersezione rappresentano i candidati biomarcatori più stabili, in quanto risultano sistematicamente rilevanti indipendentemente dall'algoritmo utilizzato.

Per tali taxa è stata analizzata la distribuzione delle abbondanze nelle due classi target (responder e non-responder) tramite rappresentazioni grafiche basate su boxplot su scala logaritmica. Inoltre, per tutte le 50 feature selezionate tramite SHAP è stata calcolata la differenza tra l'abbondanza media osservata nei due gruppi di pazienti, fornendo un'indicazione preliminare della possibile associazione biologica con la risposta terapeutica.

Infine, è stata effettuata un'ulteriore analisi di *Extreme Feature Selection*, in cui ciascun modello è stato riaddestrato utilizzando esclusivamente le 50 feature più rilevanti identificate tramite SHAP per quello specifico algoritmo. Le prestazioni sono state valutate utilizzando le metriche ROC-AUC e accuracy.

Questa pipeline consente quindi di combinare modelli di machine learning e tecniche di interpretabilità per individuare variabili biologiche rilevanti in maniera robusta, integrando analisi predittive e analisi descrittive dei dati microbiotici.

Capitolo 3

Risultati dell'Analisi Tramite Metodi di Explainable AI

In questo capitolo vengono presentati e discussi i risultati ottenuti dall'applicazione dei metodi di Explainable AI e delle tecniche di selezione delle feature, introdotti nel Capitolo 2.

L'analisi si concentra sull'interpretazione dei modelli predittivi mediante tecniche di Explainable AI, con l'obiettivo di individuare potenziali biomarcatori microbici associati alla risposta terapeutica. In particolare, si intende valutare la robustezza e la consistenza dei segnali emersi attraverso il confronto tra diversi algoritmi e differenti metriche di importanza delle variabili.

Dopo aver applicato modelli basati su alberi decisionali e averne valutato le performance — i cui risultati sono riportati nel Capitolo 2, Sezione 2.3.2 — ottenendo risultati complessivamente in linea con quanto riportato in letteratura, e nei lavori precedenti a cui si fa riferimento, si è proceduto con l'impiego di tecniche di Explainable AI per una più approfondita interpretazione dei modelli.

Per ciascun modello di machine learning vengono riportati e commentati i principali grafici ottenuti. L'analisi è strutturata come un confronto a tre livelli per quanto riguarda le tecniche di explainability, con l'obiettivo di evidenziare analogie e differenze nei risultati ottenuti.

Successivamente, l'attenzione si concentra sui risultati derivanti dall'intersezione delle 50 feature più rilevanti individuate per ciascun modello, con l'obiettivo di identificare un insieme di variabili robuste e condivise, potenzialmente interpretabili come biomarcatori affidabili.

3.1 SHAP TreeExplainer

Il primo modello di interpretabilità considerato è SHAP (SHapley Additive exPlanations), la cui descrizione è presentata nel Capitolo 1, Sottosezione 1.3.3, che rappresenta il principale strumento su cui si basa l'intera analisi.

L'obiettivo è duplice: da un lato, ottenere una misura di importanza globale, al fine di identificare quali feature contribuiscano maggiormente al processo decisionale dei modelli; dall'altro, condurre un'analisi locale per comprendere la direzione dell'impatto di ciascuna feature, ovvero verso quale classe — Responder

(Classe 0) o Non-Responder (Classe 1) — una specifica abbondanza batterica orienti la predizione.

Nel caso specifico, come descritto nel Capitolo 2, Sezione 2.4, l'output del modello è definito rispetto alla Classe 0, pertanto, i valori di SHAP devono essere interpretati in relazione a tale classe.

L'analisi è stata applicata in modo comparativo ai tre modelli addestrati: Random Forest, Extra Trees e XGBoost. Tale confronto risulta fondamentale per valutare la consistenza dei risultati: un'eventuale convergenza di modelli differenti verso gli stessi biomarcatori rappresenterebbe infatti un'evidenza a supporto della robustezza del segnale biologico individuato.

Per ciascun modello sono stati prodotti due tipi di grafici: il *Bar plot*, utilizzato per quantificare l'importanza globale delle feature, e il *Beeswarm plot*, impiegato per analizzare la distribuzione e la direzione dell'impatto dei valori di SHAP.

Il Bar plot riporterà, per ogni modello, le 20 feature più influenti, ordinate in base al valore medio assoluto dei corrispondenti SHAP values.

Il Beeswarm plot riporterà per ogni feature i punti. Il colore rappresenta il valore della feature (rosso = alta abbondanza, blu = bassa abbondanza), mentre la posizione sull'asse orizzontale indica il contributo alla predizione.

Per ragioni di leggibilità, nei grafici vengono riportati esclusivamente i livelli tassonomici di famiglia (*f_*), genere (*g_*) e specie (*s_*).

Tuttavia, al fine di garantire completezza e tracciabilità dei risultati, una tabella contenente i nomi completi delle 20 feature principali, relative a ciascun modello, e i corrispondenti valori di SHAP, è riportata in Appendice B.

3.1.1 Analisi dei risultati: Random Forest

Il primo modello analizzato attraverso il framework SHAP è il Random Forest.

Sebbene questo modello abbia mostrato performance leggermente inferiori rispetto ad Extra Trees e XGBoost (Capitolo 2, Sezione 2.3.2), la sua inclusione nell'analisi risulta rilevante per valutare la stabilità dei biomarcatori individuati al variare dell'algoritmo di apprendimento.

Bar Plot

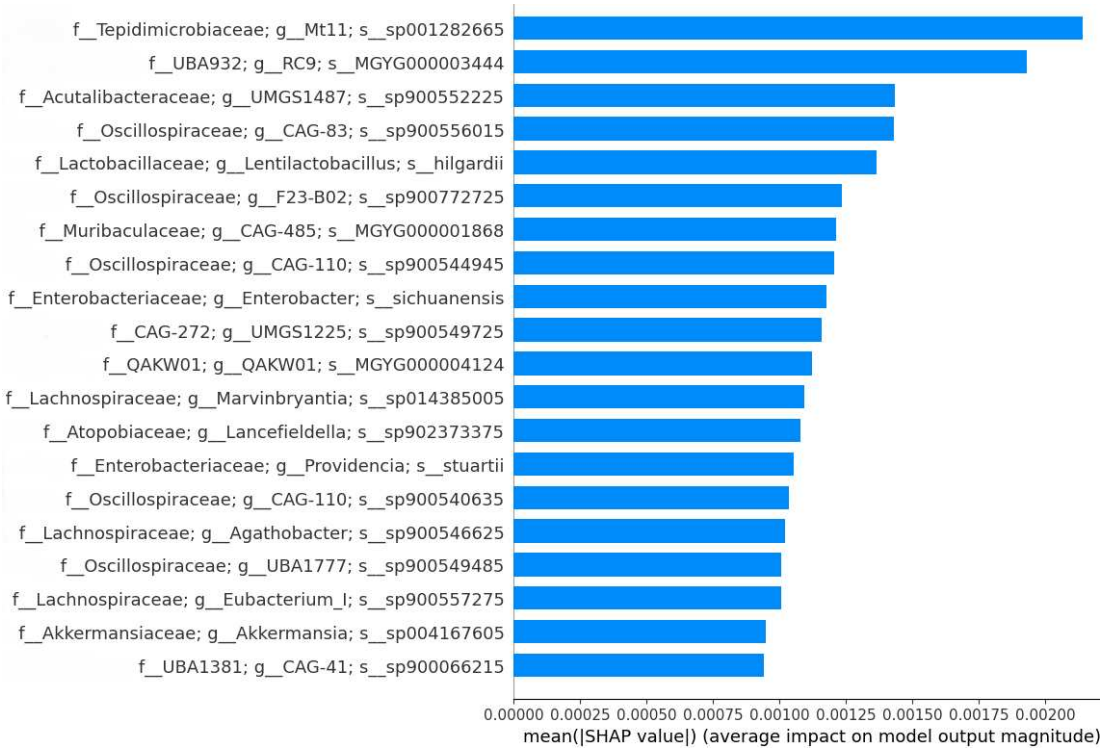

Figura 3.1: SHAP Summary Bar Plot del modello Random Forest.

Dalla Figura 3.1 possiamo osservare come il Random Forest identifichi una gerarchia di importanza delle feature guidata da due principali taxa:

1. s_sp001282665 (classe *Clostridia*, famiglia *Tepidimicrobiaceae*): questa feature rappresenta il principale driver predittivo all'interno del modello, indicando che le sue variazioni di abbondanza contribuiscono maggiormente alla discriminazione tra pazienti Responder e Non-responder.

2. s_MGYG000003444 (ordine *Bacteroidales*): si colloca al secondo posto per impatto medio sulla predizione, confermandosi come uno dei taxa più informativi nel processo decisionale del modello.

Importante osservare i valori medi assoluti di SHAP, che risultano essere numericamente contenuti, con un massimo intorno allo 0.0020. Questo suggerisce che il segnale predittivo sia distribuito in modo diffuso su un'ampia varietà di feature, e non concentrato in pochi biomarcatori dominanti.

Interessante inoltre notare la decrescita graduale dell'importanza dopo i primi 10–15 taxa. Questo andamento suggerisce che il modello non si basi su un singolo biomarcatore, ma utilizzi un segnale combinato di più specie batteriche.

Infine, si osserva la ricorrenza di taxa appartenenti a famiglie affini. Questo potrebbe indicare che la Random Forest stia intercettando pattern strutturali della comunità microbica nel suo complesso, piuttosto che l'effetto isolato di singole specie.

In conclusione, coerentemente con la natura ensemble della Random Forest, il modello sfrutta contributi distribuiti e potenziali interazioni tra più taxa.

Beeswarm Plot

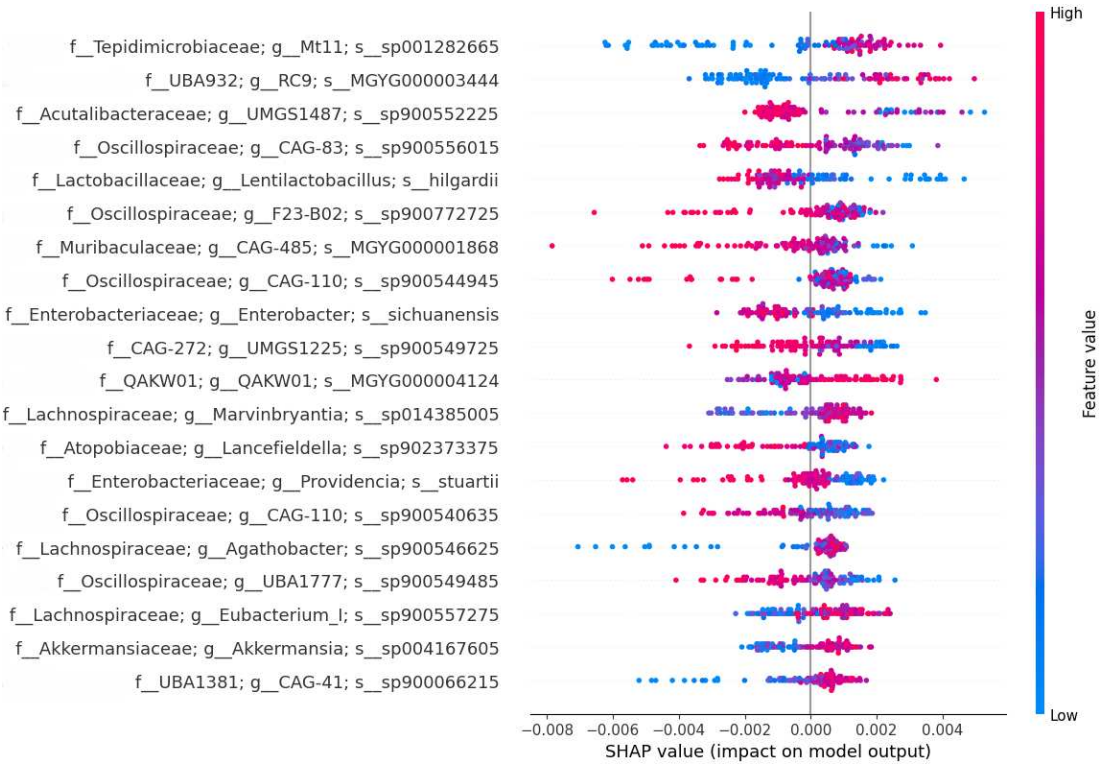

Figura 3.2: SHAP Beeswarm Plot del modello Random Forest.

Dalla Figura 3.2 si osserva che i due taxa ai vertici della classifica, s_sp001282665 e s_MGYG000003444, mostrano una netta separazione cromatica: i punti rossi, indicativi di elevata abbondanza, si posizionano prevalentemente a destra dello zero, suggerendo un contributo positivo alla predizione, mentre i punti blu, rappresentanti valori bassi di abbondanza, si collocano principalmente a sinistra. Ciò indica che un'elevata abbondanza di questi taxa incrementa la probabilità predetta di risposta, confermandone il ruolo informativo nel modello.

