UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

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

RELATORE

Prof. Fabio Palomba

Dott. Antonio Della Porta

Dott.ssa Viviana Pentangelo

Università degli Studi di Salerno

CANDIDATO

Rosa Carotenuto

Matricola: 0512113246

Anno Accademico 2024-2025

Questa tesi è stata realizzata nel

Abstract

Il diabete di tipo 1 (T1D) è una malattia autoimmune complessa, caratterizzata da una risposta immunitaria contro le cellule β pancreatiche, che compromette la produzione di insulina. La scarsità di dati relativi alle fasce pediatriche e ai giovani adulti rappresenta una sfida nell'analisi dell'espressione genica per questa malattia. In tale contesto, il machine learning offre strumenti potenti per analizzare dati complessi, ma la mancanza di interpretabilità dei modelli limita la loro applicazione clinica. Questa tesi si propone di applicare tecniche di machine learning spiegabile per migliorare la comprensione e l'interpretabilità dei modelli predittivi utilizzati nella classificazione dei pazienti sani e affetti da T1D. Sono stati esplorati modelli come SVM, Random Forest e XGBoost, con l'SVM che ha mostrato le migliori performance in termini di accuratezza e metriche generali. L'impiego di tecniche di explainability come SHAP e Anchor ha permesso una migliore interpretazione delle decisioni del modello. I risultati evidenziano come l'approccio spiegabile permetta di comprendere il comportamento del modello e i fattori che influenzano le sue predizioni, suggerendo che un approccio spiegabile potrebbe contribuire a una maggiore fiducia e affidabilità nel potenziale utilizzo clinico del machine learning nel contesto del T1D.

Indice

Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Introduzione

1.1 Contesto Applicativo

Il diabete di tipo 1 (T1D) è una malattia autoimmune caratterizzata dalla distruzione delle cellule β del pancreas che sono responsabili della produzione di insulina, un ormone essenziale per il controllo dei livelli di glucosio nel sangue. L'assenza di insulina comporta gravi squilibri glicemici e richiede una gestione rigorosa tramite somministrazione esterna di insulina, monitoraggio continuo della glicemia e aggiustamenti frequenti dell'apporto alimentare. Nonostante le terapie avanzate, il T1D comporta ancora oggi rischi significativi di complicanze a lungo termine, tra cui retinopatia, nefropatia e neuropatia. In parallelo, la ricerca scientifica ha cominciato a esplorare i fattori genetici e molecolari che influenzano la predisposizione e la progressione del T1D. L'analisi dell'espressione genica offre una finestra preziosa su questi meccanismi, consentendo ai ricercatori di mappare i cambiamenti a livello di trascrizione che caratterizzano le fasi iniziali della malattia. Le nuove tecnologie, utilizzate per misurare l'espressione di migliaia di geni contemporaneamente, hanno aperto nuove opportunità per lo studio dei biomarcatori genetici, permettendo di identificare geni associati a specifici pattern di espressione legati al T1D. Tuttavia, la raccolta di dati di espressione genica per il T1D presenta diverse sfide: i dataset

pubblici sono spesso limitati, soprattutto per le popolazioni pediatriche e giovani adulti, età in cui si ha la maggiore incidenza.

In questo contesto, il machine learning offre potenti strumenti per analizzare e interpretare dati complessi e ad alta dimensionalità, come quelli di espressione genica. L'uso di algoritmi di machine learning per identificare biomarcatori predittivi e costruire modelli di classificazione può facilitare la diagnosi precoce del T1D, oltre a contribuire alla comprensione dei fattori genetici che ne influenzano l'insorgenza e la progressione. Tuttavia, molti dei modelli attuali presentano limitazioni in termini di interpretabilità e affidabilità, soprattutto nel contesto clinico. La necessità di comprendere come e perché un modello prenda determinate decisioni è essenziale per ottenere la fiducia dei medici e dei ricercatori, e garantire che i risultati siano utilizzabili in contesti pratici.

1.2 Motivazioni e Obiettivi

Il lavoro svolto si propone di affrontare queste sfide tramite l'applicazione di modelli di machine learning sviluppando modelli predittivi capaci di distinguere tra individui sani e affetti dalla malattia. Nello specifico, l'obiettivo principale è utilizzare tecniche avanzate di preprocessamento e selezione delle feature per migliorare la qualità del dataset di espressione genica e garantire modelli robusti e interpretabili. La tesi si concentra su:

- Costruzione di modelli predittivi: Sviluppare e ottimizzare modelli di machine learning, tra cui SVM, Random Forest e XGBoost, per classificare correttamente i campioni e analizzarne le prestazioni.
- Selezione delle feature: Identificare i geni più rilevanti legati al T1D nel contesto della ricerca e analizzarne l'impatto attraverso tecniche di explainability come SHAP e Anchor, al fine di ottenere modelli interpretabili anche per utenti non esperti.

Questi obiettivi sono motivati dalla necessità di migliorare l'affidabilità e la precisione dei modelli predittivi nel campo della biologia computazionale, offrendo nuovi strumenti per la diagnosi precoce e il monitoraggio del T1D.

1.3 Risultati Ottenuti

I risultati del lavoro di tesi evidenziano l'efficacia dei modelli sviluppati nel distinguere tra individui sani e affetti da T1D, con l'SVM che ha ottenuto le migliori performance in termini di accuracy e AUC-ROC. Inoltre, l'uso delle tecniche di explainability ha permesso di identificare e spiegare le associazioni genetiche chiave associate alle predizioni dei modelli, fornendo non solo un alto livello di interpretabilità e regole comprensibili per utenti non specialisti. Le analisi hanno confermato che specifici geni, come HLA-DQB1 e PTPRN2, mostrano variazioni significative tra le classi, suggerendo il loro potenziale come discriminanti nelle predizioni.

1.4 Struttura della Tesi

La tesi è organizzata come segue:

- **Capitolo 2: Background e stato dell'arte**
  Fornisce una revisione della letteratura riguardante il diabete di tipo 1, l'uso di dati di espressione genica nella diagnosi delle malattie e le tecniche di machine learning applicate nel dominio della ricerca.

- **Capitolo 3: Costruzione del dataset**
  Descrive dettagliatamente la metodologia utilizzata per la raccolta, il preprocessamento e l'analisi dei dati.

- **Capitolo 4: Modelli di predizione e confronto**
  Presenta la costruzione dei modelli predittivi e l'analisi dei risultati ottenuti tramite l'utilizzo di metriche specifiche.

- **Capitolo 5: Explainability e discussione**
  Mostra l'applicazione delle principali tecniche di explainability e i confronti tra i risultati prodotti.

- **Capiloto 6: Conclusioni**
  Vengono discussi i risultati alla luce delle conoscenze attuali, concludendo con le implicazioni della ricerca e le potenziali direzioni future.

CAPITOLO 2

Background e Stato dell'Arte

Il T1D è legato a un malfunzionamento del sistema immunitario che attacca erroneamente le cellule beta. Questa autoimmunità risulta legata a fattori genetici e fattori ambientali. Studi di sequenziamento del genoma (WGS), studi di associazione Genome Wide (GWAS) e l'analisi di varianti genomiche ha permesso di identificare i loci genetici associati alla malattia. Tra i principali c'è il complesso maggiore di istocompatibilità (HLA) che rappresenta il fattore di rischio genetico più forte per il T1D. Tuttavia, numerosi altri geni contribuiscono alla suscettibilità e alla progressione della malattia. In questo capitolo analizzeremo in dettaglio i processi patogenetici che portano allo sviluppo del T1D esplorando i principali protagonisti della sua insorgenza. Effettueremo una panoramica sul pancreas, dove sono situate le cellule β, il sistema immunitario, i differenti tipi di immunità, le cellule che ne partecipano e le sostanze secrete in modo da comprendere al meglio il malfunzionamento delle cellule β e i meccanismi di compesazione che vengono attuati. Approfondiremo la predisposizione genetica, trattando alcuni dei loci associati alla malattia e il conseguente cambiamento di espressione genica che ne risulta. Infine, illustreremo i lavori correlati, ovvero gli studi recenti che hanno applicato tecniche avanzate di machine learning e genomica per migliorare la comprensione dei fattori genetici e ambientali alla base del T1D. Esamineremo le metodologie utilizzate, i risultati ottenuti e le

loro implicazioni, per costruire un quadro completo delle attuali conoscenze sulla patogenesi del T1D e sulle possibili applicazioni cliniche di questi approcci.

2.1 Patogenesi

Il T1D, come già detto precedentemente, è una malattia autoimmune organo specifica caratterizzata da distruzione delle cellule β . Ci sono varie prove a sostegno delle basi autoimmuni della malattia. Esse includono [1]:

1. Insulite, cioè la presenza di infiltrazione linfocitaria intorno e all'interno delle isole di Langherans, una sezione specifica del pancreas
2. La comparsa di autoanticorpi contro più autoantigeni delle isole insulari
3. La presenza di geni di suscettibilità alla malattia legati al complesso maggiore di istocompatibilità (MHC) e non legati all'MHC
4. L'aumentata propensione a sviluppare malattie autoimmuni più organo-specifiche

La patogenesi del T1D viene divisa in tre stadi. Il primo stadio è l'insulite, che come abbiamo già detto in precedenza, consiste nell'autoimmunità verso le cellule del pancreas indotta dai linfociti T citotossici, componenti cellulari del sistema immunitario. Si ha, quindi, la comparsa di autoanticorpi. Attulmente è noto che gli autoanticorpi anti-isole non causano direttamente la distruzione delle cellule pancreatiche. Essi insorgono come risultato della distruzione delle cellule β mediata dalle cellule T [1]. I cinque principali autoanticorpi correlati al T1D sono:

- Anticorpi citoplasmatici anti-insula pancreatica (ICA)
- Anticorpi anti decarbossilasi dell'acido glutammico (GADA)
- Anticorpi-2 associati all'insulinoma (IA-2A)
- Anticorpi anti-insulina (IAA)
- Anticorpi anti-trasportatore dello zinco 8 (ZnT8A)

Questi anticorpi sono presenti nel 95-98% delle persone affette da T1D. In seguito, il secondo stadio è caratterizzato da disglicemia o intolleranza al glucosio. Entrambi i primi due stadi risulato essere asintomatici. Infine, l'ultimo stadio ha sintomi di iperglicemia come poliuria[1], polidipsia[2], enuresi [3], perdita di peso, visione offuscata, alcune volte con chetoacidosi diabetica[4] o sindrome ipersmolare diabetica[5] [2]. Inoltre, uno studio che ha coinvolto parenti di primo grado dei pazienti con T1D ha riportato che IAA o GADA sono comparsi per primi, seguiti da IA-2A e ZnT8A [3].