Dal punto di vista biologico, il primo taxa appartiene alla classe *Clostridia*, il secondo all'ordine *Bacteroidales*. In letteratura, questi batteri risultano essere controversi: alcune specie possono avere effetti patogeni, mentre altre risultano benefiche [10], [53].

Anche la terza feature, s_sp900552225 (famiglia *Acutalibacteraceae*), mostra una distinzione tra punti rossi e blu; in questo caso, però, un'elevata abbondanza contribuisce negativamente alla predizione, associandosi alla classe Non-responder.

Pattern simili, seppur meno marcati, si osservano in altri taxa, suggerendo un contributo cumulativo di più feature.

Nella parte inferiore della classifica, taxa come s_sp900557275 (*Lachnospiraceae*), s_sp004167605 (*Akkermansiaceae*) e s_sp900066215 (*Monoglobales_A*) evidenziano un'associazione con risposta positiva.

Per molti taxa, l'elevata densità di punti vicino allo zero conferma che variazioni locali di abbondanza producono contributi marginali, coerentemente con un segnale predittivo distribuito basato sulla composizione complessiva del microbiota. In alcuni casi, punti rossi a sinistra indicano che un'elevata abbondanza contribuisce negativamente alla predizione. Infine, alcuni taxa non mostrano separazioni chiare tra punti rossi e blu, suggerendo effetti emergenti principalmente in interazione con altre componenti del microbiota.

3.1.2 Analisi dei risultati: Extra Trees

Passiamo ora all'analisi del secondo modello valutato tramite SHAP: Extra Trees. Questo algoritmo, pur appartenendo alla famiglia degli ensemble tree-based come la Random Forest, introduce una maggiore randomizzazione nella scelta delle soglie di split, influenzando la distribuzione dell'importanza delle feature.

Dal punto di vista delle performance, il modello Extra Trees mostra risultati intermedi rispetto agli altri modelli considerati.

In linea con quanto osservato per la Random Forest, ci si aspetta di riscontrare una certa coerenza nei taxa identificati. Una convergenza tra modelli basati su diverse strategie di costruzione degli alberi rappresenterebbe infatti un'ulteriore evidenza della robustezza del segnale biologico, suggerendo che le feature individuate siano realmente informative e non dipendenti da specifiche scelte modellistiche.

Bar Plot

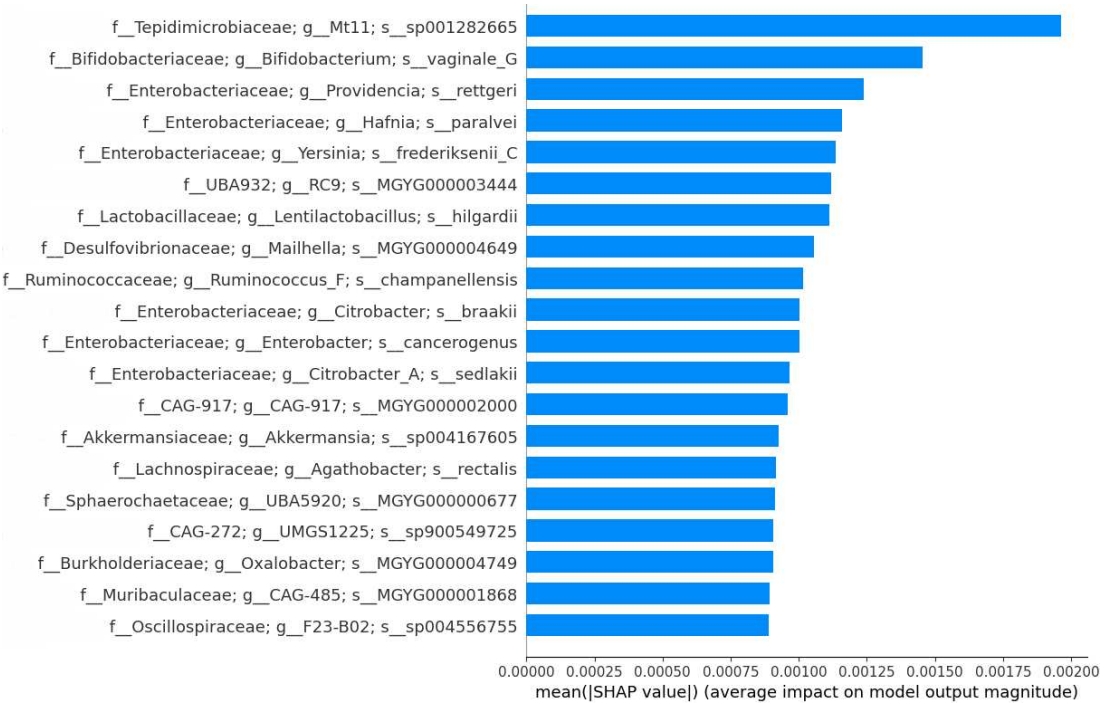

Figura 3.3: SHAP Summary Bar Plot del modello Extra Trees.

Dalla Figura 3.3 si osserva che il batterio s_sp001282665 (ordine *Clostridia*, famiglia *Tepidimicrobiaceae*) si conferma stabilmente al primo posto anche nel modello Extra Trees, con un'importanza media nettamente superiore a tutte le altre variabili.

A differenza di quanto rilevato per la Random Forest, nel modello Extra Trees si nota una variazione nella gerarchia dei taxa: nelle prime posizioni compaiono nuovi taxa come s_vaginale_G (famiglia *Bifidobacteriaceae*) e diversi rappresentanti della famiglia *Enterobacteriaceae*, tra cui s_rettgeri (genere *Providencia*) e s_paralvei (genere *Hafnia*). Questo suggerisce che il maggiore grado di randomizzazione introdotto da Extra Trees consenta di catturare segnali associati a taxa meno evidenti in altri modelli.

Il taxon s_MGYG000003444 (ordine *Bacteroidales*), pur scendendo leggermente di posizione rispetto alla Random Forest, rimane saldamente nella top 10, confermando la sua rilevanza nel processo decisionale.

Il decremento dei valori SHAP tra la seconda e la ventesima posizione appare piuttosto graduale, indicando una distribuzione del contributo delle feature più omogenea.

Anche per Extra Trees, i valori SHAP complessivi risultano contenuti e molto simili a quelli della Random Forest, suggerendo una scala di impatto delle feature comparabile tra i due modelli.

Beeswarm Plot

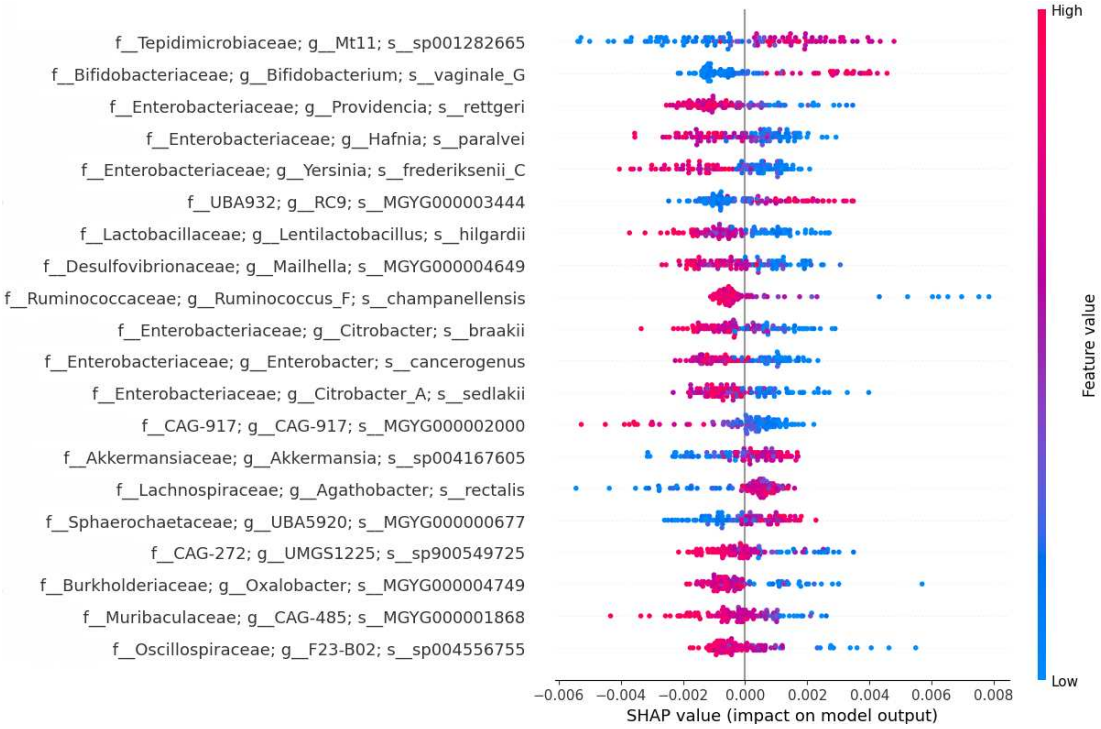

Figura 3.4: SHAP Beeswarm Plot del modello Extra Trees.

Dalla Figura 3.4 si osserva che, per i principali biomarcatori già identificati nel modello precedente, come s_sp001282665 e s_MGYG000003444, emerge una

distribuzione coerente: i punti rossi, rappresentativi di un'elevata abbondanza, risultano nettamente spostati verso destra, indicando una relazione positiva tra l'abbondanza di tali taxa e la probabilità di classificazione come Responder, anche nel modello Extra Trees.

È particolarmente interessante analizzare la direzione dell'impatto per i taxa collocati nelle prime posizioni del ranking. Per il taxon s_vaginale_G (famiglia *Bifidobacteriaceae*) i punti rossi si concentrano sul lato positivo dell'asse delle ascisse, suggerendo un contributo favorevole alla risposta. Questo è coerente con quanto riportato in letteratura, dove il genere *Bifidobacterium* è frequentemente associato a effetti benefici per la salute e a una modulazione positiva della risposta immunitaria dell'ospite [24].

Al contrario, per alcuni rappresentanti della famiglia *Enterobacteriaceae*, anch'essi ben posizionati nel ranking, si osserva una tendenza dei valori elevati verso sinistra, indicando una possibile associazione con una minore probabilità di risposta. Questo risulta in linea con studi che collegano la presenza di questi batteri nel microbiota intestinale a stati di disbiosi e a esiti meno favorevoli della risposta terapeutica, in particolare nelle terapie oncologiche e immunomodulanti [5].

Le feature posizionate nella parte inferiore della classifica mostrano valori SHAP prevalentemente concentrati attorno allo zero, suggerendo un contributo limitato alla predizione e una distribuzione più diffusa del segnale tra numerose variabili.

Infine, solo un numero ristretto di taxa presenta una chiara associazione tra elevata abbondanza e contributo positivo alla predizione. Tra questi, oltre ai biomarcatori principali, emergono s_sp004167605 (famiglia *Akkermansiaceae*) e s_rectalis (famiglia *Lachnospiraceae*), che, sebbene con distribuzioni centrate intorno allo zero, mostrano una presenza significativa di punti rossi orientati verso la direzione di risposta. Questi risultati risultano coerenti con la letteratura, che associa tali taxa a uno stato favorevole del microbiota intestinale [42], [36].

3.1.3 Analisi dei risultati: XGBoost

Passiamo ora all'analisi dell'ultimo modello valutato tramite il framework SHAP: XGBoost. Questo algoritmo ha mostrato la migliore capacità predittiva nella nostra sperimentazione.

Basato su una logica di Gradient Boosting, XGBoost è particolarmente efficace nel catturare interazioni complesse tra i taxa, fornendo una mappatura dei biomarcatori estremamente dettagliata e precisa.

Bar Plot

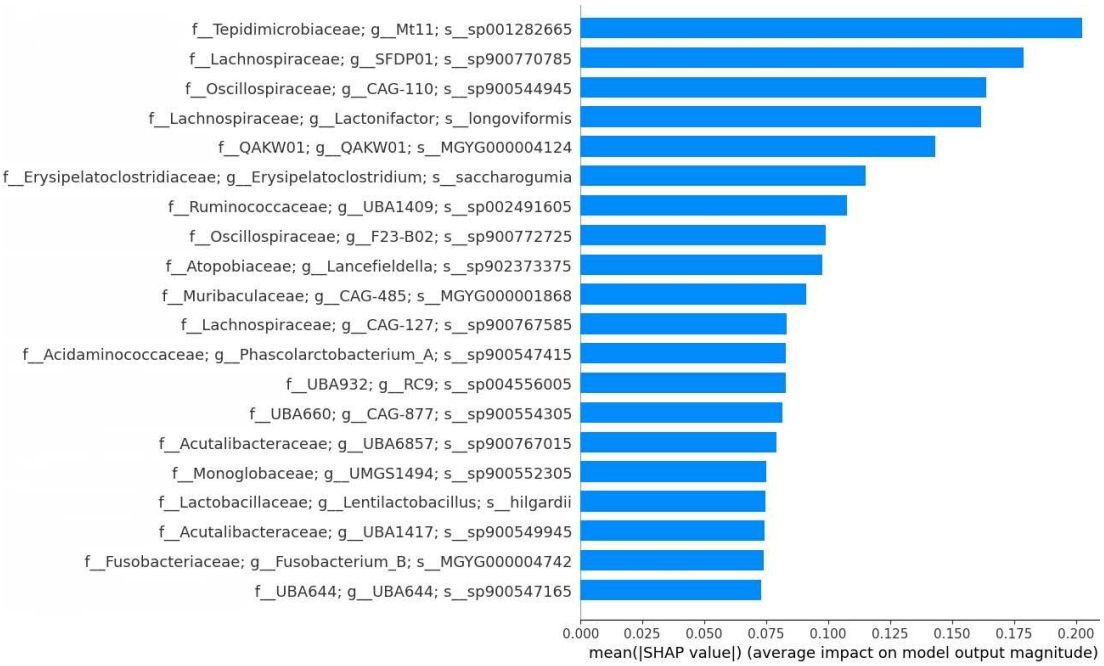

Figura 3.5: SHAP Summary Bar Plot del modello XGBoost.

Dalla Figura 3.5 emerge una gerarchia delle feature fortemente concentrata su un numero limitato di taxa dominanti.

In particolare, il taxon s__sp001282665 (ordine *Clostridia*, famiglia *Tepidimicrobiaceae*) si conferma stabilmente al primo posto, con un'importanza media assoluta nettamente superiore rispetto a tutte le altre variabili, evidenziando il suo ruolo come principale driver predittivo della risposta terapeutica.

Tra i taxa più rilevanti compaiono anche s__sp900770785 (famiglia *Lachnospiraceae*) e s__sp900544945 (ordine *Oscillospirales*), che nei modelli Random Forest ed Extra Trees non emergevano o avevano un impatto medio significativamente inferiore. La loro presenza nelle prime posizioni indica una diversa distribuzione dell'importanza delle feature in XGBoost.

Un aspetto particolarmente rilevante riguarda la scala dei valori SHAP: rispetto ai modelli Random Forest ed Extra Trees, XGBoost presenta valori medi assoluti sensibilmente più elevati (fino a circa 0.200). Ciò indica che il modello attribuisce un peso maggiore a poche feature, rendendo la predizione fortemente dipendente da taxa altamente informativi.

Si osserva inoltre una decrescita relativamente rapida dell'importanza dopo le prime posizioni, in contrasto con l'andamento più graduale degli altri modelli. Questo comportamento suggerisce una maggiore concentrazione del contributo predittivo e una minore distribuzione dell'importanza tra le feature.

In conclusione, i risultati sono coerenti con la natura del Gradient Boosting, che tende a focalizzarsi su un insieme selezionato di feature ad alto impatto, producendo una rappresentazione più concentrata del segnale predittivo.

Beeswarm Plot

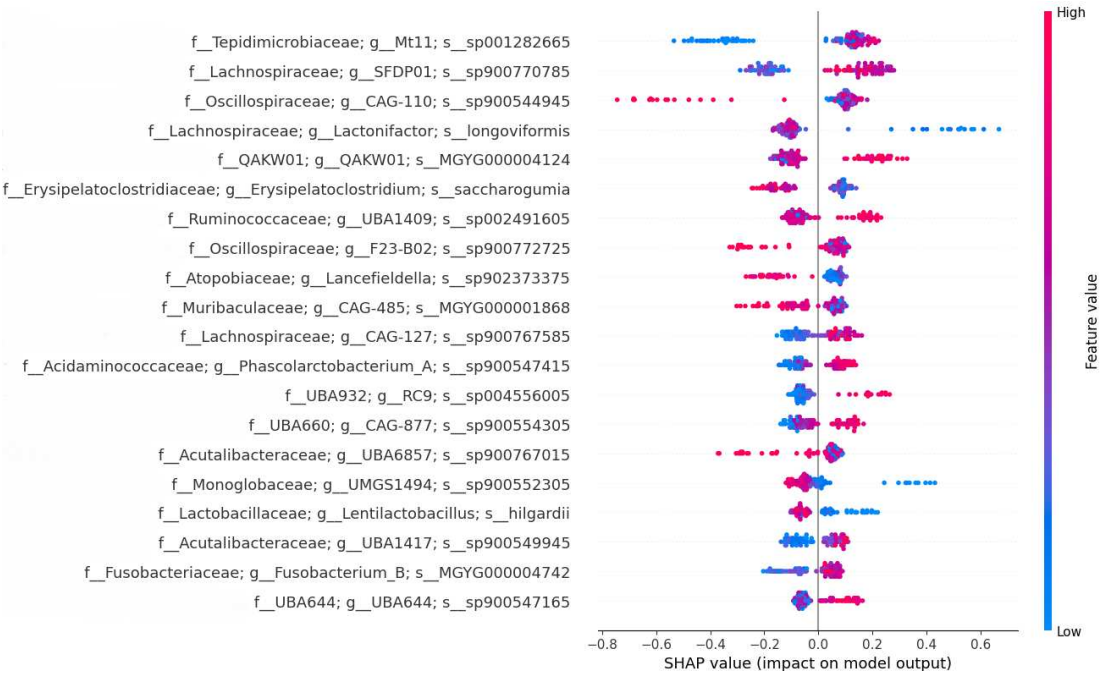

Figura 3.6: SHAP Beeswarm Plot del modello XGBoost.

Dalla Figura 3.6 si osserva, rispetto alle Figure 3.2 e 3.4, una separazione più marcata tra contributi positivi e negativi, con una minore concentrazione di punti intorno allo zero.

In particolare, per i taxa più importanti, la distribuzione appare chiaramente polarizzata: valori elevati della feature si collocano prevalentemente nella regione dei valori SHAP positivi, mentre valori bassi si associano a contributi negativi. Questo indica una relazione coerente tra abbondanza del taxon e direzione dell'effetto sulla predizione. Il taxon s_sp001282665 (ordine *Clostridia*) si conferma favorevole a una risposta positiva. Anche altri taxa ai vertici della classifica, come s_sp900770785 (famiglia *Lachnospiraceae*) e s_sp900544945 (ordine *Oscillospirales*), risultano associati a una possibile risposta positiva all'immunoterapia [36], [47].

Rispetto ai modelli precedenti, XGBoost mostra cluster di punti più compatti e ben separati, suggerendo una maggiore capacità di distinguere regioni dello spazio delle feature associate a effetti differenti sulla risposta.

Alcuni taxa presentano però una distribuzione più eterogenea, con punti sia a destra sia a sinistra dello zero. Questo indica che il loro contributo non è univoco, ma dipende dal contesto, suggerendo possibili interazioni con altre variabili o effetti non lineari.

Infine, alcuni taxa mostrano che valori elevati della feature si associano a contributi negativi, indicando una relazione inversa con la probabilità di risposta. Ciò evidenzia la presenza di driver con effetti opposti all'interno del microbiota.

Nel complesso, XGBoost evidenzia una struttura più definita e polarizzata dei contributi delle feature, coerente con una maggiore concentrazione del segnale predittivo su specifici taxa.

3.1.4 Sintesi Comparativa di SHAP

L'integrazione dei risultati SHAP per i tre modelli analizzati permette di delineare un quadro chiaro sull'importanza dei taxa nel determinare le predizioni dei modelli. Pur trattandosi di contributi alle singole predizioni e non di impatto globale sulle performance, emergono pattern significativi e consistenti.

Il taxon s_sp001282665 (ordine *Clostridia*, famiglia *Tepidimicrobiaceae*) si conferma il principale driver predittivo in tutti e tre i modelli.

Anche s_MGYG000003444 (ordine *Bacteroidales*) occupa posizioni di rilievo, suggerendo la robustezza di questi biomarcatori indipendentemente dall'architettura scelta.

Nei modelli Random Forest ed Extra Trees, il segnale predittivo è distribuito su numerosi taxa, con valori medi SHAP contenuti, indicando che le predizioni derivano da contributi cumulativi di più specie.

Nel modello XGBoost, invece, l'importanza è concentrata sui principali taxa, con valori SHAP più elevati e una decrescita rapida dopo le prime posizioni, indicando una maggiore dipendenza dai taxa dominanti.

I taxa principali mostrano effetti coerenti tra i modelli: un'elevata abbondanza di s_sp001282665 o s_MGYG000003444 aumenta la probabilità di classificazione come Responder. Alcuni taxa secondari presentano effetti meno definiti o non lineari, suggerendo possibili interazioni con altre feature.

Extra Trees identifica nuovi taxa rilevanti nelle prime posizioni, come s_vaginale_G (famiglia *Bifidobacteriaceae*) e rappresentanti della famiglia *Enterobacteriaceae*, catturando segnali meno evidenti per Random Forest.

XGBoost mostra una maggiore polarizzazione tra contributi positivi e negativi, con cluster di punti compatti e ben separati, enfatizzando pochi taxa dominanti.

SHAP conferma l'importanza di driver chiave comuni, ma evidenzia anche differenze nella distribuzione del segnale tra modelli.

3.2 Permutation Feature Importance

Il secondo metodo utilizzato per l'analisi dell'importanza delle feature è la Permutation Feature Importance (descritta nel Capitolo 1, Sezione 1.4). Questa tecnica agisce direttamente sulle performance del modello: l'importanza di una feature viene infatti misurata osservando quanto il punteggio di ROC-AUC diminuisce quando i valori di tale variabile vengono mescolati casualmente, distruggendo così il legame tra la feature e il target.

L'obiettivo è identificare quali taxa risultino indispensabili per il modello: una feature è considerata critica se la sua permutazione causa una riduzione significativa delle prestazioni, indicando che l'algoritmo si affida fortemente a quell'informazione per generalizzare correttamente.

Analogamente a quanto fatto per SHAP, l'analisi della Permutation Feature Importance viene applicata in modo comparativo ai tre modelli considerati (Random Forest, Extra Trees e XGBoost). Questo passaggio risulta cruciale per convalidare i biomarcatori precedentemente identificati: se un taxon emerge come rilevante sia in termini di contributo (SHAP) sia in termini di necessità predittiva (Permutation), e ciò avviene in modo consistente tra algoritmi differenti, la robustezza del suo ruolo come biomarcatore della risposta terapeutica può ritenersi ulteriormente rafforzata.

La Permutation Feature Importance viene applicata alle top 50 feature individuate tramite SHAP per ciascun modello, come specificato nel Capitolo 2, Sezione 2.4.3.

Viene infine presentato e discusso un grafico che riporta la variazione media della metrica ROC-AUC a seguito della permutazione casuale dei valori di ciascuna feature. Le barre rappresentano l'importanza relativa dei taxa, evidenziando quali risultino maggiormente indispensabili per mantenere la capacità predittiva del modello.

Per ragioni di leggibilità, i taxa verranno riportati utilizzando la nomenclatura tassonomica compresa tra il livello di ordine (*o_*) e quello di specie (*s_*).

3.2.1 Analisi dei risultati: Random Forest

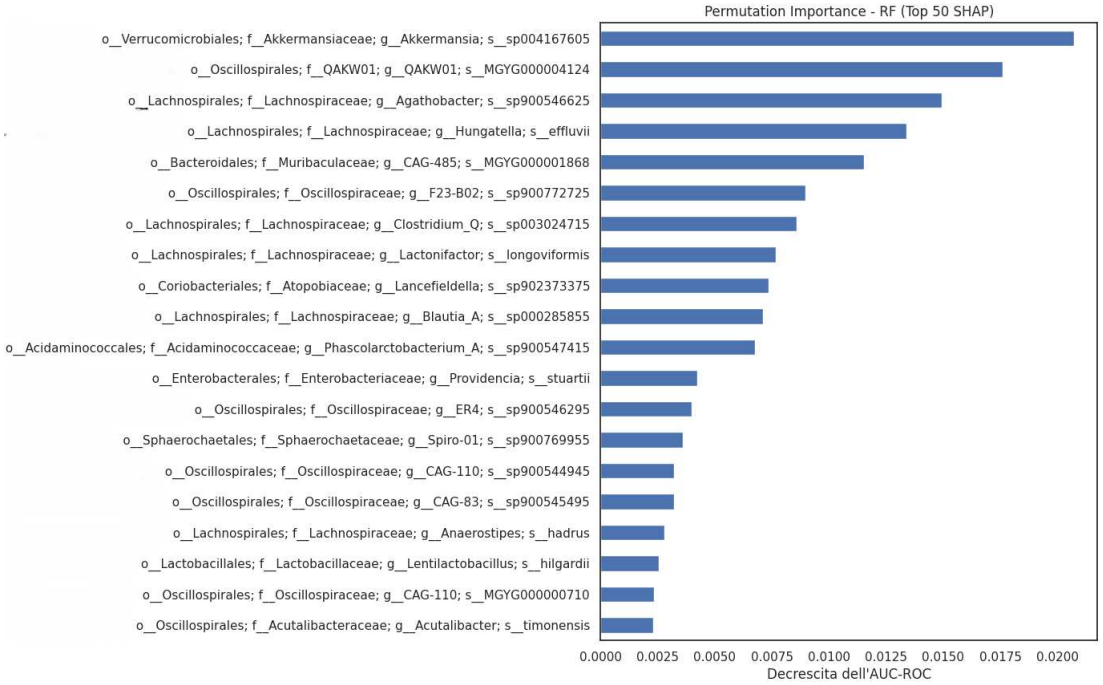

Figura 3.7: Permutation Feature Importance per il modello Random Forest.