2.2 Pancreas

Il pancres un organo pieno a struttura lobulare costituito per il 97% da componente esocrina e dal 3% da componente endocrina. È situato in diagonale dietro allo stomaco e presenta una forma allungata. La componente endocrina, che è quella che nel caso specifico ci interessa, è formata da quelle che sono le isole di Langherans. Esse sono composte da alcune centinaia di cellule endocrine e capillari. Le principali cellule endocrine degli isolotti sono:

- Cellule α che producono glucagone, adibito all'aumento del livello del glucosio nel sangue ed altri ormoni, come la colecistochinina o l'endorfina
- Cellule β che producono insulina, adibita alla diminuzione del livello di glucosio nel sangue
- Cellule δ che producono somatostatina

Nel caso del T1D, l'insorgenza è dovuta proprio alla distruzione delle cellule β delle isole da parte dei linfociti T autoreattivi. Al momento della diagnosi circa il

[1]La **poliuria** è una condizione in cui vengono prodotte abbondanti quantità di urine associata a un aumento della frequenza della minzione.

[2]La **polidipsia** è una condizione caratterizzata da un'eccessiva sensazione di sete.

[3]La **enuresi** è una condizione che porta alla perdita del controllo della vescica e il rilascio involontario di urina.

[4]La **chetoacidosi** diabetica è una complicanza acuta del TD1 che porta a sintomi come nausea, vomito, dolore addominale e un caratteristico odore fruttato dell'alito.

[5]La **sindrome ipersmolare diabetica** è una complicazione del diabete mellito caratterizzata da sintomi come l'iperglicemia grave, la disidratazione estrema e alterato stato di coscienza.

70-80 % del patrimonio beta cellulare dell'organismo risulta distrutto, sebbene questa percentuale possa essere molto variabile [4].

2.3 Sistema immunatario

Illustreremo qui le principali caratteristiche del sistema immunitario utili nella comprensione dei meccansmi alla base della disfunzione delle cellule β pancreatiche e dei meccanismi di compensazione attuati.

Immunità. L'immunità può essere distinta in immunità innata e immunità adattiva. L'immunità innata è chiamata così perché è presente dalla nascita e non deve essere acquisita mediante l'esposizione ai patogeni. Fornisce una risposta immediata ai microrganismi invasori. I componenti dell'immunità innata riconoscono, però, solo un numero limitato di invasori rispetto all'immunità adattiva. Non dispone di memoria immunoligica come l'immunità adattiva. Inoltre, l'immunità innata induce quello che è il fenomeno dell'infiammazione. L'immunità innata e l'immunità adattiva si distinguono anche per il tipo di cellule, il tipo di sostanze prodotte e il tipo di meccanismi effettori. L'immunità adattiva, invece, non è presente alla nascita, ma inizia a svilupparsi dopo che il sistema immunitario viene a contatto con invasori esterni e riconosce sostanze esogene. È dotata di memoria immunologica, ma non è dotata della stessa rapidità di azione dell'immunità innata ed entra in gioco solo dopo che i meccanismi dell'immunità innata hanno fallito. Inoltre l'immunità adattiva può essere divisa in immunità umorale e immunità cellulo-mediata.

Cellule dell'immunità innata. Le cellule dell'immunità innata sono distinte in due gruppi: fagociti e cellule citotossiche. I fagociti sono cellule capaci di adoperare la fagocitosi: internalizzano e digeriscono il patogeno distruggendolo. Le cellule citotossiche, invece, sono capaci di uccidere le cellule infettate da microrganismi. Le cellule dell'immunità innata sono:

- Macrofagi: Derivano dai monociti, insieme ai neutrofili e alle cellule dendritiche fanno parte della prima linea di difesa dell'immunità innata. Sono capaci di

attività fagocitica. Sono in grado di effettuare la presentazione dell'antigene ai linfociti T.

- Cellule dendritiche: Hanno come pricipale obiettivo quello di captare particelle nel fluido extracellulare, come batteri o virus. Anch'esse sono in grado di effettuare la presentazione dell'antigene ai linfociti T e, proprio per le loro capacità di riconoscimento di particelle estranee, sono tra le più efficaci.

- Granuliciti: Si dividono in neutrofili, basofili ed eosinofili. I neutrofili sono i principali effettori dell'immunità innata. Operano tramite fagocitosi e attività battericide attraverso le sostanze presenti nei loro granuli. I basofili hanno una moderata attività fagocitica e gli eosinofili sono coinvolti nell'eliminazione di alcuni microrganismi.

- Mastociti: Sono anch'esse cellule ricche di granuli, utili nella risposta antiparassitaria.

- Cellule Natural Killer (NK): Possiedono una grande quantità di granuli citotossici. La loro principale funzione è quella di esercitare la citotossicità verso cellule infettate da virus. Esprimono recettori attivatori o inibitori e dall'equilibrio di questi recettori dipende la loro l'attivazione. I recettori inibitori riconoscono strutture self mentre i recettori attivatori riconoscono sia le cellule infettate sia cellule che non esprimono molecole che le identificano come self.

Il riconoscimento di un patogeno da parte dei macrofagi e della cellule dendritiche attiva i meccanismi di fagocitosi e i meccanismi pro-infiammatori. Una volta riconosciuto il patogeno, queste cellule sintetizzano e secernano una serie di citochine.

Cellule dell'immunità adattiva. Le cellule dell'immunità adattiva sono essenzialmente i linfociti B e i linfociti T. I linfociti sono piccole cellule facilmente distinguibili a livello del sangue per avere un grosso nucleo che occupa quasi tutta la cellula e un citosol ristretto. Il linfocita "resting" o "naive" è un linfocita che non ha ancora incontrato l'antigene. Nel momento in cui il linfocita incontra l'antigene va incontro ad una fase di espansione clonale, cioè inizia a proliferare in maniera sotenuta. Abbiamo due tipi di linfociti: linfociti T, che si specializzeranno in linflociti T citotossici

e linfociti T helper, e i linfociti B, che si differenziano in plasmacellule adibite alla produzione di iummunoglobuline, detti anche anticorpi. I linfociti T, a differenza dei linfociti B, non sono in grado di riconoscere l'antigene direttamente, ma lo fanno solo nel momento in cui avviene la presentazione dell'antigene tramite altre cellule. I recettori delle cellule B e T vanno incontro ad un processo di riarrangiamento: il DNA presente nella linea germinale non risulta essere uguale al DNA presente a livello somatico nei linfociti. Proprio grazie a questi riarrangiamenti, queste cellule sono in grado di riconoscere un enorme numero di antigeni. C'è quindi bisogno di eliminare quei linfociti che sono detti autoreattivi, cioè riconoscono le strutture self. Questo meccanismo viene detto selezione clonale negativa o delezione clonale. Nel caso del T1D i linfociti autoreattivi sfuggono a questo processo.

Immunità umorale. L'immunità umorale dipende dalle immunoglobuline e quindi è operata dai linfociti B. Le cellule B vengono attivate in due modi:

- Riconoscimento dell'antigene da parte del loro recettore di membrana, non chè un'immunoglobulina.
- Da una cellula TH2 tramite contatto cellula cellula e il rilascio di alcune citochine.

Dopo una prima fase di proliferazione, il linfocita B si differenzia in plasmacellula, una cellula adibita ad una vasta produzione di immunoglobuline.

Immunità cellulo-mediata. L'immunità cellulo mediata ha inizio quando una cellula presentante l'antigene (APC) presenta l'antigene ad un linfocita T, attivandolo. Come abbiamo già detto, i linfociti T si distinguono in linfociti T helper, ovvero i linfociti T CD4, oppure linfociti T citotossici, ovvero i linfociti T CD8. CD4 e CD8 sono dei marker espressi sulla superficie dele cellule T. Un linfocita T CD4 può andare incontro a due destini differenti:

- Trasformarsi in una cellula TH2.
- Trasformarsi in una cellula TH1.

Le cellule TH2 hanno principalmente la funzione di attivare le cellule B. Le celule TH1 migrano nei siti di infezione e riconoscono i macrofagi infettati dai patogeni. Se i macrofagi non riescono ad eliminare i patogeni, hanno bisogno delle cellule TH1, che hanno la capacità di rilasciare sostanze capaci di potenziare la fagocitosi. Se avviene l'attivazione di un linfocita T CD8, probabilmente si è verificata un'infezione intracellulare, come un'infezione virale, e quindi c'è bisogno dell'intervento di un linfocita citotossico.

Presentazione dell'antigene. Il riconoscimento dell'antigene da parte dei linfociti T avviene solo quando le APC, come le cellule dendritiche, effettuano la presentazione dell'antigene a tali cellule. La presentazione dell'antigene avviene tramite delle molecole, chiamate molecole della istocompatibilità o molecole MHC. Risultano codificate in un locus genomico detto complesso magiore della istocompatibilità e sono in grado di trasportare frammenti di antigene di natura peptidica. Le molecole MHC possono essere di due classi:

- MHC di classe I
- MHC di classe II

Le MHC di classe I sono presenti sulle membrane di tutte le cellule nucleate, invece le MHC di classe II sono espresse solo sulle cellule del sistema immunitario, in particolare sulle cellule che presentano l'antigene di tipo specializzato. Le molecole MHC di classe I sono formate da due catene polipeptidiche: una catena α (catena pesante) ed una subunità chiamata β2–microglobulina (catena leggera). Invece, le molecole MHC classe II sono formate da una catena α e una catena β.

2.3.1 Alcune componenti molecolari del sistema immunitario

Le citochine sono molecole di natura polipeptidiche secrete da alcune cellule del sistema immunitario e da altre cellule quando interagiscono con, per esempio, alcuni patogeni. Le principali categorie comprendono:

- Le chemochine
- I fattori stimolanti le colonie (CSFs)

- Interferoni
- Interleuchine
- I fattori di crescita trasformante (TGFs)
- Fattori di necrosi tumorali

Anche se l'interazione del linfocita con uno specifico antigene stimola la secrezione di citochine, le citochine di per sé non sono antigene-specifiche. Uniscono l'immunità innata a quella acquisita e generalmente influenzano l'entità delle risposte infiammatorie o immunitarie. Nell'ambito della nostra ricerca analizzeremo [5]:

Chemochine. Le chemochine inducono la chemiotassi e la migrazione di leucociti. Troviamo recettori per le chemochine sulle cellule T della memoria, su monociti/macrofagi, sulle cellule dendritiche e sulle cellule T quiescenti.

Interferoni. Gli interferoni sono una famiglia di proteine che hanno attività antivirale e agiscono come modulatori immunitari. Quelli che interessano nell'ambito della patogenesi del T1D sono:

- Interferone α (IFNα): È prodotto dai leucociti e i principali effetti sono:
  - Inibizione della replicazione virale
  - Aumento dell'espressione delle MHC di classe I
- Inteferone γ (IFNγ): È prodotto dalle cellule NK, cellule citotossiche di tipi I e cellule T-helper di tipo 1. Tra principali effetti sono:
  - Aumento delle MHC di classe I e II e dell'espressione del recettore Fc, recettore per il frammento costante delle immunoglobuline.
  - Attivazione dei macrofagi e delle cellule NK
  - Inibizione della proliferazione delle cellule T-helper di tipo 2

Interleuchine. Le interleuchine sono prodotte da un gran numero di cellule, hanno molteplici effetti sullo sviluppo cellulare e sulla regolazione delle risposte immunitarie. Nel contentesto del T1D, analizziamo:

- Interleuchina 1 (IL-1) α e β: Sono prodotte da cellule B, cellule dendritiche, macrofagi, monociti, cellule natural killer. Tra i principali effetti troviamo:
  - Costimolazione dell'attivazione delle cellule T promuovendo la produzione di citochine
  - Miglioramento della proliferazione e della maturazione delle cellule B
  - Miglioramento della citotossicità delle cellule
  - Induzione della produzione dell'IL-1, TNF e altro.
  - Attività proinfiammatoria mediante induzione di chemochine
  - Induzione dei mediatori della fase acuta dell'infiammazione

Fattori di necrosi tumorale.

- TFN α: È prodotto dalle cellule B, cellule dendritiche, macrofagi, mastociti, monociti, cellule NK, cellule TH. Tra i principali effetti troviamo:
  - Induzione della secrezione di numerose citochine che stimolano l'infiammazione
  - Attivazione dei macrofagi
  - Attività antivirale
  - Citotossicità per le cellule tumorali
- TFN β: È prodotto dalle cellule T citotossiche e dalle TH1 Tra i principali effetti troviamo:
  - Citotossicità per le cellule tumorali
  - Attività antivirale
  - Promozione della fagocitosi mediata dai neutrofili e dai macrofagi

2.4 Insorgenza del diabete di tipo 1

Problematiche delle cellule β. L'identificazione di cellule immunitarie innate e l'espressione di molecole infiammatorie nel pancreas trapiantato di soggetti con T1D supporta l'ipotesi di un ruolo chiave del sistema immunitario innato e dell'infiammazione nella ricorrenza del T1D in questi pazienti. La risposta immunitaria che scatena l'insulite, infatti, porta l'attivazione di specifici recettori, chiamati recettori del riconoscimento dei pattern (PRR), sulla membrana delle cellule β del pancreas. Ciò viene tradotto nella produzione di interferone di tipo I (come l'IFNα) da parte delle cellule β che fa partire il reclutamento delle varie cellule immunitarie. I macrofagi sono i primi soccorritori che iniziano la produzione di TFN. Ciò inesca il NF-κ B, un fattore trascrizionale con principale funzione proapoptotica[6]. Si ha, in seguito, l'aumento della permeabilità vascolare che porta l'infiltrazione di linfociti T naive e attivati. Prevalentemente sono presenti linfociti T CD8+, ma anche cellule B CD20+, linfociti T CD4+ e macrofagi CD68+. La risposta delle cellule β alle citochine, come IL1-β e IFNγ, presenti in questa fase è l'attivazione dei percorsi anti-infiammatori. Inoltre, le citochine pro-infiammatorie disturbano l'attività metabolica ed elettrica delle cellule β e la sintesi e il contenuto dei granuli di insulina. La disfunzione delle cellule β risulta presente anni prima dell'insorgenza del T1D. Le cellule β non sono solo bersagli passivi, ma partecipano attivamente e possibilmente amplificano i processi patogeni: impiegano meccanismi di compesazione in risposta allo stress immunitario, che diventano deleteri a lungo termine. Durante l'aumento della sintesi dell'insulina si ha anche un aumento di proteine mal ripiegate e prodotti ribosomiali all'interno e all'esterno delle cellule. Queste, vengono ad essere riconosciute e portano all'attivazione della via HLA-1. La funzionalità alterata delle cellule β, influenzata da segnali proinfiammatori e cambiamenti nell'espressione genica, quindi, porta alla perdita della produzione di insulina [2].

Popolazione. Nonostante il T1D possa insorgere a chiunque e a qualsiasi età ci sono delle precisazioni da fare. L'insorgenza della patologia si ha a due picchi: nei bambini tra i 4 e i 7 anni e nei bambini tra i 10 e i 14 anni. Il rischio cambia anche

[6]Con **apoptosi** si intende la morte cellulare programmata

a seconda della popolazione: abbiamo un maggior tasso di incidenza nell'Europa del Nord, in paesi come Finlandia e Svezia. Abbiamo uno dei più alti tassi di T1D in Sardegna, specificamente solo nella regione perchè nel resto dell'Italia non risulta essere così rilevante. I caucasici risultano, anch'essi, moto colpiti. Troviamo invece una ridotta incidenza tra le poplazioni dell'America latina, l'Asia e Afroamericane.

2.4.1 Predisposizione genetica

Il rischio di T1D nei fratelli dei pazienti è 15 volte superiore al rischio di T1D nella popolazione generale, il che suggerisce che i fattori genetici svolgono un ruolo importante nella suscettibilità alla malattia [6]. Nella maggior parte dei casi, i pazienti affetti da T1D ereditano le caratteristiche da entrambi i genitori. In caso di uomini con T1D, la probabilità che la progenie possa contrarre la patologia è 1 su 17. Invece, se si tratta di una donna la probabilità che la progenie, se nata prima che la donna abbia compiuto 25 anni, è di 1 su 25; se invece la progenie è nata dopo i 25 anni di età, il rischio scende a 1 su 100. I rischi risultano raddoppiati in caso l'avvento della malattia per un genitore risulta essere prima degli 11 anni. In più, se entrambi i partner hanno il TD1 allora il rischio risulta essere tra 1 su 10 e 1 su 4 [7]. La maggior parte dei rischi riguardanti l'insorgenza della malattia derivano da polimorfismi a singoli nucleotidi (SNP) dei geni. La presenza di SNP in un determinato gene può modificare la struttura, il livello di espressione o la funzionalità della proteina codificata, rendendola unica per quell'individuo.

Complesso maggiore di istocompatibilità (HLA). Le MHC sono codificate in uno specifico locus genico. Questo locus codifica per un centinaio di geni, molti dei quali hanno funzioni specializzate nella risposta immunitaria. Il locus MHC è chiamato Human Leukocyte Antigen (HLA) ed è localizzato sul cromosoma 6. I geni di classe I nell'uomo sono: A, B e C che codificano per la catena α; la β2–microglobulina è codificata su un diverso cromosoma. I geni di classe II sono: DP, DQ e DR ognuno dei quali codifica per una catena α e una β tranne per il gene DR che codifica per due catene β differenti. Le molecole MHC di classe I e II sono molecole polimorfiche[7] e

[7]Con **polimorfismo** si intende la presenza in una popolazione più genotipi o fenotipi per un determinato carattere. Con **fenotipo** si intende la manifestazione di un gene in un essere vivente.

poligeniche [8]. Le molecole MHC sono presenti nella popolazione con più varianti polimorfiche: per le MHC di classe II la maggior parte dei polimorfismi si trovano nei geni che codificano per le catene β, a livello del gene DR, e i polimorfismi per le catene α sono molto meno numerosi; invece, per le MHC di classe I hanno il maggior numero di polimorfismi nel gene B. Il polimorfismo nelle molecole MHC ha lo scopo di diversificare al massimo la capacità di legare peptidi derivanti da antigeni differenti. Invece, la poligenia è dovuta al fatto che per ogni classe esistono tre geni [8]. C'è un alto rischio di contrarre il T1D in caso siano presenti mutazioni nei geni HLA di classe II, che rappresenta il 50% del rischio di contrazione. Alune varianti in tutti e tre i loci possono influenzare il rischio dello sviluppo di un primo autoanticorpo. Quindi i principali determinanti genetici sono i polimorfismi della classe II dei geni DQ. DR e, in misura minore, DP. Alleli della classe I, HLA-B, sono associati all'insorgenza della malattia [9], in particolare quando si ha l'interazione con i geni della classe II DR e DQ [10].

Geni non-HLA. Gli studi di associazione genome-wide (GWAS) hanno aumentato le conoscenze sulle basi genetiche del T1D: sono stati identificate altre SNP che non sono localizzate nei geni HLA. Il lavoro è stato effettuato da varie collaborazioni nazionali e internazionali, come il Type 1 Diabetes Genetics Consortium (T1DGC), The Environmental Determinants of Diabetes in the Young (TEDDY), Diabetes Autoimmunity Study in the Young (DAISY), Diabetes in the Newborn Study (BABYDIAB) e il Wellcome Trust Case Control Consortium (WTCCC) [11]. Di seguito sono riportati i geni più fortemente associati. In primo luogo, il gene dell'insulina (INS) ha una forte associazione con la malattia. INS si trova sul cromosoma 11p15; i polimorfismi dell'insulina regolano la quantità di mRNA[9] dell'insulina nel timo. I polimorfismi di questo gene possono influenzare lo sviluppo della tolleranza immunitaria all'insulina [12]. Tutti i polimorfismi si trovano al di fuori delle sequenze di codificanti del DNA, suggerendo che la suscettibilità al T1D deriva dalla modulazione della trascrizione dell'INS [13]. Un altro gene rilevante è il PTPN22 che codifica per la fosfatasi

[8] Con **poligenia** si intende il fenomeno per cui più geni determinano uno stesso carattere.

[9] L'**RNA messaggero (mRNA)** permette il trasferimento dell'informazione dai geni ai ribosomi, adibiti alla sintesi delle proteine.

linfoide-specifica (LYP) sul cromosoma 1p13. LA LYP è coinvolta nella prevenzione dell'attivazione spontanea delle cellule T. Il polimorfismo di questo gene promuove la sopravvivenza dei linfociti T autoreattivi nel timo [13, 10]. Il gene CTLA4, che si trova sul cromosoma 2q33, invece, codifica per una molecola che risulta essere un regolatore negativo per l'attivazione delle cellule T citotossiche [13, 10]. Il gene IL2RA, che si trova sul cromosoma 10p15, codifica la subunità α del recettore dell'IL-2, che è espresso sui linfociti [14, 15]. Il gene IFIH1 è localizzato sul cromosoma 2q24.2. Codifica per la proteina MDA5, proteina 5 associata alla differenziazione del melanoma, un sensore citoplasmatico dell'RNA a doppio filamento nei virus. Infatti, anche le infezioni virali fanno parte della patogenesi della malattia. La proteina MDA5 attiva la cascata di risposte immunitarie antivirali [16]. Altri SNP geni importanti, che non verranno approfonditi, sono CTSH (catepsina H), GLIS3 [17], SH2B3 (proteina adattatrice), ERBB3 (recettore tirosin chinasico erbB-3) e UBASH3A (proteina A associata all'ubiquitina contenente il dominio SH3) [18] sempre legati all'insorgenza dell'autoimmunità delle cellule delle isole. Insieme a questi possiamo notare che il rischio di IAA come primo autoanticorpo è stato associato a geni come INS, SH2B3 e ERBB3 mentre GADA come primo autoanticorpi è stato associato a geni come SH2B3 [19]. Oltre questi, sono presenti anche altri geni che portano allo sviluppo degli autoanticorpi.