Dalla Figura 3.7 si nota un cambiamento nel ranking delle feature rispetto a quanto emerso dall'analisi SHAP (Figura 3.1), a conferma che le due metodologie catturano aspetti differenti dell'importanza delle variabili.

L'analisi evidenzia come un numero ristretto di taxa contribuisca in maniera predominante alla capacità predittiva del modello. Le prime feature in classifica determinano infatti una riduzione significativamente più marcata della metrica ROC-AUC rispetto alle successive. In particolare, a partire dalle prime cinque posizioni, i valori tendono a stabilizzarsi in un intervallo compreso tra 0.0075 e 0.0025, suggerendo una forte dipendenza del modello da un sottoinsieme limitato di variabili altamente informative.

Il principale driver del modello risulta essere il taxon s_sp004167605, appartenente al genere *Akkermansia*, che assume un ruolo centrale nel processo di discriminazione tra le classi. È interessante notare come tale feature occupasse una posizione meno rilevante nell'analisi SHAP, evidenziando come la sua importanza sia legata più alla necessità predittiva globale del modello che al contributo marginale locale.

Si osserva inoltre una presenza ricorrente di taxa appartenenti alle famiglie *Lachnospiraceae* e *Oscillospiraceae*, suggerendo un contributo consistente e distribuito di tali gruppi microbici.

Nonostante le differenze, vi è una sostanziale coerenza con i risultati ottenuti tramite SHAP: alcune feature risultano rilevanti in entrambe le analisi. Tra queste si segnalano, ad esempio, il taxon s_MGYG000004124 (ordine *Bacteroidales*) e diversi esponenti dell'ordine *Clostridia*, come s_sp900546625. La conferma trasversale di tali taxa ne rafforza il potenziale ruolo come biomarcatori.

Alcune delle feature maggiormente rilevanti secondo SHAP, però, non compaiono tra le prime posizioni nella Permutation Feature Importance. Questa discrepanza può essere attribuita alla presenza di correlazioni tra variabili: feature con elevato contributo marginale alle predizioni possono risultare ridondanti, poiché il modello è in grado di compensarne la perturbazione tramite altre variabili informative. Di conseguenza, la loro permutazione non determina un degrado significativo della performance complessiva.

Infine, la progressiva diminuzione dell'importanza nelle posizioni inferiori indica come molte feature abbiano un impatto marginale sulla performance del modello, pur contribuendo complessivamente alla sua robustezza.

3.2.2 Analisi dei risultati: Extra Trees

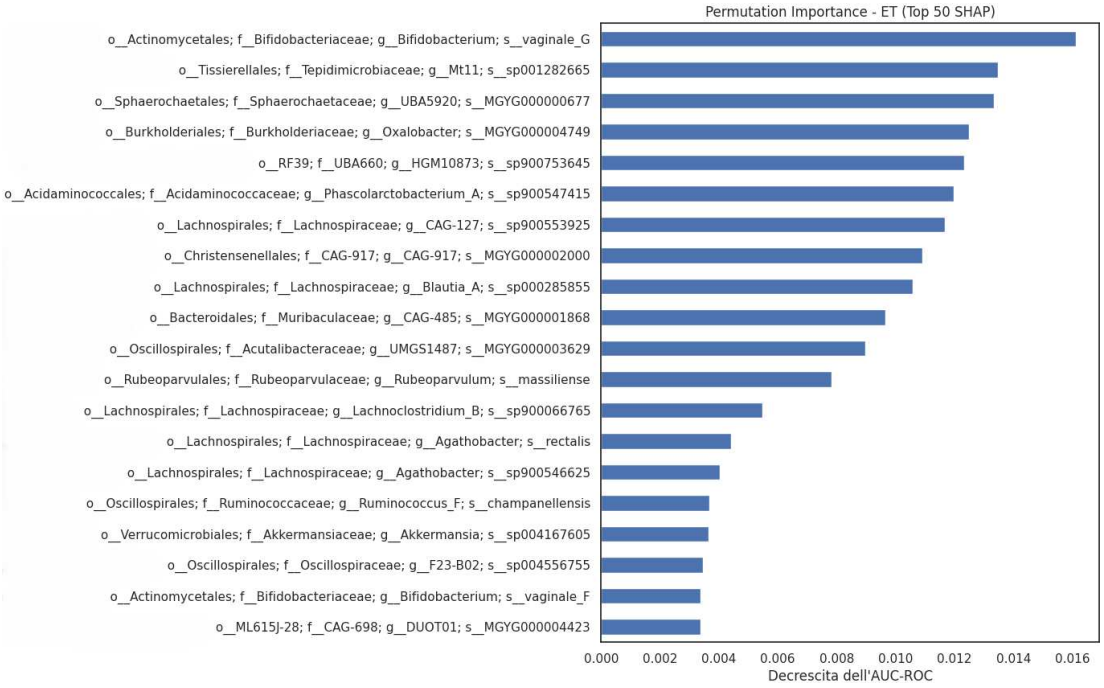

Figura 3.8: Permutation Feature Importance per il modello Extra Trees.

Dall'analisi della Figura 3.8 emerge una maggiore coerenza nel ranking delle feature rispetto all'analisi SHAP (3.3), rispetto al modello precedente.

Tra i batteri al vertice della classifica, il taxon s_vaginale_G (genere *Bifidobacterium*) risulta assolutamente indispensabile: la sua permutazione provoca una decrescita dell'AUC-ROC superiore allo 0.016, confermando il ruolo centrale nella discriminazione del modello. Anche il taxon s_sp001282665 (famiglia *Tepidimicrobiaceae*) mantiene stabilmente le prime posizioni anche nella Permutation Importance, sottolineando la sua doppia funzione di driver predittivo e variabile cruciale per la robustezza del modello.

Inoltre, si osserva la presenza in alta classifica di taxa come s_MGYG000000677 e s_MGYG000004749, che non dominavano le classifiche SHAP. Questo indica che, pur avendo un impatto marginale più contenuto sulle singole predizioni, queste specie apportano informazioni uniche che non possono essere compensate da altri taxa correlati nella foresta di alberi casuali.

Infine, la graduale decrescita dei valori e la loro entità relativamente contenuta suggeriscono la presenza di correlazioni tra le feature e confermano un segnale distribuito, tipico di sistemi in cui più variabili contribuiscono collettivamente alla performance del modello.

3.2.3 Analisi dei risultati: XGBoost

L'analisi della Permutation Feature Importance si conclude con l'applicazione al modello XGBoost. Essendo XGBoost il classificatore con il valore di AUC-ROC più elevato (0.631 ± 0.032), l'importanza attribuita ai singoli taxa da questo

modello rappresenta un'indicazione particolarmente utile su quali microrganismi siano rilevanti per la diagnosi della risposta terapeutica.

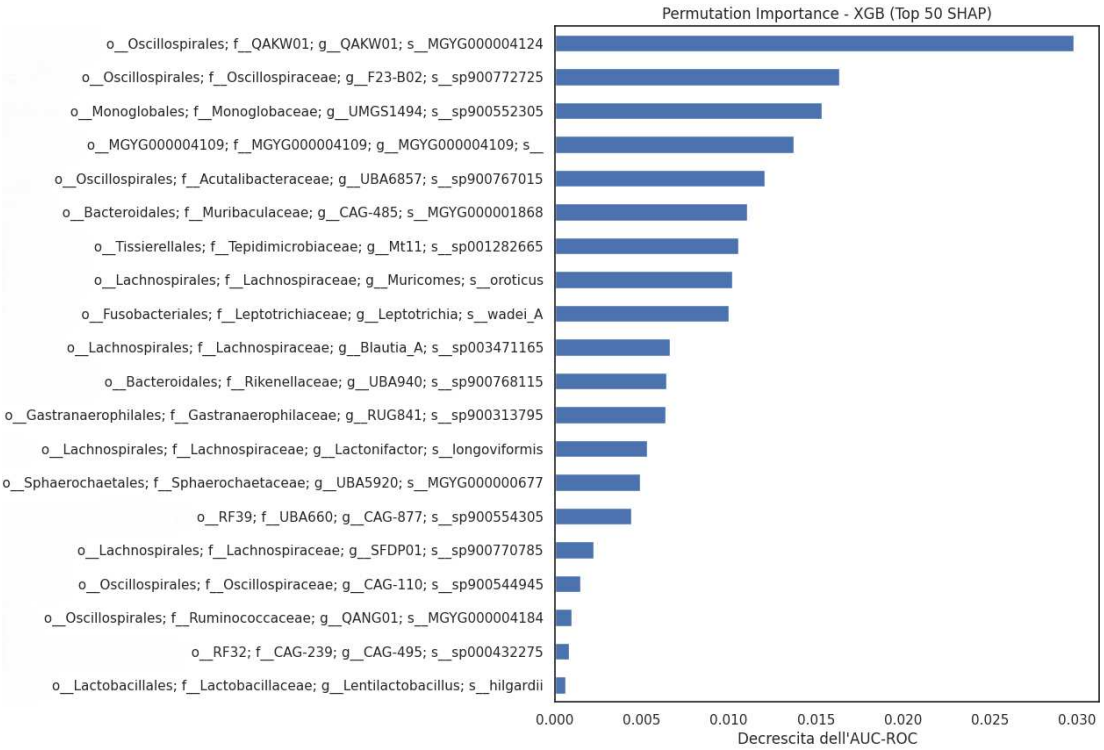

Figura 3.9: Permutation Feature Importance per il modello XGBoost.

Anche in questo caso si osserva una maggiore coerenza tra le feature individuate dal metodo SHAP (3.5) e quelle riportate dalla Permutation Importance (Figura 3.9).

Il taxon s__MGYG000004124 (ordine *Bacteroidales*) emerge come la feature più critica. Infatti, la sua permutazione provoca una riduzione dell'AUC-ROC prossima a 0.030, significativamente superiore rispetto agli altri modelli. Questo indica che, nella struttura di boosting sequenziale di XGBoost, tale specie rappresenta un'informazione centrale per la predizione, confermata anche dalla top 5 di SHAP.

Un aspetto interessante riguarda il taxon s__sp001282665 (ordine *Clostridia*), che pur risultando il driver con il maggiore impatto sulle singole predizioni nello SHAP Bar Plot, nella Permutation Importance occupa una posizione più defilata, con un impatto comunque significativo (0.010). Ciò suggerisce che, per XGBoost, il segnale derivante dai Clostridia sia in parte ridondante: il modello ha appreso pattern simili da altri taxa correlati, riducendo l'effetto della permutazione del singolo taxon sulla performance complessiva.

Inoltre, il modello XGBoost evidenzia l'importanza di taxa come s__sp900772725 (ordine *Oscillospirales*) e s__sp900770785 (ordine *Lachnospiraceae*). La loro posizione elevata nella Permutation Importance suggerisce che forniscano informazioni complementari, contribuendo a distinguere i casi più complessi di Responder e Non-Responder.

Infine, l'intervallo di valori di importanza (asse x) in XGBoost è più ampio e differenziato rispetto a Random Forest ed Extra Trees. Questo riflette come il modello attribuisca pesi più marcati ad alcune variabili chiave, rendendo più evidente la gerarchia tra taxa predominanti e feature secondarie.

3.2.4 Sintesi Comparativa della Permutation Feature Importance

L'integrazione dei risultati derivanti dalla Permutation Importance per i tre modelli analizzati permette di delineare un quadro robusto sulla dipendenza strutturale dei classificatori rispetto ai biomarcatori microbici. Mentre l'analisi SHAP si focalizzava sul contributo alla singola predizione, la Permutation Importance evidenzia l'indispensabilità delle feature per la tenuta delle performance globali.

L'analisi comparativa della PFI tra i tre modelli mostra come la dipendenza strutturale degli algoritmi vari significativamente, nonostante SHAP indicasse driver comuni.

1. Dall'analisi del grafico 3.9 emerge che XGBoost possiede la gerarchia più definita. Il taxon s_MGYG000004124 (ordine *Bacteroidales*) non solo occupa il primo posto, ma la sua permutazione provoca una riduzione dell'AUC-ROC quasi doppia rispetto ai leader degli altri modelli.