Espressione genica Con espressione genica intendiamo il processo tramite il quale la cellula trasforma le informazioni contenute nel proprio DNA in molecole, proteine o RNA, tramite la quale si svolgono le funzioni necessarie alle attività metaboliche e strutturali. L'espresione genica è regolata da meccanismi raffinati. Riguardo il T1D, studi di GWAS hanno idenificato oltre 60 regioni a rischio in tutto il genoma umano caratterizzate da SNP che garantiscono la predisposizione alla patologia. Vi è un'evidenza cescente che gli SNP associati alla malattia possono alterare l'espressione genica attraverso interazioni spaziali tra loci distali, cioè loci effettivamente lontali dal gene target, ma che possono essere vicini all'interno del nucleo quando il DNA si trova sottoforma di cromatina. Questi loci spesso contengono elementi regolatori come enhancer, cioè sequenze di DNA che, quando attivate, possono aumentare il livello di trascrizione di geni specifici, e silencer, regioni del DNA che sopprimono

l'espressione di geni nelle vicinanze. Inoltre, oltre i loci HLA, come abbiamo già accenato, altri studi GWAS hanno associato il locus della Catepsina H (CTHS) come un locus suscettibile all'aumento del rischio di sviluppo della malattia. Abbiamo, infatti, una sovraespressione del gene CTSH in alcune cellule del pancreas nei pazienti con T1D in comparazione a gruppi di controllo. Questa ricerca è stata effettuata tramite single cell RNA sequencing (scRNA) [17]. Abbiamo anche una ridotta espressione di una variabile solubile di CTL4, uno dei geni citati precedentemente [13]. Dufort et al. mediante RNA-seq mostra livelli più elevati di cellule B, livelli più bassi di neutrofili e livelli più bassi di peptide C [20]. L'iperespressione di HLA-I è stata descritta sia individui con T1D di recente diagnosi che in donatori di positivi agli auto anticorpi (AAb+) prima che la malattia si presentasse. È interessante notare che l'HLA-I era espresso principalmente dalle cellule α indipendentemente dallo stato della malattia. È stato anche riscontrato che le cellule risultassero funzionalmente compromesse, esprimevano segnali proinfiammatori e avevano un'espressione genica alterata. Tuttavia, le cellule β sono le più colpite da questo processo. L'aumento dell'espressione di HLA-II e dei componenti della via di elaborazione e presentazione dell'antigene HLA-II è stato osservato nelle cellule β di individui con T1D, e sembra essere unico per il T1D. Ciò è stato riscontrato nel 90% dei pazienti [21] e mostra il ruolo diretto delle cellule β nella patologia del T1D, agendo come APC professionali nella risposta autoimmune. L'esposizione delle cellule β umane alle citochine proinfiammatorie IFNα1, IL1-β e IFNγ provoca un'espressione differenziale dei geni, in particolare all'aumento dell'espressione di HLA-I, che, insieme al stress intracellulare e all'apoptosi delle cellule β, può portare ad un aumento della presentazione di neoantigeni, contribuendo così al reclutamento di cellule T autoreattive che attaccano selettivamente le cellule β [2].

2.5 Lavori correlati

In questa sezione, esamineremo i lavori correlati che hanno contribuito alla comprensione dei meccanismi genetici e immunitari del T1D, con un focus sull'applicazione di tecniche di machine learning e sull'analisi trascrittomica. In *Classification of Gene Expression Dataset for Type 1 Diabetes Using Machine Learning Methods*, Noor AlRefaai

et al. [22] analizzano l'uso di diverse tecniche di machine learning per classificare i geni associati al T1D utilizzando dataset di espressione genica. Vengono descritte cinque fasi che includono il preprocessing, il ranking, la selezione delle feature, la classificazione e la valutazione. Il preprocessing ha incluso la gestione di dati mancanti tramite la media della colonna per sostituire i valori assenti; successivamente è stata applicata la Min-Max Normalization per ridurre i valori in un intervallo tra 0 e 1 standardizzandoli. Per la selezione delle feature sono stati utiizzati Chi-squared test, ANOVA, Mutual Information e Principal Component Analysis. Tra i vari algoritmi la combinazione tra SVM e Chi-squared ha prodotto i risultati migliori, con un'accuracy dell'89.9%. La ricerca si conclude dicendo che attraverso un'attenta selezione di feature e preprocessing è possibile migliorare la classificazione dei geni associati al T1D fornendo strumenti utili per la diagnosi precoce e la gestione della malattia. In *Predicting Diabetes Mellitus With Machine Learning Techniques*, Zou et al. [23] esplorano l'uso di tre algoritmi di machine learning, Random Forest (RF), Decision Tree (J48) e Neural Network (NN) a due strati per la predizione del diabete mellito usando due dataset: un dataset di esami fisici di persone sane e diabetiche in Cina, dove RF ha raggiunto un'accuracy dell'80.8%, e il dataset Pima Indians Diabetes, dove RF ha raggiunto 76.0%. La rete neurale, invece, ha ottenuto risultati misti, con un'accuray leggermente inferiore a quella di RF, mentre J48 ha avuto una performance inferiore agli altri due. La fase di preprocessing ha incluso la rimozione di campioni con dati mancanti e l'applicazione di tecniche per la riduzione della dimensionalità come Principal Component Analysis e Minumun Redundancy Maximum Relevance. Il primo ha ridotto il numero di feature trasformandole in componenti principali, ma non ha apportato miglioramenti consistenti alle metriche dei modelli; il secondo è stato utilizzato per la selezione delle feature più rilevanti, producendo buoni risultati e migliorando la precisone rispetto all'uso di tutte le feature. È stato osservato, quindi, che l'indice glicemico è la feature più importante tra quelle scelte, ma non è sufficiente per dare la massima accuratezza. In *Predictive Supervised Machine Learning Models for Diabetes Mellitus*, L.J. Muhammad et al. [24] hanno esplorato l'uso di diversi algoritmi di machine learning per la predizione del diabete mellito di tipo 2. Il dataset utilizzato è stato reperito presso il Murtala Mohammed Specialist Hospital in Nigeria e sono stati selezionati sei algoritmi di machine learning tra cui: Random Forest, che

ha raggiunto la precisione più alta (88.76%), seguito da Gradient Boosting (86.76%) e SVM (85.29%). Il preprocessing si è basato sulla pulizia dei dati per rimuovere eventuali valori mancanti e la conversione dei dati in un formato CSV per facilitarne l'analisi. Inoltre, i dati sono stati analizzati per identificare le correlazioni tra variabili: è risultato, anche qui, che il livello di glucosio risulta essere una feature di forte impatto. La ricerca conclude dicendo che modelli come Random Forest e Gradient Boosting, che combinano più classificatori, offrono una migliore accuretazza e capacità predittiva per il diabete di tipo 2 e che questi strumenti possono essere utilizzati come strumenti di supporto nella diagnosi clinica fornendo informazioni preziose. In *Modeling Type 1 Diabetes Progression Using Machine Learning and Single-Cell Transcriptomic Measurements in Human Islets*, Abhijeet R. Patil et al. [25], esaminano l'utilizzo di modelli di machine learning per prevedere lo sviluppo del T1D analizzando i cambiamenti trascrittomici nelle isole pancreatiche. Sono stati utilizzati dati di sequenziamento RNA a singola cellula (scRNA-seq) da donatori di pancreas: il team ha sviluppato modelli di machine learning per identificare signature genetiche precoci associate al T1D. L'algoritmo che è stato utilizzato è XGBoost, scelto per la capacità di gestire dati ad alta dimensionalità. Sono stati costruiti tre classificatori binari, uno per distinguere tra cellule di donatori T1D e AAb+, cioè positivi per gli autoanticorpi, uno per T1D e controlli sani e un ultimo per AAb+ e controlli sani. XGBoost ha ottenuto performance elevate, con un'accuracy media del 99% per la classificazione T1D vs. controlli sani. Il preprocessing dei dati di scRNA-seq ha incluso la filtrazione delle cellule di bassa qualità, seguita da tecniche di riduzione della dimensionalità. Le cellule sono state annotate usando marcatori genetici specifici per identificare i diversi tipi di cellule. Successivamente i dati sono stati divisi tra dati di training e dati di testing e sono stati applicati metodi di ottimizzazione degli iperparametri per affinare il modello. Il modello ha rilevato una signature genetica comune tra diversi tipi cellulari nelle isole associate al T1D, con enfasi su geni della classe HLA I. Inoltre, i pathway più rappresentati includono la risposta immunitaria e il processamento antigenico, suggerendo che questi cambiamenti molecolari sono parte integrante della progressione della malattia. Si conclude spiegando che l'approccio con XGBoost e scRNA-seq non solo permette la classificazione accurata delle cellule T1D, ma può anche essere un mezzo di identificazione per cambiamenti molecolari precoci.

CAPITOLO 3

Costruzione del dataset

In questo capitolo viene descritto il processo di costruzione del dataset utilizzato per l'addestramento dei modelli di machine learning finalizzati alla predizione del diabete di tipo 1 (T1D). La scelta dei dati è stata guidata dall'esigenza di rappresentare campioni appartenenti alla fascia di età in cui il T1D tende a insorgere, cioè in età pediatrica e in giovani adulti, in quanto analizzare questi dati proprio al momento dell'insorgenza della malattia risulta particolarmente utile per scopi preventivi. Considerata questa specifica tipologia di campioni, la disponibilità di dati pubblici è risultata limitata; di conseguenza, è stato necessario trattare ogni singolo campione come un singolo paziente, allo scopo di ampliare il numero di osservazioni e mantenere la rappresentatività del dataset rispetto alla fascia di età studiata. I dati grezzi sono stati reperiti dalla piattaforma GEO (Gene Expression Omnibus), da cui sono stati selezionati due dataset che rispondono ai criteri sopra descritti. Il capitolo fornisce una panoramica delle principali tecniche di preprocessing adottate per migliorare la qualità dei dati. In particolare, è stata effettuata la normalizzazione e la correzione del batch effect per ridurre le variazioni non biologiche introdotte dalle due diverse piattaforme di microarray utilizzate. Ogni tecnica è spiegata in dettaglio, evidenziando come queste operazioni siano fondamentali. Al termine del preprocessing, i due dataset selezionati sono stati uniti, combinando le espressioni

geniche in base ai geni in comune. Questo processo ha permesso di creare un singolo dataset integrato, che rappresenta un quadro più ampio delle espressioni geniche associate al T1D e costituisce la base per l'addestramento dei modelli di machine learning selezionati nello studio.

3.1 Introduzione alla scelta del dataset

I dati sono stati recuperati dalla piattaforma Gene Expression Ombinus (GEO) [26]. Sono stati scelti due dataset reperibili dalla piattaforma con ID di accesso GSE9006 e GSE43488. Entrambi sono costituiti dall'espressione dei geni presenti nelle cellule mononucleari del sangue periferico (PBMC)[1]. In entrambi i casi, si tratta di dati di microarray, una tecnologia utilizzata per misurare l'espressione genica di migliaia di geni in un singolo esperimento.

GSE9006. Sono stati prelevati campioni di sangue da 43 pazienti con diagnosi di T1D, 12 con diagnosi di T2D e 24 soggetti sani come gruppo di controllo. In seguito, da 20 dei pazienti con T1D sono stati ottenuti ulteriori campioni a 1 e 4 mesi dalla diagnosi. I pazienti avevano un'età compresa tra i 2 e i 18 anni. I dati sono stati raccolti tramite Affymetrix U133A Gene Array (GPL96) e Affymetrix Human Genome U133B Array (GPL97). È stata scelta la piattaforma GPL96 in modo da ridurre a complessità, in quanto la seconda copre sonde aggiuntive, ma mantenere comunque i geni principali.

GSE43488. Sono stati prelevati campioni provenienti da 10 bambini positivi agli autoanticorpi diabetici e 10 bambini utilizzati come control; 18 bambini prediabetici e 18 bambini utilizzati come control. I control presentavano la stessa categoria di rischio HLA-DQB1, genere,luogo e data di nascita, ma risultavano negativi agli autoanticorpi per il T1D. I dati sono stati raccolti tramite Affymetrix Human Genome U219

[1]Con **PBMC** si fa riferimento a tutte le cellule del sangue periferico con un singolo nucleo rotondo. Sono formate per la maggior parte da linfociti e monociti; vengono spesso utilizzate per studi di immunologia e per indagare malattie autommuni.

Array (GPL13667) e sono stati esclusi i campioni solamente positivi agli anticorpi e mantenuti i campioni prediabetici in modo da favorire l'etereogeneità del dataset.

Per la costruzione del dataset, i livelli di espressione genica sono stati ottenuti dai file CEL disponibili sulla piattaforma GEO. I dati clinici e demografici dei pazienti sono stati estratti dai file series matrix associati, anch'essi disponibili su GEO, e collegati alle espressioni geniche per arricchire il dataset. Data la scarsità di informazioni pubbliche, ogni campione è stato trattato come indipendente e distinto: questo è stato possibile poiché erano presenti notevoli differenze tra i campioni. Non sono state introdotte variabili cliniche in quella che sarebbe stata poi la futura predizione su questi dati. Inoltre, i due dataset sono stati trattati prima separatamente per poi effettuare un merge sui geni mappati in comune.

Come per i dati di espressione, anche per l'estrazione dei dati clinici si è lavorato separatamente per poi unificarli. Si è deciso di mantenere l'età, il sesso dei pazienti e di introdurre due etichette: "Healthy", in caso in cui i soggetti fossero sani, "Type 1 Diabetes", in caso in cui i soggetti fossero prediabetici o già soggetti al T1D.

Tecnologie. Durante la gestione dei dati sono state utilizzate diverse tecnologie chiave. R, utilizzato attraverso l'interfaccia grafica di R-Studio, ha facilitato la visualizzazione dei dati. È stata utilizzata la piattaforma Bioconductor per permettere l'analisi e la modellazione dei dati, grazie anche all'utilizzo del pacchetto affy per la lettura dei dati dai file CEL. Parallelamente, è stato utilizzato Pyton con la libreria pandas per manipolazioni ulteriori.

3.2 Preprocessing.

Normalizzazione. Il dataset GSE9006 è stato sottoposto a un processo di normalizzazione, poiché il GSE43488 risultava già normalizzato con il metodo RMA (Robust Multi-array Average). RMA è una tecnica usata per migliorare la comparabilità dei dati di microarray. Attraverso la correzione del background, cioè eliminazone del rumore di fondo che può interferire con il segnale reale proveniente dalle sonde, e la logaritmizzazione dei segnali, RMA stabilizza la varianza dei dati. Infine, applica una normalizzazione quantile per uniformare i dati tra i diversi array, consentendo un'analisi più accurata e affidabile dei livelli di espressione genica tra i campioni.

Mappatura dei geni. Dopo aver effettuato la lettura dei file CEL e l'eventuale normalizzazione, ogni sonda è stata mappata al gene di riferimento utilizzando le librerie `hgu133a.db` e `hgu219.db` rispettivamente per GSE9006 e GSE43488. Durante il mapping dei geni, sono sorti due problemi: diverse sonde erano associare a geni uguali e diversi geni erano mappati dalla stessa sonda. Si è scelto quindi di numerare le sonde in ordine crescente e di usare la funzione `collapsedRows()` contenuta all'interno del pacchetto WGCNA, che riduce la ridonanza nei dati di espressione genica [27].

Riduzione del batch effect. Il batch effect è una problematica comune che si verifica quando i dati provengono da esperimenti condotti in condizioni diverse o, come nel caso di questo studio, da piattaforme di microarray differenti. Tali differenze tecniche possono introdurre variazioni che non riflettono vere differenze biologiche, ma sono piuttosto legate a variazioni sperimentali. In questo studio, poiché i campioni di espressione genica derivavano da due piattaforme diverse, era probabile che queste differenze tecniche influenzassero i risultati, mascherando o distorcendo le reali variazioni biologiche tra i campioni. Per valutare l'effettiva presenza del batch effect, è stata applicata la PCA (Principal Component Analysis). La PCA è una tecnica di riduzione della dimensionalità che consente di trasformare variabili correlate in componenti principali non correlate, mantenendo la maggior parte dell'informazione nei dati e riducendo le variabili a quelle più rilevanti. In questo caso, i dati sono stati

ridotti a due componenti principali, che sono state successivamente rappresentate graficamente. Come mostrato in Figura 3.1, i campioni si raggruppavano in due cluster distinti, suggerendo la presenza di variazioni tecniche legate al batch effect.

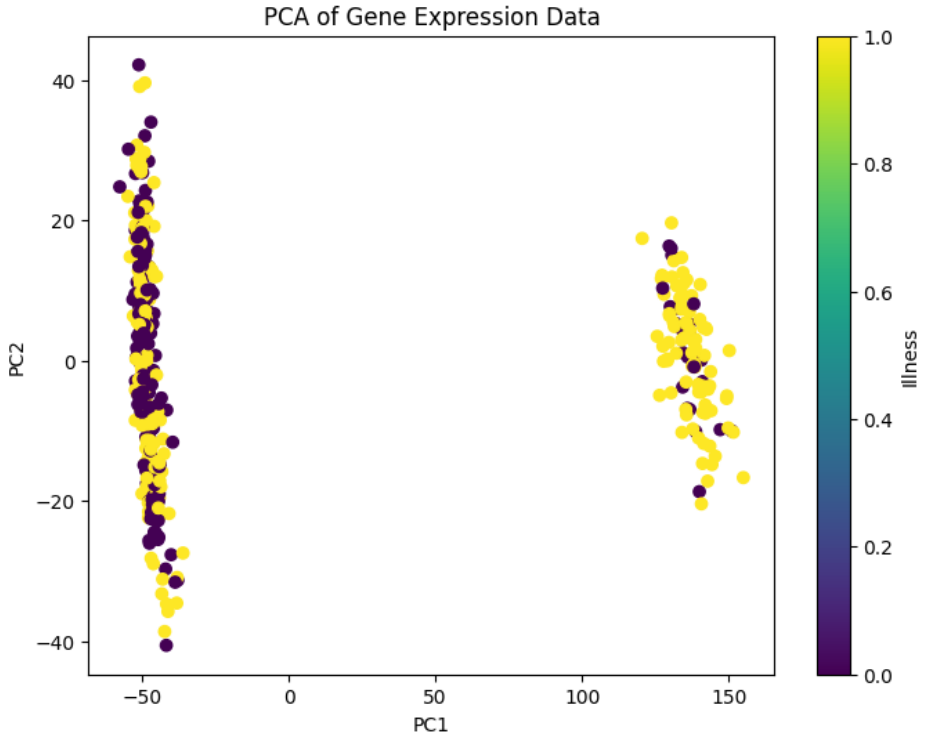

Figura 3.1: Grafico prima della correzione del batch effect.

Dopo aver applicato la correzione del batch effect, il grafico in Figura 3.2 mostra una distribuzione molto più uniforme dei campioni, senza evidenti separazioni che possano essere attribuite a differenze tecniche. In contrasto con il grafico precedente, non ci sono più raggruppamenti netti o separazioni lungo le componenti principali. Questo risultato suggerisce che il batch effect è stato efficacemente ridotto, consentendo una maggiore focalizzazione sulle variazioni biologiche reali nei dati, che ora risultano meno influenzati da fattori tecnici legati. La correzione del batch effect è stata quindi cruciale per migliorare le analisi successiva.

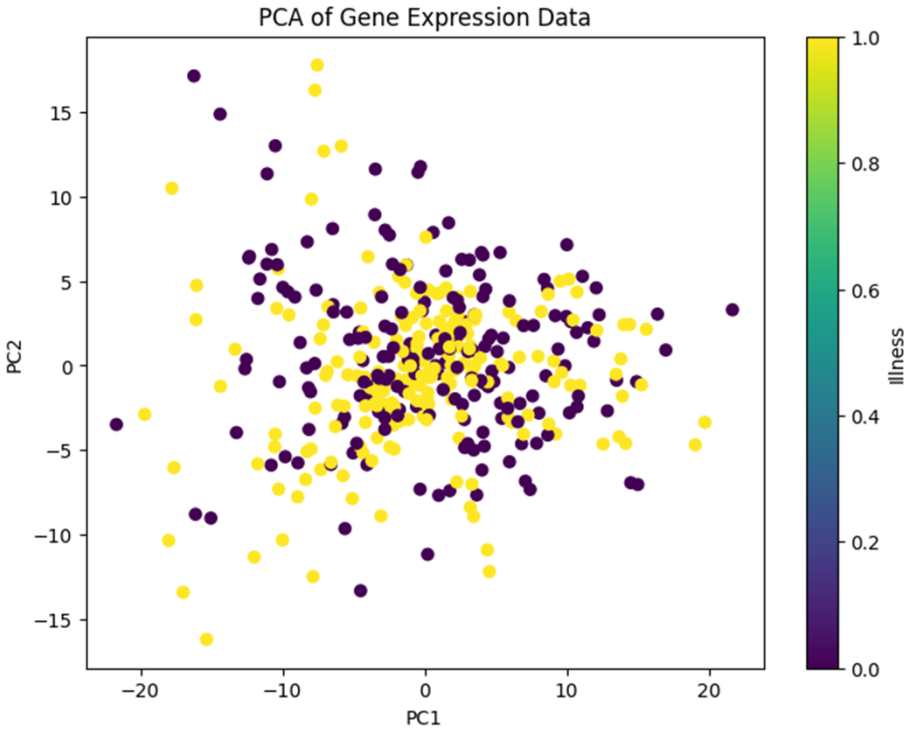

Figura 3.2: Grafico dopo la correzione del batch effect.

Definizione del threshold. L'applicazione di una soglia (threshold) sui dati di espressione genica è una tecnica fondamentale per escludere i valori meno rilevanti e migliorare la qualità complessiva delle analisi successive. In questa ricerca, è stato deciso di fissare una soglia pari a 5, al fine di escludere quei geni la cui espressione risultava troppo bassa e presumibilmente non rilevante per lo studio in corso. La scelta di tale soglia è stata motivata dall'analisi della distribuzione dei dati, visualizzata tramite un istogramma. Come mostrato in Figura 3.3, l'istogramma della distribuzione dei valori di espressione genica prima dell'applicazione del threshold presenta picchi principali compresi tra i valori di 4 e 8.