2. Dal grafico 3.7, relativo a Random Forest, il taxon al vertice è s_sp004167605 (genere *Akkermansia*), con una decrescita di circa 0.021. Pur non essendo tra i top assoluti nello SHAP, rappresenta la variabile più difficile da sostituire per il Random Forest. Al secondo posto troviamo i *Bacteroidales*, s_MGYG000004124, confermando la sua importanza trasversale.

3. Per Extra Trees (grafico 3.8) la classifica cambia leggermente: il primo posto è occupato da s_vaginale_G (genere *Bifidobacterium*). Tuttavia, Extra Trees mantiene in posizione elevata il taxon s_sp001282665 (ordine *Clostridia*), secondo nella Permutation, suggerendo che questa architettura rifletta più fedelmente i driver identificati inizialmente tramite SHAP.

Il confronto tra i tre modelli evidenzia differenze strutturali: il leader di SHAP, s_sp001282665 (*Clostridia*), perde importanza nella PFI di XGBoost e Random Forest, ma resta alto in Extra Trees. Ciò indica che i primi due modelli hanno appreso pattern sostitutivi da taxa correlati, mentre Extra Trees rimane più dipendente dalla singola feature. Inoltre, taxa come *Akkermansia* e *Bifidobacterium*, pur non emergendo tra i top SHAP, risultano fondamentali per la stabilità dei modelli.

Questi risultati sottolineano la diversa dipendenza dei modelli dai singoli taxa e pongono le basi per ulteriori verifiche tramite Feature Ablation.

3.3 Ablation Feature Importance

L'ultima fase dell'analisi è rappresentata dalla Feature Ablation Importance, il test più radicale per la valutazione dell'importanza delle variabili. Essa consiste

nella rimozione fisica e definitiva di uno o più taxa dal dataset, seguita dal completo ri-addestramento dei modelli (Capitolo 1, Sezione 1.5).

L'Ablation permette di osservare come l'architettura decisionale dell'algoritmo si riorganizzi in assenza di determinati biomarcatori. Se le performance del modello (AUC-ROC) subiscono un calo drastico a seguito della rimozione di un taxon, ciò costituisce una prova empirica della sua rilevanza strutturale e della mancanza di alternative informative all'interno del microbiota intestinale.

L'analisi di Ablation viene condotta con una duplice finalità: da una parte, verificare se la rimozione dei taxa identificati come leader da SHAP e Permutation FI comporti effettivamente la perdita del potere predittivo; dall'altra, valutare se il modello sia in grado di mantenere performance stabili appoggiandosi a taxa secondari, rivelando così la resilienza dell'ecosistema microbico nel contesto della risposta terapeutica.

Il test è applicato in modo comparativo su Random Forest, Extra Trees e XGBoost, considerando per ciascun modello le 20 top feature identificate da SHAP. Per ogni taxon viene riportato e commentato un grafico che mostra l'impatto della sua rimozione sulle performance del modello (la descizione della metodologia è specificata nel Capitolo 2, Sezione 2.4.4).

La variazione di performance è stata calcolata come differenza tra l'AUC del modello completo e quella ottenuta dopo la rimozione della feature. Valori positivi indicano una riduzione delle prestazioni in assenza della feature, evidenziandone un contributo positivo. Al contrario, valori negativi suggeriscono che la feature possa introdurre rumore o ridondanza, portando a un leggero miglioramento delle prestazioni.

3.3.1 Analisi dei risultati: Random Forest

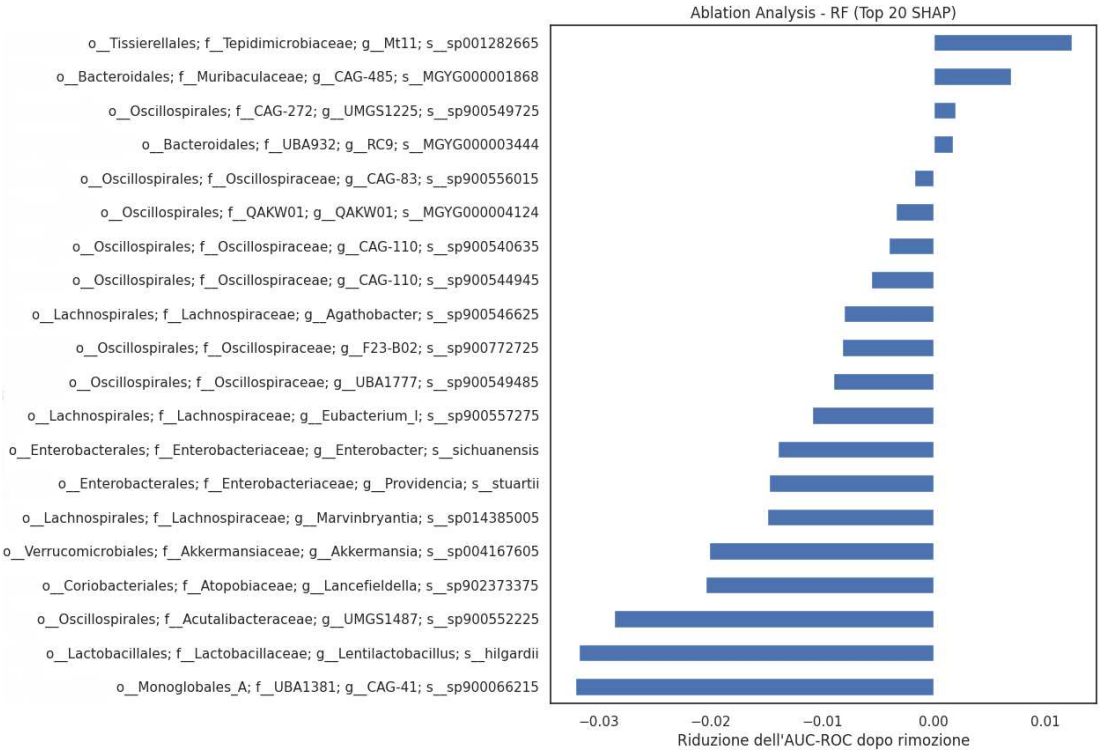

Figura 3.10: Ablation Feature Importance per il modello Random Forest.

Dalla Figura 3.10 si osserva subito come solo quattro feature risultino significative per il modello. In particolare, le prime due feature apportano una differenza di performance di circa 0.01. Tra queste sono presenti le due principali individuate da SHAP (s_sp001282665 e s_MGYG000003444).

Al contrario, la maggior parte delle feature mostra valori di ΔAUC prossimi a zero o negativi, suggerendo un contributo limitato o ridondante. In questi casi, la loro esclusione porta a un lieve miglioramento delle prestazioni, suggerendo la presenza di ridondanza informativa o rumore. Tra queste si trovano anche batteri che risultavano rilevanti nella Permutation Analysis. In particolare, un impatto negativo marcato è legato a una feature appartenente alla famiglia *Akkermansia*, che nella Permutation aveva mostrato un effetto significativo.

Infine, le feature con un impatto basso nella Permutation mantengono in questa analisi un effetto leggermente negativo o comunque contenuto.

Nel complesso, l'analisi evidenzia che, sebbene le feature selezionate tramite SHAP siano globalmente rilevanti, il loro contributo effettivo alla performance del modello non è uniforme. Ciò sottolinea l'importanza di combinare metodi di interpretabilità basati su importanza statistica con approcci di tipo causale, come la Feature Ablation Analysis, per ottenere una valutazione più robusta del ruolo delle singole variabili.

3.3.2 Analisi dei risultati: Extra Trees

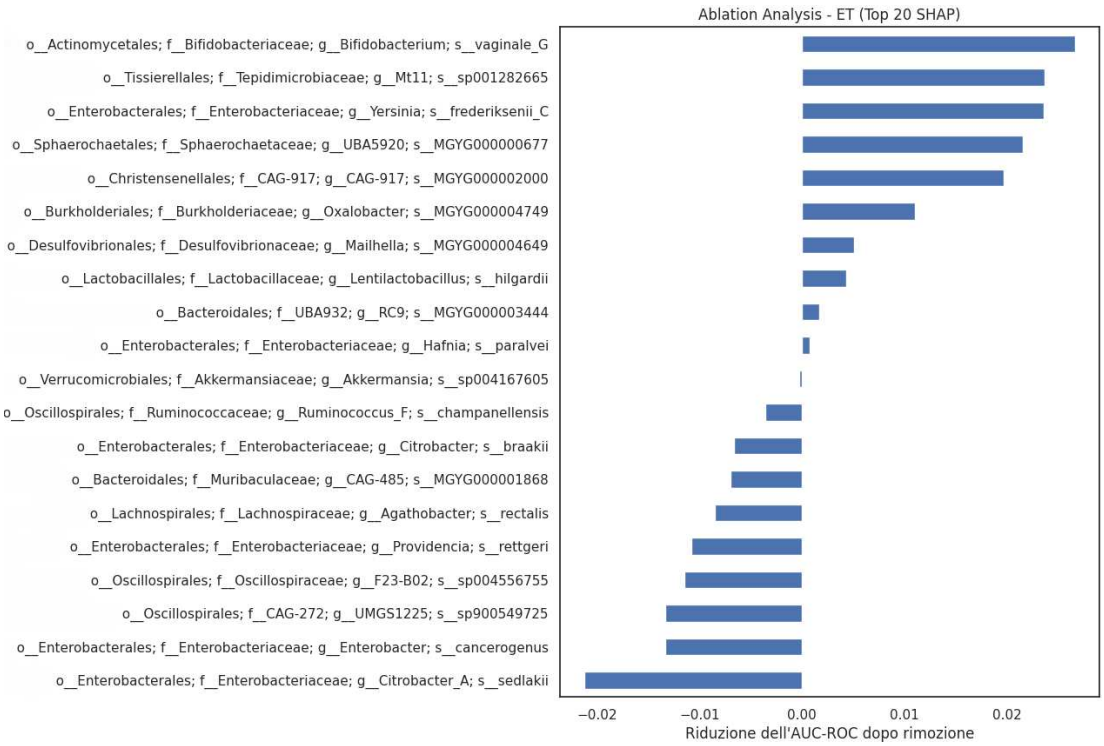

Figura 3.11: Ablation Feature Importance per il modello Extra Trees.

Dalla Figura 3.11 si osserva una separazione abbastanza bilanciata tra le feature con ΔAUC positivo, indicativo di un calo della performance, e quelle con ΔAUC negativo.

L'analisi conferma che, per il modello Extra Trees, le feature risultate rilevanti secondo SHAP e successivamente validate tramite Permutation Importance mantengono la loro importanza anche nell'Ablation. Questo suggerisce che tali variabili rappresentino effettivamente i driver principali della risposta per questo modello.

In particolare, il taxon s__vaginale_G (genere *Bifidobacterium*) si conferma al primo posto, con una riduzione della performance di circa 0.02. Anche il taxon s__sp001282665 (famiglia *Tepidimicrobiaceae*) si mantiene in seconda posizione, con un impatto leggermente inferiore ma comunque significativo.

Inoltre, si osserva come taxa che non comparivano tra le feature rilevanti nella Permutation confermino la loro scarsa utilità anche nell'Ablation, come il taxon s__sedlakii (genere *Citrobacter_A*).

Infine, anche per questo modello si osservano variazioni di performance complessivamente contenute, seppur più marcate rispetto al Random Forest, con valori massimi, sia positivi che negativi, intorno a 0.02.

3.3.3 Analisi dei risultati: XGBoost

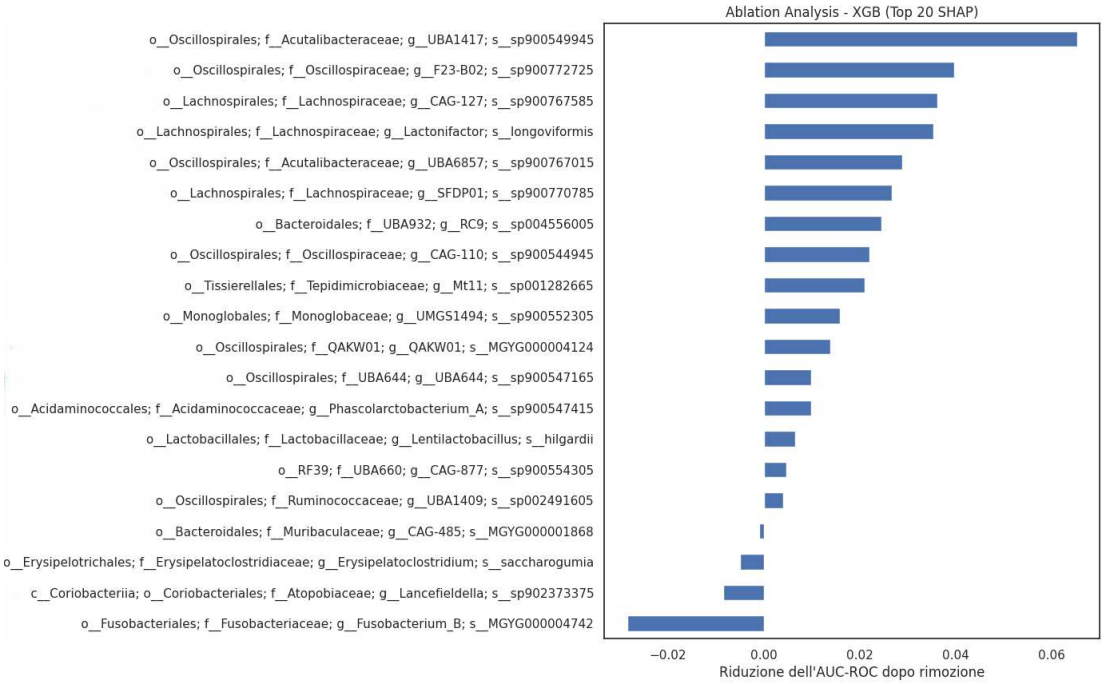

Figura 3.12: Ablation Feature Importance per il modello XGBoost.

XGBoost è il modello che ha mostrato le migliori performance e capacità nei metodi precedenti, presentando sia valori SHAP sia cali di performance (Permutation) nettamente più elevati rispetto agli altri due modelli. Anche nell'Ablation si osservano riduzioni della performance che raggiungono valori intorno a 0.06.

Dalla Figura 3.12 si osserva come la maggior parte delle feature comporti un calo della performance, mentre solo quattro su venti risultano poco informative per il modello. Questo rappresenta un segnale positivo, indicando che SHAP, in combinazione con XGBoost, seleziona feature che hanno un impatto significativo anche a livello globale sulle performance.

La feature che determina il maggiore calo di performance è il taxon s_sp900549945, appartenente alla famiglia *Acutalibacteraceae*, che nello SHAP occupava posizioni più basse. Questo suggerisce una possibile importanza globale della feature non completamente catturata dall'analisi locale.

Inoltre, feature che occupavano posizioni elevate nello SHAP e valori intermedi nella Permutation risultano rilevanti anche nell'Ablation, indicando che la loro rimozione comporta un calo delle performance. È il caso, ad esempio, del taxon s_longoviformis (famiglia *Lachnospiraceae*) e del taxon s_sp900770785 (sempre *Lachnospiraceae*), che mostrano un impatto significativo anche in questa analisi.

Il taxon s_sp001282665 (famiglia *Tepidimicrobiaceae*), che occupava la prima posizione nello SHAP, mantiene un'importanza rilevante anche nell'Ablation, confermando il suo ruolo centrale.

Infine, le feature che mostrano un miglioramento della performance (valori di ΔAUC negativi) non comparivano tra quelle rilevanti nella Permutation,

confermando il loro contributo marginale o ridondante. Un caso interessante è rappresentato dal taxon s_MGYG000001868 (famiglia *Muribaculaceae*), che presenta un ΔAUC leggermente negativo, mentre nella Permutation mostrava un valore intorno a 0.01, suggerendo una possibile ridondanza informativa.

3.3.4 Sintesi Comparativa dell'Ablation Feature Importance

L'analisi comparativa della Feature Ablation Importance evidenzia differenze significative nella dipendenza dei modelli dalle singole feature, offrendo una prospettiva complementare rispetto a SHAP e Permutation Importance.

Dai risultati emerge come Random Forest presenti una maggiore robustezza alla rimozione delle variabili: solo un numero limitato di feature provoca una riduzione significativa delle performance, mentre la maggior parte mostra un impatto nullo o addirittura leggermente negativo. Questo comportamento suggerisce una forte ridondanza informativa e una distribuzione più uniforme del segnale tra i taxa.

Il modello Extra Trees si colloca in una posizione intermedia: alcune feature risultano chiaramente indispensabili, in particolare quelle già identificate come rilevanti da SHAP e Permutation Importance, mentre altre mostrano un contributo marginale. Questo indica una maggiore coerenza con i driver individuati nelle analisi precedenti, pur mantenendo una certa capacità di compensazione tra variabili.

Diversamente, XGBoost evidenzia una dipendenza più marcata da un sottoinsieme ristretto di feature. La rimozione di alcune variabili chiave comporta infatti cali di performance significativamente più elevati rispetto agli altri modelli, confermando una struttura predittiva più gerarchica e meno ridondante. In questo caso, le feature selezionate da SHAP risultano, nella maggior parte dei casi, effettivamente indispensabili anche a livello globale.

Nel complesso, l'analisi di Ablation permette di distinguere chiaramente tra modelli caratterizzati da un segnale distribuito, come Random Forest, e modelli più selettivi e dipendenti da feature specifiche, come XGBoost. Extra Trees rappresenta invece un compromesso tra questi due comportamenti.

Questi risultati confermano l'importanza di integrare approcci diversi di interpretabilità: mentre SHAP identifica i driver locali delle predizioni e la Permutation Importance ne valuta il contributo globale, la Feature Ablation fornisce una verifica empirica della reale indispensabilità delle variabili, completando il quadro interpretativo.

3.4 Individuazione delle Variabili Principali

Per aumentare la robustezza dei risultati, è stata calcolata anche l'intersezione tra le 50 feature più rilevanti identificate dai tre modelli secondo il metodo SHAP. I taxa appartenenti a questa intersezione rappresentano i candidati biomarcatori più stabili, in quanto risultano sistematicamente rilevanti indipendentemente dall'algoritmo utilizzato.

L'intersezione ha permesso di ottenere i seguenti 8 biomarcatori.

Precisiamo che il seguente elenco non vuole indicare una classifica di importanza:

1. p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_Muribaculaceae; g_CAG-485; s_MGYG000001868;
2. p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_UBA932; g_RC9; s_MGYG000003444;
3. p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Lactobacillaceae; g_Lentilactobacillus; s_hilgardii;
4. p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_CAG-272; g_UMGS1225; s_sp900549725;
5. p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_CAG-110; s_sp900544945;
6. p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_UBA1777; s_MGYG000002084;
7. p_Firmicutes_A; c_Clostridia; o_Tissierellales; f_Tepidimicrobiaceae; g_Mt11; s_sp001282665;
8. p_Firmicutes_C; c_Negativicutes; o_Acidaminococcales; f_Acidaminococcaceae; g_Phascolarctobacterium_A; s_sp900547415;

Questo processo di intersezione ha permesso di individuare le feature che hanno mostrato consistenza in tutti e tre i modelli. Partendo da un totale di 4630 feature, per ciascun modello sono state selezionate le 50 più rilevanti secondo il metodo SHAP. L'intersezione tra questi insiemi ha portato a 8 feature comuni, rappresentando così i biomarcatori più stabili.

Dal punto di vista statistico, ridurre da 4630 a 8 feature che risultano rilevanti in maniera consistente tra modelli diversi indica una robustezza significativa, suggerendo che questi taxa non sono selezionati per caso né dipendono dal modello.

Dal punto di vista biologico, la maggior parte dei biomarcatori identificati appartiene al phylum *Firmicutes*, distribuiti tra diverse classi e ordini, mentre due appartengono a *Bacteroidota*.

Per valutare come questi batteri siano distribuiti tra pazienti Responder e Non-Responder, abbiamo realizzato un boxplot delle loro abbondanze relative. Ogni subplot mostra le abbondanze di un taxon su scala logaritmica, evidenziando differenze tra valori piccoli e grandi. I boxplot rappresentano mediana, quartili e variabilità dei dati, mentre i punti neri indicano i singoli campioni. Questa visualizzazione permette di osservare eventuali differenze tra i due gruppi e di identificare taxa potenzialmente discriminanti.

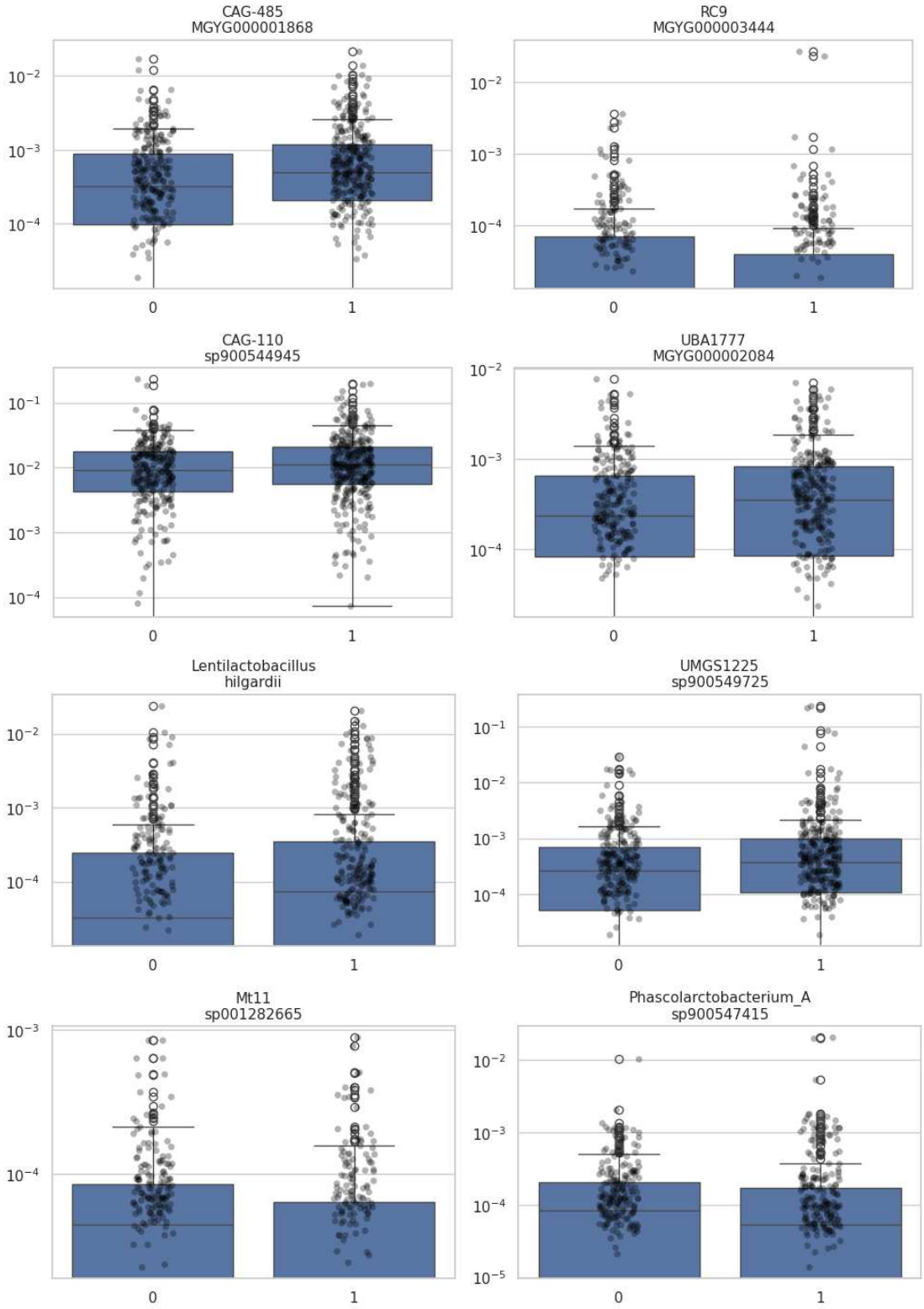

Figura 3.13: Distribuzione delle abbondanze relative dei taxa selezionati nei gruppi Responder e Non-Responder.

Dalla Figura 3.13 non emerge una distinzione netta tra le due classi, a causa della significativa sovrapposizione tra le distribuzioni. Tuttavia, la maggior parte

dei taxa selezionati mostra una tendenza a presentare abbondanze più elevate nei pazienti appartenenti alla classe 1 (Non-Responder).

Nello specifico, i taxa appartenenti all'ordine *Bacteroidales* risultano abbondanti nei Non-Responder, in accordo con la letteratura, dove la loro presenza è spesso associata a una minore risposta al trattamento [27]. Al contrario, il genere g_Mt11 mostra una maggiore abbondanza nei Responder. Questo taxon appartiene alla famiglia *Tepidimicrobiaceae*, per la quale le informazioni biologiche sono ancora limitate, rendendo complessa un'interpretazione funzionale dettagliata.

Al fine di completare l'analisi, è stata condotta un'Ablation Feature Importance su ciascun modello, rimuovendo iterativamente ognuna delle otto feature per valutarne il contributo alla performance predittiva.

Il grafico seguente mostra il confronto tra i diversi modelli in relazione all'importanza delle singole feature.