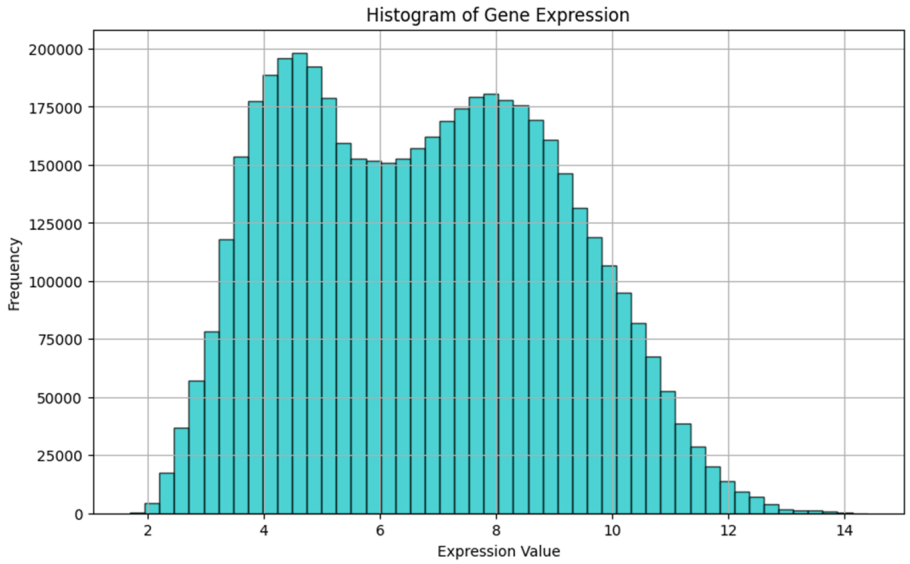

Figura 3.3: Distribuzione dei valori prima dell'applicazione del threshold.

L'adozione di un threshold pari a 5 permette quindi di escludere i valori inferiori, che tendono a rappresentare espressioni deboli o geni scarsamente attivi, i quali potrebbero influire negativamente sull'interpretazione dei risultati. In Figura 3.4, possiamo osservare l'effetto dell'applicazione del threshold: i valori inferiori alla soglia sono stati eliminati, portando a una distribuzione dei dati più focalizzata sulle espressioni geniche rilevanti.

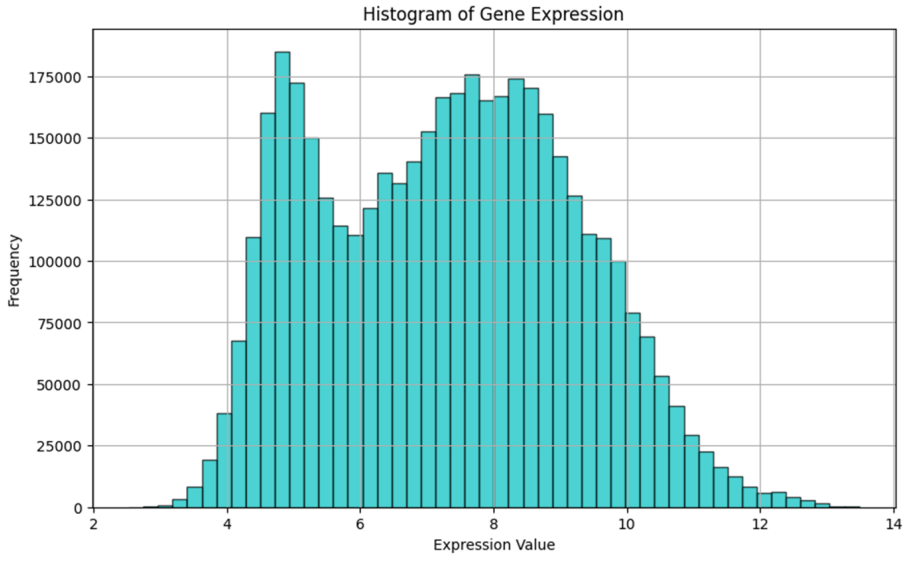

Figura 3.4: Distribuzione dei valori dopo l'applicazione del threshold.

Questa operazione di filtering non solo riduce il rumore presente nei dati, ma migliora anche l'efficienza computazionale nelle fasi successive dell'analisi, come la selezione delle feature e l'addestramento dei modelli.

Feature Selection. La selezione delle feature è una fase cruciale nell'analisi dei dati, in quanto permette di identificare le caratteristiche più rilevanti per il problema in esame, riducendo al contempo la dimensionalità del dataset. In questo studio, la selezione delle feature è stata eseguita utilizzando un approccio guidato dalla letteratura scientifica e dalle risorse bioinformatiche, in particolare tramite l'utilizzo della piattaforma KEGG (Kyoto Encyclopedia of Genes and Genomes) [28, 29, 30], riconosciuto nell'ambito biologico e bioinformatico. Attraverso KEGG, è stato possibile identificare i geni maggiormente associati alla malattia in esame. Sono stati presi in considerazione sia i geni che sono direttamente correlati alla patologia, come accennato nella sezione 2.4.1, sia i geni appartenenti al pathway di insorgenza della malattia. Questo approccio combinato consente di catturare sia i fattori diretti che i meccanismi sottostanti l'insorgenza della patologia, rendendo l'analisi più completa. KEGG PATHWAY Database è stato utilizzato per recuperare i geni coinvolti nei processi biologici correlati all'insorgenza della malattia. In particolare, è stato considerato il pathway con entry `hsa04940`, che rappresenta il percorso biologico chiave nel contesto della malattia. Inoltre, è stato consultato anche il KEGG DISEASE Database, in cui è disponibile l'entry `H00408`, che raccoglie i geni strettamente legati alla patologia. I geni presi in considerazione sono riportati nella Tabella 3.1 e nella Tabella 3.2

Tabella 3.1: Geni pathtway di insorgenza.

Entry	Simbolo	Nome
3630	INS	insulin
2571	GAD1	glutamate decarboxylase 1
2572	GAD2	glutamate decarboxylase 2
5798	PTPRN	protein tyrosine phosphatase receptor type N
5799	PTPRN2	protein tyrosine phosphatase receptor type N2
1363	CPE	carboxypeptidase E
3329	HSPD1	heat shock protein family D (Hsp60) member 1
3382	ICA1	heat shock protein family D (Hsp60) member 1
3108	HLA-DMA	major histocompatibility complex, class II, DM alpha
3109	HLA-DMB	major histocompatibility complex, class II, DM beta
3111	HLA-DOA	major histocompatibility complex, class II, DO alpha
3112	HLA-DOB	major histocompatibility complex, class II, DO alpha
3113	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1
3115	HLA-DPB1	HLA-DPB1; major histocompatibility complex, class II, DP beta 1
3117	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1
3118	HLA-DQA2	major histocompatibility complex, class II, DQ alpha 2
3119	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1
3122	HLA-DRA	major histocompatibility complex, class II, DR alpha
3123	HLA-DRB1	major histocompatibility complex, class II, DR beta 1
3125	HLA DRB3	major histocompatibility complex, class II, DR beta 3
3126	HLA-DRB4	HLA-DRB4; major histocompatibility complex, class II, DR beta 4
3127	HLA-DRB5	major histocompatibility complex, class II, DR beta 5
941	CD80	CD80 molecule
942	CD86	CD86 molecule
940	CD28	CD28 molecule
3592	IL12A	interleukin 12A
3593	IL12B	interleukin 12B
3558	IL2	interleukin 2

3458	IFNG	interferon gamma
3105	HLA-A	major histocompatibility complex, class I, A
3106	HLA-B	major histocompatibility complex, class I, B
3107	HLA-C	major histocompatibility complex, class I, C
3134	HLA-F	HLA-F; major histocompatibility complex, class I, F
3135	HLA-G	major histocompatibility complex, class I, G
3133	HLA-E	major histocompatibility complex, class I, E
356	FASLG	Fas ligand
355	FAS	fas cell surface death receptor
5551	PRF1	perforin 1
3002	GZMB	granzyme B
4049	LTA	lymphotoxin alpha
7124	TFN	tumor necrosis factor
3552	IL1A	interleukin 1 alpha
3553	IL1B	interleukin 1 beta

Tabella 3.2: Geni associati alla malattia.

Simbolo	Nome
INS	insulin
SUMO4	small ubiquitin like modifier 4
IL2RA	interleukin 2 receptor subunit alpha
CTLA4	cytotoxic T-lymphocyte associated protein 4
HNF1A	transcription factor 1, hepatocyte nuclear factor 1-alpha
CCR5	C-C motif chemokine receptor 5
HLA-DRB1	major histocompatibility complex, class II, DR beta 1
HLA-DQB1	major histocompatibility complex, class II, DQ beta 1
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1
PTPN22	protein tyrosine phosphatase non-receptor type 22
PTPN2	protein tyrosine phosphatase non-receptor type 2

ERBB3	erb-b2 receptor tyrosine kinase 3
IL2	interleukin 2
IL21	interleukin 21
IFIH1	interferon induced with helicase C domain 1
CLEC16A	C-type lectin domain containing 16A
BACH2	BTB domain and CNC homolog 2
CTSH	cathepsin H
SH2B3	SH2B adaptor protein 3
C12orf30	N-alpha-acetyltransferase 25, NatB auxiliary subunit
CD226	CD226 molecule
ITPR3	inositol 1,4,5-trisphosphate receptor type 3
CRYP27B1	cytochrome P450 family 27 subfamily B member 1

Bilanciamento del dataset. Il bilanciamento del dataset è una fase critica quando la variabile target, o variabile dipendente, presenta una distribuzione diseguale tra le classi. In questo studio, dopo il preprocessing, si è riscontrato uno squilibrio nella distribuzione delle classi: la classe “Type 1 Diabetes” aveva un numero maggiore di campioni rispetto a “Healthy”. Questo sbilanciamento avrebbe potuto influenzare negativamente il processo di addestramento del modello, portando a previsioni distorte verso la classe più rappresentata. Per risolvere questo problema, è stata adottata la tecnica dell’undersampling, che consiste nel ridurre il numero di campioni della classe maggioritaria (“Type 1 Diabetes”) in modo da uguagliare il numero di campioni della classe minoritaria (“Healthy”). È stato scelto `StratifiedShuffleSplit` di scikit-learn per eseguire questa operazione, poiché garantisce che la suddivisione mantenga la proporzione delle etichette in modo casuale e stratificato, preservando l’integrità del dataset durante il processo di riduzione. Prima del bilanciamento, erano presenti 210 campioni con etichetta “Type 1 Diabetes” e 193 campioni con etichetta “Healthy”, evidenziando uno squilibrio verso la classe “Type 1 Diabetes”. Dopo l’applicazione della tecnica di undersampling, entrambe le classi sono state ridotte a 193 campioni, creando così un dataset bilanciato.

CAPITOLO 4

Modelli di predizione e confronto

In questo capitolo verranno illustrati i principali algoritmi di machine learning che sono stati selezionati e implementati per affrontare l'obiettivo della nostra ricerca: la classificazione binaria per la predizione del T1D. In particolare, ci concentreremo sui modelli XGBoost, Support Vector Machine (SVM) e Random Forest, che sono stati scelti dopo una valutazione della letteratura scientifica riportata in 2.5. Esamineremo, quindi, le motivazioni che ci hanno spinto a selezionare questi specifici modelli, le metodologie adottate per il loro addestramento e i risultati ottenuti sia nelle fasi di training che di testing.

XGBoost. XGBoost (Extreme Gradient Boosting) è un algoritmo di machine learning che si distingue per le sue prestazioni elevate e per la sua efficienza sia in termini di velocità che di accuratezza. È utilizzato sia per problemi di regressione che per problemi di classificazione; inoltre, è particolamente efficiente quando si lavora con dataset complessi e con alta dimensionalità. Si basa sul gradient boosting, una tecnica di machine learning che permette la costruzione di un mdello predittivo attraverso una combinazione di modelli deboli, solitamente alberi decisionali. L'idea si basa sul migliorare iterativamente un modello correggendo gli errori commessi dai modelli precedenti. Si parte con una previsione iniziale; in seguito, ad ogni passo viene costruito un nuovo albero decisionale che cerca di correggere gli errori commessi dal modello complessivo precedente. Invece che addestrare un nuovo albero sui valori originali, l'algoritmo addestra ogni nuovo albero sui residui o errori dei modelli precedenti. Essi rappresentano la differenza tra la previsione fatta dal modello e il valore reale. Il gradiente è la direzione che va a migliorare le previsioni. L'aggiornamento del modello avviene in modo che il modello complessivo venga aggiornato combinando i modelli precedenti con quello nuovo, solitamente sommando il nuovo albero con un peso, chiamato tasso di apprendimento (learning rate), per migliorare la generalizzazione. Questo processo si ripete finché non si raggiunge un numero massimo di iterazioni o finché l'errore non è sufficientemente ridotto. Tra i vantaggi di questo algoritmo c'è l'alta flessibilità, che permette di applicarlo a molte tipologie di problemi; tra i principali svantaggi, invece, abbiamo il tempo di addestramento che può oneroso, specialmente su dataset molto grandi. I parametri che sono stati ottimizzati durante l'addestramento sono:

- "colsample_bytree": Controlla quante caratteristiche saranno considerate a ogni split durante la costruzione degli alberi.
- "learning_rate": Rappresenta il tasso di apprendimento. Valori più bassi comportano aggioramenti più piccoli dei pesi del modello, il che richiede un numero maggiore di alberi, ma permette di catturare pattern più complessi.
- "max_depth": Controlla la profondità massima di ciascun albero.
- "n_estimators": Indica il numero di alberi che saranno addestrati nel modello.

- "subsample": Rappresenta la frazione del dataset che ogni albero deve campionare per l'addestramento.
- "min_child_weight": Controlla la somma minima dei pesi delle osservazioni necessarie per dividere un nodo.
- "gamma": Controlla la sensibilità del modello alla creazione di nuovi nodi durante la costruzione degli alberi.
- "reg_alpha": È il coefficiente di regolarizzazione L1 applicato sui pesi delle foglie: aggiunge una penaità proporzionale alla somma dei valori assoluti dei pesi delle foglie.
- "reg_lambda": È il coefficiente di regolarizzazione L2 applicato sui pesi delle foglie degli alberi: aggiunge una penalità proporzionale al quadrato dei valori dei pesi delle foglie.

SVM L'algoritmo SVM (Support Vector Machine) è un algoritmo di apprendimento supervisionato utilizzato principalmente per compiti di classificazione. L'SVM lavora cercando di trovare un iperpiano che separi i dati appartenenti a diverse classi nel modo più efficace possibile. L'obiettivo è quello di massimizzare il margine, ovvero la distanza tra l'iperpiano e i punti di dati più vicini ad esso, chiamati support vectors. In un caso semplice, l'SVM cerca di tracciare una linea di separazione che divide due classi di dati: se le due classi di dati sono linearmente separabili, l'SVM riesce a trovare l'iperpiano che massimizza il margine. In altri casi, i dati non possono essere separati linearmente: l'SVM utilizza una tecnica chiamata kernel trick, che trasforma i dati in uno spazio di dimensioni più alte, dove può essere trovata una separazione lineare. Uno dei vantaggi principali dell'utilzzo di SVM è l'efficacia con dati complessi: può essere particolarmene utile quando il numero di feature risulta elevato rispetto al numero di campioni. Inoltre, SVM tende a generalizzare bene, porprio perché cerca di massimizzare il margine tra le classi, il che riduce il rischio di overfitting. Uno dei principali svantaggi è che quando si utilizzano kernel non lineari può risultare computazionalmente costoso per dataset molto grandi, poiché richiede

operazioni complesse per trasformare i dati nello spazio più ampio. I parametri che sono stati ottimizzati durante l'addestramento sono:

- "C (Cost)": Controlla quanto il modello penalizza gli errori di classificazione sui dati di training.
- "Kernel": L'SVM può utilizzare diversi tipi di kernel per separare i dati quando la separazione lineare non è possibile.
- "gamma": Definisce quanto lontano l'influenza di un singolo punto di dati si estende; viene utilizzato quando si impiegano kernel non lineari.

Random Forest Il Random Forest è un algoritmo di apprendimento supervisionato, comunemente utilizzato per compiti di classificazione e regressione, e si basa sull'idea di combinare più modelli semplici per creare un modello robusto e accurato. Il modello Random Forest appartiene alla famiglia degli ensemble learning, dove si costruisce una collezione di alberi decisionali e si combina il risultato di ciascuno utilizzando il voto di maggioranza per le classificazioni, oppure la media per le regressioni, al fine di ottenere una previsione finale. La costruzione del random forest si basa sulla costruzione di un numro predeterminato di alberi decisionali, ciascuno dei quali viene addestrato su un sottoinsieme casuale di dati di training, selezionato tramite campionamento con ripetizione. Questa tecnica, nota come bootstrap aggregation, permette di generare alberi diversi tra loro, riducendo la varianza del modello. Tra i vantaggi del Random Forest c'è la capacità di generalizzare bene su dati mai visti. È noto anche per essere un modello che richiede poca ottimizzazione dei parametri rispetto ad altri modelli complessi. Presenta, però, alcuni limiti: il tempo di esecuzione può essere significativo considerando dataset molto grandi. I parametri che sono stati ottimizzati durante l'addestramento sono:

- "n_estimators": Come in XGBoost, indica il numero di alberi che saranno addestrati nel modello.
- "max_depth": Come in XGBoost, controlla la profondità massima di ciascun albero.

- "min_sample_split": Controlla il numero minimo di campioni richiesto per dividere un nodo. Se il numero di campioni in un nodo è inferiore a questo valore, il nodo non verrà diviso ulteriormente.
- "min_samples_leaf": Indica il numero minimo di campioni che devono essere presenti in un nodo foglia.
- "bootstrap": Controlla se il campionamento con ripetizione viene utilizzato per creare i sottoinsiemi di dati per ogni albero. Se impostato su True, ogni albero viene addestrato su un sottoinsieme casuale dei dati di training; se impostato su False, ogni albero viene addestrato sull'intero dataset senza campionamento con ripetizione.

4.1 Addestramento e validazione

In questo studio l'addestramento e la validazione dei modelli sono stati divisi in tre fasi principali:

- **Identificazione dei range per i parametri**: Per esplorare lo spazio dei parametri al meglio è stato scelto di utilizzare Optuna: permette di andare a specificare un intervallo di valori per ogni iperparametro. Optuna è una libreria open-source per l'ottimizzazione automatica degli iperparametri. Si basa sulla ricerca bayesiana, che permette di identificare rapidamente le combinazioni di iperarametri più promettenti; effettua una ricerca intelligente basata sui risultati delle prove precedenti. Nello studio è stato scelto di andare a massimizzare F1-score, la media armonica tra precision e recall. Oltre ad utilizzare optuna, è stato utilizzato anche `StratifiedKFold`: divide il dataset in modo che ogni fold contenga la stessa proporzione di etichette di classe presenti nel dataset. Per ogni fold, viene utilizzata una parte dei dati come set di addestramento e la parte rimanente come validazione. Questo processo viene ripetuto fino a quando ogni fold ha avuto la possibilità di essere utilizzata come set di validazione. C'è da precisare che queste operazioni sono state effettute su l'80% del dataset.

- **Ottimizzazione e Addestramento Finale**: Una volta individuati i range di parametri più interessanti, è stato utilizzato `GridSearchCV` per eseguire un'ottimizzazione finale e sistematica dei modelli. Sono state esplorate combinazioni dei parametri selezionati per ogni algoritmo. Vengono riportati, di seguito, le combinazini di parametri che hanno portato i risultati migliori:

Tabella 4.1: Valori per XGBoost.

Parametro	Valore
colsample_bytree	0.5
learning_rate	0.01
max_depth	15
n_estimators	150
subsample	0.6
min_child_weight	1
gamma	0.1
reg_alpha	0
reg_lambda	1

Tabella 4.2: Valori per SVM.

Parametro	Valore
C	100
gamma	3.0
kernel	rbf

Tabella 4.3: Valori per Random Forest.

Parametro	Valore
n_estimators	400
max_depth	25
min_samples_split	6
min_samples_leaf	1
bootstrap	True

Per questi parametri è stata utilizzata una cross-validation di 10 fold in modo da testare ogni combinazone di iperparametri su più sottoinsiemi di dati, riducendo il rischio che i parametri scelti siano ottimali sono per una singola suddivisione.

- **Validazione**: Dopo aver identificato i parametri ottimali, è stato effettutato l'addestramento finali del modello utilizzando l'80% del dataset. Questa suddivisione ha permesso di sfruttare a pieno i dati disponibili, mantenendo un 20% di dati non utilizzati per la valutazione finale. La scelta di riservare una parte del dataset come set di test indipendente è fondamentale per valutare un modo oggettivo le prestazioni dei modelli e per garantire che il modello sia generalizzabile e in grado di fornire buone prestazioni anche su dati non visti, diminuendo il rischio di overfitting.

4.2 Metriche e confronto

Per valutare le prestazioni dei modelli di classificazione, sono state utilizzate diverse metriche, ciascuna delle quali fornisce una prospettiva diversa sulle prestazioni del modello. Classificheremo con:

- **TP** (True Positive): Numero di campioni positivi correttamente classificati.
- **TN** (True Negative): Numero di campioni negativi correttamente classificati.
- **FP** (False Positive): Numero di campioni negativi classificati come positivi.
- **FN** (False Negatives): Numero di campioni positivi classificati come negativi.