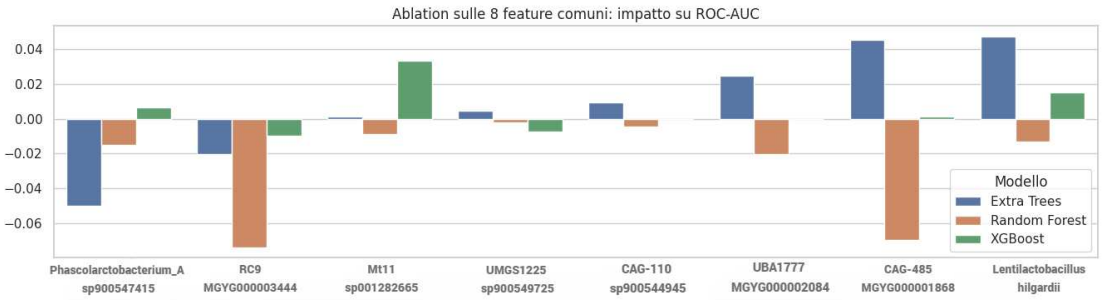

Figura 3.14: Ablation Feature Importance per i modelli considerati, espressa come variazione della ROC-AUC.

L'analisi di ablation, riportata in Figura 3.14, permette di valutare il contributo marginale di ciascun biomarcatore alla capacità predittiva dei vari modelli. La variazione della performance è stata calcolata come:

$$\Delta AUC = AUC_{\text{baseline}} - AUC_{\text{senza feature}}$$

Quindi valori positivi indicano una diminuzione delle prestazioni a seguito della rimozione della feature, mentre valori negativi suggeriscono un contributo minore o ridondante.

Dai risultati emerge che solo un numero limitato di feature mostra un impatto positivo e relativamente stabile su più modelli. In particolare, alcuni biomarcatori risultano rilevanti per modelli come Extra Trees e, in misura minore, XGBoost, mentre il Random Forest presenta prevalentemente valori negativi di ΔAUC, indicando che molte feature non contribuiscono in modo significativo, o possono addirittura degradare le prestazioni.

Si osserva inoltre una marcata variabilità tra i modelli, infatti, feature considerate importanti da un algoritmo risultano trascurabili per altri. Questo comportamento riflette le differenze strutturali tra i modelli.

Nel complesso, l'analisi di ablation evidenzia che, nonostante tutte le feature siano state selezionate come rilevanti tramite SHAP, solo un sottoinsieme contribuisce in modo consistente al miglioramento delle performance, mentre le restanti possono introdurre ridondanza o rumore.

3.4.1 Extreme Feature Selection

L'ultima fase dell'analisi ha lo scopo di valutare la capacità predittiva dell'insieme di feature selezionate tramite SHAP.

Per ciascun modello è stato costruito un dataset ridotto contenente esclusivamente le 50 feature più rilevanti identificate per quello specifico algoritmo. Ciascun modello è stato quindi riaddestrato utilizzando tale sottoinsieme di feature, e ne sono state valutate le prestazioni in termini di accuracy e ROC-AUC.

Nella tabella seguente è riportato un confronto tra i risultati ottenuti dopo la feature selection e quelli iniziali, ottenuti addestrando i modelli sull'intero set di feature (Tabella 2.1). In particolare, vengono confrontati i valori di accuracy e ROC-AUC.

Modello	**Accuracy**	**ROC-AUC**
	Tutte le feature	
Random Forest	0.554 ± 0.034	0.585 ± 0.030
Extra Trees	0.585 ± 0.025	0.596 ± 0.034
XGBoost	0.594 ± 0.009	0.631 ± 0.032
	Top50 feature (SHAP)	
Random Forest	0.570	0.615
Extra Trees	0.579	0.594
XGBoost	0.632	0.643

Tabella 3.1: Confronto delle performance dei modelli addestrati sulle feature complete e sulle Top50 feature individuate da SHAP.

Dalla Tabella 3.1 si osserva come la riduzione del numero di feature da 4630 a 50 non comporti una perdita di performance, ma in diversi casi porti a un miglioramento. In particolare, il modello XGBoost mostra un incremento sia in termini di accuracy (da 0.594 a 0.632) sia di ROC-AUC (da 0.631 a 0.643), suggerendo che le feature selezionate catturano efficacemente l'informazione più rilevante per la predizione.

Anche Random Forest beneficia della selezione delle feature, con un aumento della ROC-AUC da 0.585 a 0.615, mentre Extra Trees mantiene prestazioni sostanzialmente stabili, indicando una minore sensibilità alla riduzione dello spazio delle feature.

Nel complesso, i risultati evidenziano come la selezione delle feature tramite SHAP consenta di ridurre significativamente la dimensionalità del problema mantenendo, e in alcuni casi migliorando, le prestazioni dei modelli. Questo suggerisce che molte delle feature originali siano ridondanti o poco informative, mentre le Top 50 identificano le componenti più rilevanti del segnale.

La riduzione della dimensionalità contribuisce inoltre a limitare il rumore e a migliorare la capacità di generalizzazione. In questo contesto, XGBoost si

conferma il modello più performante, evidenziando una maggiore capacità di catturare pattern complessi anche a partire da un insieme ridotto di feature.

Complessivamente, l'analisi ha permesso di individuare un insieme ristretto di biomarcatori stabili, selezionati tramite l'intersezione delle feature più rilevanti identificate dai diversi modelli. L'utilizzo combinato di SHAP, analisi delle distribuzioni e tecniche di ablation ha consentito di valutare non solo la rilevanza statistica dei taxa, ma anche il loro contributo effettivo alla capacità predittiva.

Nonostante l'elevata dimensionalità iniziale del dataset, è quindi possibile isolare un numero limitato di feature informative, alcune delle quali risultano coerenti con quanto presente in letteratura. Tuttavia, la sovrapposizione tra le distribuzioni e la variabilità osservata tra modelli indicano che il fenomeno analizzato è complesso e non riconducibile a pochi biomarcatori dominanti.

Conclusioni

In questo lavoro di tesi sono stati applicati modelli di Explainable Artificial Intelligence (XAI) con l'obiettivo di individuare potenziali biomarcatori batterici associati alla risposta all'immunoterapia.

Il dataset analizzato presenta un'elevata complessità, essendo caratterizzato da 567 campioni e 4630 feature. In una fase preliminare sono stati testati modelli basati su reti neurali convolutive 1D, che tuttavia non hanno mostrato performance soddisfacenti, probabilmente a causa della limitata numerosità del campione rispetto all'elevata dimensionalità dei dati. Si è quindi scelto di adottare modelli tree-based, più adatti a questo tipo di contesto, ottenendo performance complessive intorno al 60%, in linea con studi precedenti presenti in letteratura.

L'utilizzo di tecniche di Explainable AI, in particolare SHAP, ha permesso di identificare le feature più influenti per ciascun modello. Metodi complementari come Permutation Importance e Ablation hanno consentito di valutare il contributo effettivo di tali feature alle performance predittive.

L'analisi comparativa tra i modelli ha portato all'identificazione di 8 potenziali biomarcatori stabili, appartenenti principalmente al phylum *Firmicutes* e, in misura minore, a *Bacteroidota*. Alcuni di questi taxa risultano coerenti con evidenze presenti in letteratura, suggerendo un possibile ruolo biologico nella modulazione della risposta all'immunoterapia, mentre altri appartengono a gruppi ancora poco caratterizzati, evidenziando la necessità di ulteriori studi.

Nel complesso, i risultati ottenuti mostrano come la combinazione di modelli di machine learning e tecniche di interpretabilità possa rappresentare uno strumento efficace per l'identificazione di biomarcatori in contesti ad alta dimensionalità. Tuttavia, la limitata dimensione del campione rispetto al numero di feature rappresenta una criticità importante, suggerendo la necessità di dataset più ampi per migliorare la robustezza dei risultati.

Inoltre, la presenza di numerosi taxa non ancora completamente caratterizzati dal punto di vista biologico rende complessa l'interpretazione funzionale dei risultati.

Studi futuri potrebbero integrare ulteriori tecniche di Explainable AI, come LIME, e considerare modelli più avanzati, oltre a includere informazioni cliniche aggiuntive per una comprensione più completa del fenomeno.

Infine, è importante sottolineare come la composizione del microbioma intestinale vari significativamente tra individui e popolazioni, limitando la generalizzabilità dei risultati. Il ruolo del microbioma nella risposta agli inibitori dei checkpoint immunitari appare quindi altamente complesso e dipendente da molteplici fattori. Studi futuri dovranno considerare campioni di dimensioni maggiori e modellare in

modo più approfondito le interazioni tra microbioma e variabili cliniche nel corso del trattamento.

Appendice A

Descrizione dei Metadati

Dato	Descrizione
samples	Identificativo univoco del paziente
country	Paese di provenienza
continent	Continente di provenienza
sex	Sesso del paziente (male o female)
age	Età del paziente (espressa come intervallo es. 50-60)
atb	Assunzione di antibiotici (NA, no, yes)
cancer	Tipologia di Cancro
therapy	Tipologia di terapia somministrata
SRA	Codice identificativo Sequence Read Archive
Platform	Piattaforma di sequenziamento utilizzata
avgSpotLen	Lunghezza media sequenze
Bases	Numero totale di basi sequenziate
BioProject	Identificativo dello studio di provenienza del campione
LibraryLayout	Layout della libreria (SINGLE/PAIRED)
ORR	Tipologia di risposta (es. completa/parziale)
response	Risposta clinica alla terapia (R/NR)
medoids	Cluster principale (1 o 2)
submedoids	Sottocluster del cluster principale

Tabella A.1: Elenco completo e descrizione dei Metadati presenti nel Dataset, non considerati nella sperimentazione.

Appendice B

Tassonomia completa delle Top 20 feature per modello secondo SHAP

In questa appendice viene invece fornita la tassonomia completa (dal kingdom alla species) per garantire una descrizione dettagliata e la piena riproducibilità dei risultati. I valori tra parentesi rappresentano l'importanza media SHAP.

Tassonomia delle Top20 SHAP per Random Forest con relativo valore di importanza.

1. Bacteria; Firmicutes_A; Clostridia; Tissierellales; f_Tepidimicrobiaceae; g_Mt11; s_sp001282665 (0.002139)
2. k_Bacteria; p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_UBA932; g_RC9; s_MGYG000003444 (0.001929)
3. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Acutalibacteraceae; g_UMGS1487; s_sp900552225 (0.001433)
4. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_CAG-83; s_sp900556015 (0.001431)
5. k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Lactobacillaceae; g_Lentilactobacillus; s_hilgardii (0.001365)
6. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_F23-B02; s_sp900772725 (0.001236)
7. k_Bacteria; p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_Muribaculaceae; g_CAG-485; s_MGYG000001868 (0.001212)
8. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_CAG-110; s_sp900544945 (0.001205)
9. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Enterobacter; s_sichuanensis (0.001177)

10. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_CAG-272; g_UMGS1225; s_sp900549725 (0.001158)

11. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_QAKW01; g_QAKW01; s_MGYG000004124 (0.001121)

12. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Lachnospirales; f_Lachnospiraceae; g_Marvinbryantia; s_sp014385005 (0.001095)

13. k_Bacteria; p_Actinobacteriota; c_Coriobacteriia; o_Coriobacteriales; f_Atopobiaceae; g_Lancefieldella; s_sp902373375 (0.001080)

14. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Providencia; s_stuartii (0.001052)

15. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_CAG-110; s_sp900540635 (0.001034)

16. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Lachnospirales; f_Lachnospiraceae; g_Agathobacter; s_sp900546625 (0.001021)

17. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_UBA1777; s_sp900549485 (0.001007)

18. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Lachnospirales; f_Lachnospiraceae; g_Eubacterium_I; s_sp900557275 (0.001006)

19. k_Bacteria; p_Verrucomicrobiota; c_Verrucomicrobiae; o_Verrucomicrobiales; f_Akkermansiaceae; g_Akkermansia; s_sp004167605 (0.000947)

20. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Monoglobales_A; f_UBA1381; g_CAG-41; s_sp900066215 (0.000943)

Tassonomia delle Top20 SHAP per Extra Trees con relativo valore di importanza.

1. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Tissierellales; f_Tepidimicrobiaceae; g_Mt11; s_sp001282665 (0.001963)

2. k_Bacteria; p_Actinobacteriota; c_Actinomycetia; o_Actinomycetales; f_Bifidobacteriaceae; g_Bifidobacterium; s_vaginale_G (0.001456)

3. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Providencia; s_rettgeri (0.001240)

4. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Hafnia; s_paralvei (0.001159)

5. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Yersinia; s_frederiksenii_C (0.001136)

6. k_Bacteria; p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_UBA932; g_RC9; s_MGYG000003444 (0.001119)

7. k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Lactobacillaceae; g_Lentilactobacillus; s_hilgardii (0.001113)

8. k_Bacteria; p_Desulfobacterota; c_Desulfovibrionia; o_Desulfovibrionales; f_Desulfovibrionaceae; g_Mailhella; s_MGYG000004649 (0.001057)

9. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Ruminococcaceae; g_Ruminococcus_F; s_champanellensis (0.001016)

10. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Citrobacter; s_braakii (0.001001)

11. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Enterobacter; s_cancerogenus (0.001001)

12. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Enterobacterales; f_Enterobacteriaceae; g_Citrobacter_A; s_sedlakii (0.000967)

13. k_Bacteria; p_Firmicutes_A; c_Clostridia_A; o_Christensenellales; f_CAG-917; g_CAG-917; s_MGYG000002000 (0.000959)

14. k_Bacteria; p_Verrucomicrobiota; c_Verrucomicrobiae; o_Verrucomicrobiales; f_Akkermansiaceae; g_Akkermansia; s_sp004167605 (0.000927)

15. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Lachnospirales; f_Lachnospiraceae; g_Agathobacter; s_rectalis (0.000917)

16. k_Bacteria; p_Spirochaetota; c_Spirochaetia; o_Sphaerochaetales; f_Sphaerochaetaceae; g_UBA5920; s_MGYG000000677 (0.000913)

17. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_CAG-272; g_UMGS1225; s_sp900549725 (0.000906)

18. k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Burkholderiales; f_Burkholderiaceae; g_Oxalobacter; s_MGYG000004749 (0.000906)

19. k_Bacteria; p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_Muribaculaceae; g_CAG-485; s_MGYG000001868 (0.000894)

20. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_F23-B02; s_sp004556755 (0.000889)

Tassonomia delle Top20 SHAP per XGBoost con relativo valore di importanza.

1. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Tissierellales; f_Tepidimicrobiaceae; g_Mt11; s_sp001282665 (0.202391)

2. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Lachnospirales; f_Lachnospiraceae; g_SFDP01; s_sp900770785 (0.178990)

3. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_CAG-110; s_sp900544945 (0.163710)

4. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Lachnospirales; f_Lachnospiraceae; g_Lactonifactor; s_longoviformis (0.161622)

5. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_QAKW01; g_QAKW01; s_MGYG000004124 (0.143055)

6. k_Bacteria; p_Firmicutes; c_Bacilli; o_Erysipelotrichales; f_Erysipelatoclostridiaceae;g_Erysipelatoclostridium; s_saccharogumia (0.115135)

7. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Ruminococcaceae; g_UBA1409; s_sp002491605 (0.107562)

8. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Oscillospiraceae; g_F23-B02; s_sp900772725 (0.099084)

9. k_Bacteria; p_Actinobacteriota; c_Coriobacteriia; o_Coriobacteriales; f_Atopobiaceae; g_Lancefieldella; s_sp902373375 (0.097657)

10. k_Bacteria; p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_Muribaculaceae; g_CAG-485; s_MGYG000001868 (0.091024)

11. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Lachnospirales; f_Lachnospiraceae; g_CAG-127; s_sp900767585 (0.083133)

12. k_Bacteria; p_Firmicutes_C; c_Negativicutes; o_Acidaminococcales; f_Acidaminococcaceae; g_Phascolarctobacterium_A; s_sp900547415 (0.083033)

13. k_Bacteria; p_Bacteroidota; c_Bacteroidia; o_Bacteroidales; f_UBA932; g_RC9; s_sp004556005 (0.082809)

14. k_Bacteria; p_Firmicutes; c_Bacilli; o_RF39; f_UBA660; g_CAG-877; s_sp900554305 (0.081665)

15. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Acutalibacteraceae; g_UBA6857; s_sp900767015 (0.078931)

16. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Monoglobales; f_Monoglobaceae; g_UMGS1494; s_sp900552305 (0.074991)

17. k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Lactobacillaceae; g_Lentilactobacillus; s_hilgardii (0.074717)

18. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_Acutalibacteraceae; g_UBA1417; s_sp900549945 (0.074460)

19. k_Bacteria; p_Fusobacteriota; c_Fusobacteriia; o_Fusobacteriales; f_Fusobacteriaceae; g_Fusobacterium_B; s_MGYG000004742 (0.073829)

20. k_Bacteria; p_Firmicutes_A; c_Clostridia; o_Oscillospirales; f_UBA644; g_UBA644; s_sp900547165 (0.072968)

Bibliografia

[1] Aas, K., Jullum, M. and Løland, A., *Explaining individual predictions when features are dependent: More accurate approximations to Shapley values*, Artificial Intelligence, vol. 298, 103502, 2021.

[2] Aitchison, J., *The Statistical Analysis of Compositional Data*, London: Chapman and Hall, 1986.

[3] Ali, S., Abuhmed, T., El-Sappagh, S. et al., *Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence*, Information Fusion, vol. 99, pp. 1-52, Elsevier, 2023.

[4] Almeida, A., Nayfach, N., Boland, M. et al., *A unified catalog of 204,938 reference genomes from the human gut microbiome*, Nature biotechnology, vol. 39, no.1, 2021.

[5] Baldelli, V., Scaldaferri, F., Putignani, L. and Del Chierico, F., *The Role of Enterobacteriaceae in Gut Microbiota Dysbiosis in Inflammatory Bowel Diseases*, Microorganisms., PubMed Central, vol.9, no. 697, 2021.

[6] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J. et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion, vol. 58, pp. 82-115, Elsevier, 2020.

[7] BNova, *XAI o eXplainable AI: cos'è e come funziona*, 2025. URL: `https://www.bnova.it/intelligenza-artificiale/explainable-artificial-intelligence/`.

[8] Breiman, L., *Random Forests*, Machine Learning, vol. 45, pp. 5-32, 2001.

[9] Caldarini, D., *Machine learning e applicazioni per l'industria. Cosa è il machine learning e come può essere utilizzato per migliorare l'industria?*, Tech Blog SMC. URL: `https://techblog.smc.it/it/2020-05-25/machine-learning-industry`

[10] Candeliere, F.,Musmeci, E., Amaretti, A. et al., *Profiling of the intestinal community of Clostridia: taxonomy and evolutionary analysis*, Microbiome Research Reports, vol. 2, no. 13, 2023.

[11] Chen, M., *Cos'è il Machine Learning?*, Oracle Italia, 2024. URL: https://www.oracle.com/it/artificial-intelligence/machine-learning/what-is-machine-learning/

[12] Chen, T. and Guestrin, G., *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, arXiv:1603.02754, 2016.

[13] *Cross-validation: evaluating estimator performance*, scikit-learn documentation. URL:https://scikit-learn.org/stable/modules/cross_validation.html.

[14] Davar, D.,Dzutsev, A.K., McCulloch, J.A. et al., *Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients*, Science (New York, N.Y.), vol. 371, 2021.

[15] Derosa, L., Routy, B., Thomas, A.M, *Intestinal Akkermansia muciniphila predicts clinical response to PD-1 blockade in patients with advanced non-small-cell lung cancer*, Nature medicine, vol. 28, 2022.

[16] Fatima, S.S., Wooldridge, M. and Jennings, N.R., *A Linear Approximation Method for the Shapley Value*, Artificial Intelligence, vol. 172, pp. 1673-1699, Elsevier, 2008.

[17] Frankel, A.E., Coughlin, L.A., Kim, J., et al., *Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients*, Neoplasia (New York, N.Y.), vol. 19, 2017.

[18] Galli, S., *Understanding Permutation Feature Importance for Model Interpretation*, Train In Data, 2024. URL: https://www.blog.trainindata.com/permutation-feature-importance/.

[19] Geurts, P., Ernst, D. and Wehenkel, L., *Extremely randomized trees*, Mach Learn, Springer Science + Business Media, vol. 63, pp. 3-42, 2006.

[20] Gopalakrishnan, V., Spencer, C.N, Nezi, L. et al., *Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients*, Science (New York, N.Y.) vol. 359, 2018.

[21] Greenacre, M., Martínez-Álvaro, M. and Blasco, A., *Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation*, Frontiers in Microbiology, vol.12, 2021.

[22] Hameed, I., Sharpe, S., Barcklow, D. et al., *BASED-XAI: Breaking Ablation Studies Down for Explainable Artificial Intelligence*, 2022, arXiv:2207.05566.

[23] Hickman, B. and Korpela, K., *Impact of data compositionality on the detection of microbiota responses*, Gut Microbes, vol. 17, no. 1, 2025.

[24] Hidalgo-Cantabrana, C., Delgado, S., Ruiz, L. et al., *Bifidobacteria and Their Health-Promoting Effects*, Microbiol Spectr., PubMed Central, 2017.

[25] IBM, *Che cos'è l'AI spiegabile?*. URL: `https://www.ibm.com/it-it/think/topics/explainable-ai`.

[26] Lee, K. A., Thomas, A. M., Bolte, L. A. et al., *Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma*, Nature Medicine, 2022.

[27] Limeta, A., Ji, B., Levin, M. et al., *Meta-analysis of the gut microbiota in predicting response to cancer immunotherapy in metastatic melanoma*, JCI Insight, vol. 5, no. 23, 2020.

[28] Liu, B., Chau, J., Dai, Q., *Exploring Gut Microbiome in Predicting the Efficacy of Immunotherapy in Non-Small Cell Lung Cancer*, Cancers vol. 14, no. 21, 2022.

[29] Longo, L., Brci, M., Cabitza, F. et al., *Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions*, arXiv:2310.19775, 2023.

[30] Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L., *Bracken: estimating species abundance in metagenomics data*, PeerJ. Computer science, vol. 3, 2017.

[31] Lu, Y., Yuan, X., Wang, M., *Gut microbiota influence immunotherapy responses: mechanisms and therapeutic strategies*, Journal of hematology & oncology, vol. 15, 2022.

[32] Lundberg, S.M., Erion, G., Chen, H. et al., *From local explanations to global understanding with explainable AI for trees*, Nature Machine Intelligence, 2020.

[33] Lundberg, S.M. and Lee, S.-I., *A Unified Approach to Interpreting Model Predictions*, 31st Conference on Neural Information Processing Systems (NIPS), 2017, arXiv:1705.07874.

[34] Lundberg, S.M., Erion, G.G. and Lee, S.-I., *Consistent Individualized Feature Attribution for Tree Ensembles*, 2019, arXiv:1802.03888.

[35] Matson, V., Fessler, J., Bao, R. et al., *The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients*, Science (New York, N.Y.), vol. 359, 2018.

[36] McCulloch, J.A., Davar, D., Rodrigues, R.R. et al., *Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1*, Nat Med, PubMed Central, vol. 28, no. 3, 2022.

[37] Merrick, L., *Randomized Ablation Feature Importance*, 2019, arXiv:1910.00174.

[38] Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2022. URL: `https://christophm.github.io/interpretable-ml-book/`

[39] Molnar, C., König, G., Bischl, B. and Casalicchio, G., *Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach*, arXiv:2006.04628.

[40] Novielli, P., Romano, D., Magarelli, M. et al., *Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification*, Frontiers in Microbiology, 2024.

[41] Olaniran, O. R., Olaniran, S. F., Allohibi, J. et al., *Mixed effect gradient boosting for high-dimensional longitudinal data*, Scientific Reports, vol. 15, 2025.

[42] Ottman, N., Geerlings, S.Y., Aalvink, S. et al., *Action and function of Akkermansia muciniphila in microbiome ecology, health and disease*, Best Practice & Research Clinical Gastroenterology, vol. 31, no. 6, pp. 637-642, 2017.

[43] Pan, A. Y., *Statistical analysis of microbiome data: The challenge of sparsity*, Current Opinion in Endocrine and Metabolic Research, vol. 19, pp. 35-40, 2021.

[44] *Permutation Feature Importance*, scikit-learn documentation. URL: `https://scikit-learn.org/stable/modules/permutation_importance.html`.

[45] Peters, B.A., Wilson, M., Moran, U. et al., *Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients*, Genome medicine, vol. 11, 2019.

[46] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S. et al., *Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development*, Clinical and Translational Science, vol. 17, no. 11, 2024.

[47] Qi, X., Liu, Y., Hussein, S. et al., *The Species of Gut Bacteria Associated with Antitumor Immunity in Cancer Therapy*, Cells, vol. 11, 2022.

[48] Qin, L., Zhu, Y., Liu, S. et al., *The Shapley Value in Data Science: Advances in Computation, Extensions, and Applications*, Mathematics, 2023.

[49] Routy, B., Le Chatelier, E., Derosa, L. et al., *Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors*, Science (New York, N.Y.), vol. 359, 2018.

[50] Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I. et al., *A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME*, Advanced Intelligent Systems, vol. 7, no. 1, 2025.

[51] Sarker, I. H., *Machine Learning: Algorithms, Real-World Applications and Research Directions*, SN Computer Science, Springer Nature, pp. 2:160, 2021.

[52] Sundararajan, M. and Najmi, A., *The Many Shapley Values for Model Explanation*, arXiv:1908.08474, 2020.

[53] Tan, H., Zhai, Q. and Chen, W., *Investigations of Bacteroides spp. towards next-generation probiotics*, Food Research International, vol. 116, pp. 637-644, 2019.

[54] Wenink, E., *DeepLIFT – AI Ethics Tool Landscape*, 2021. URL: `https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/deeplift/`

[55] Wood, D.E., Lu, J., and Langmead, B., *Improved metagenomic analysis with Kraken 2*, Genome biology, vol. 20, no.1, 2019.

[56] Yang, J., *Fast TreeSHAP: Accelerating SHAP Value Computation for Trees*, arXiv:2109.09847, 2021.