Accuracy L'accuracy è la percentuale di previsioni corrette rispetto al totale delle previsioni effettuate. È utile quando il dataset è bilanciato, ma può essere fuorviante in presenza di classi squilibrate, dove una classe domina sull'altra.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Recall Il recall rappresenta la percentuale di campioni positivi correttamente identificati tra tutti i campioni positivi reali. È particolarmente importante quando si

vuole minimizzare il numero di falsi negativi, ovvero evitare di perdere casi positivi. Un recall elevato indica che il modello è efficace nell'identificare la maggior parte dei campioni positivi.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision La precision misura la percentuale di previsioni positive corrette sul totale delle previsioni positive. È particolarmente utile quando l'obiettivo è minimizzare i falsi positivi. Una precisione elevata indica che pochi campioni negativi sono stati classificati come positivi, ovvero una bassa incidenza di falsi positivi.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

F1 Score L'F1 Score è la media armonica tra precision e recall. È utile quando si ha bisogno di un compromesso tra precision e recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC-ROC L'AUC-ROC misura la capacità del modello di distinguere tra classi. La ROC Curve è un grafico che traccia il tasso di veri positivi (TPR o Sensibilità) contro il tasso di falsi positivi (FPR) a vari livelli di soglia. L'area sotto questa curva (AUC) fornisce un'indicazione della performance del modello: un valore vicino a 1 indica un ottimo modello, mentre un valore vicino a 0.5 indica che il modello non è migliore del caso. L'AUC è calcolata integrando l'area sotto la curva ROC. Una AUC elevata indica che il modello è in grado di separare efficacemente le classi.

True Positive Rate

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

False Positive Rate

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Risultati. In questa sezione vengono presentati e analizzati i risultati ottenuti dai modelli. Le performance sono valutate in base alle metriche descritte in precedenza,

fornendo un quadro completo dell'efficacia dei modelli nel contesto dell'applicazione considerata. I dati e le osservazioni riportati mettono in luce sia i punti di forza che le eventuali limitazioni del modello, permettendo di comprendere a fondo le sue capacità predittive. Di seguito viene riportata la tabella riassuntiva delle metriche di valutazione, che mostra i risultati ottenuti dai modelli per ciascuna misura considerata.

Tabella 4.4: Metriche dei modelli.

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

Partendo dall'accuracy notiamo come sia il modello Random Forest che XGBoost hanno raggiunto un'accuratezza del 79.5%, mentre il modello SVM si distingue leggermente, ottenendo un'accuratezza del 80.8%. Questo suggerisce che l'SVM ha una capacità leggermente migliore nel classificare correttamente i campioni nel complesso, anche se la differenza rispetto a Random Forest e XGBoost è marginale. Tuttavia, l'accuracy da sola non risulta sufficiente, poiché non distingue tra i vari tipi di errori (falsi positivi e falsi negativi) e può essere fuorviante se utilizzata come unica misura. Passando alla precision, vediamo che anche qui l'SVM raggiunge il valore più alto, con 81.3%, mentre Random Forest e XGBoost seguono con 79.8%. Una precision più alta indica che l'SVM è più accurato quando classifica un campione come positivo, riducendo così i falsi allarmi. Il recall rivela un quadro simile, con l'SVM che mantiene un valore più alto, 80.8%, rispetto a Random Forest e XGBoost, entrambi con 79.5%. Questo risultato indica che l'SVM riesce a identificare correttamente una percentuale più alta dei veri positivi, il che è cruciale in contesti in cui è importante catturare tutti i casi positivi, come in questo studio. L'F1-Score conferma la solidità dell'SVM con un valore di 80.7%, leggermente superiore ai 79.4% di Random Forest e XGBoost.

Questo punteggio evidenzia come l'SVM riesca a mantenere un buon compromesso tra l'accuratezza delle previsioni positive e la capacità di catturare tutti i casi positivi. L'F1-score ci dice che l'SVM è il modello più adatto, poiché è quello che meglio bilancia precision e recall. Passando ad AUC-ROC, vengono riportate in Figura 4.1, Figura 4.2 e Figura 4.3 le curva di ROC per ciascun modello.

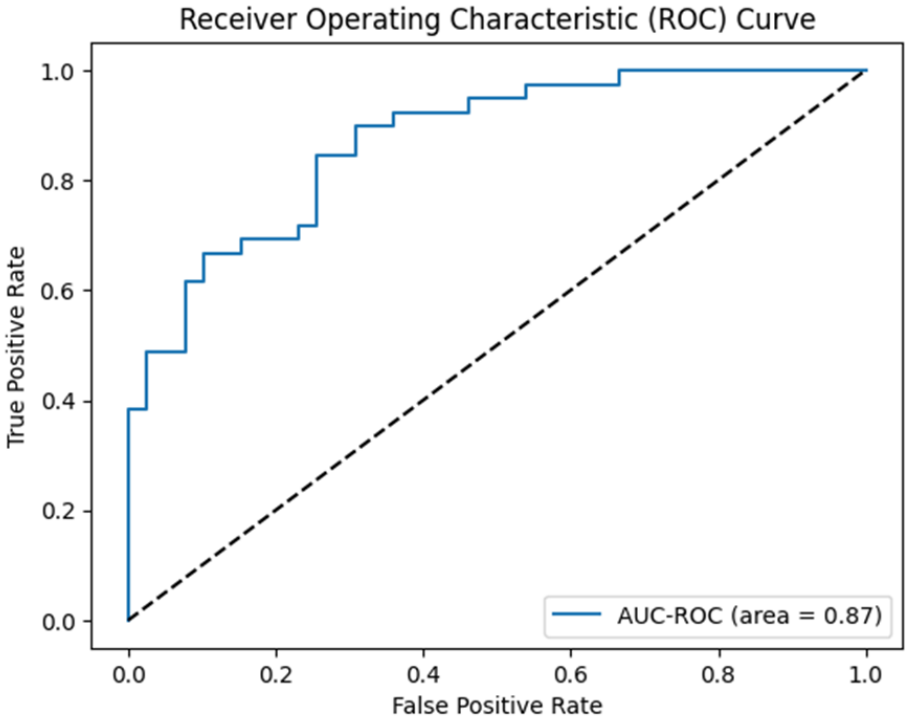

Figura 4.1: Curva di ROC per XGBoost.

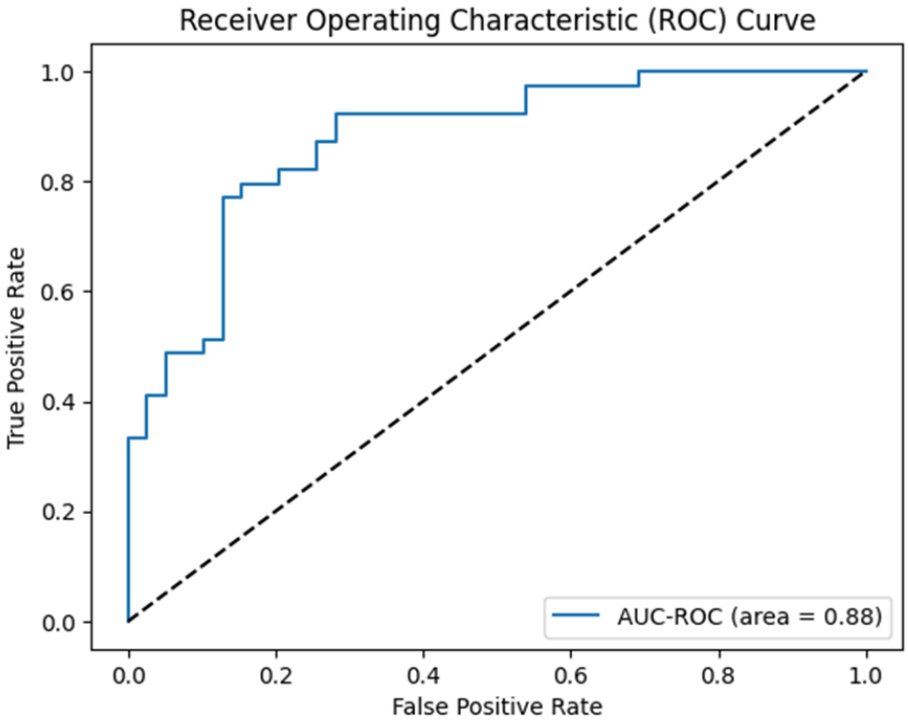

Figura 4.2: Curva di ROC per SVM.

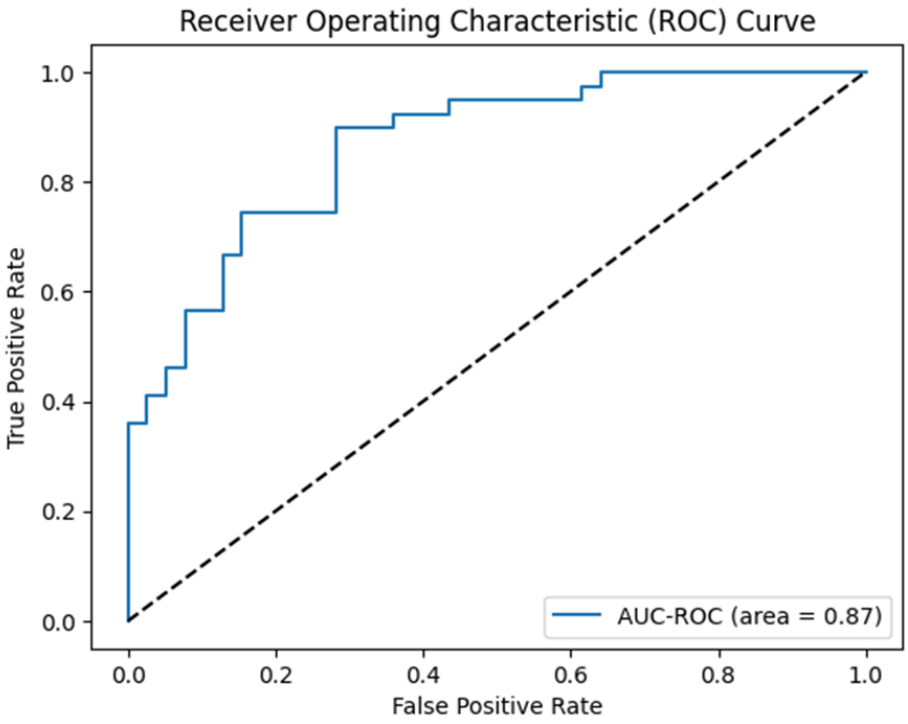

Figura 4.3: Curva di ROC per Random Forest.

L'AUC-ROC rafforza ulteriormente la posizione dell'SVM come modello migliore, con un valore di 87.6%, superiore a quello di 87.4% per Random Forest e 87.2% per XGBoost. Un valore di AUC-ROC elevato come quello dell'SVM suggerisce che il modello riesce a separare efficacemente le classi positive e negative, fornendo una maggiore flessibilità per ottimizzare le prestazioni a seconda delle priorità. Nel complesso, l'SVM si dimostra il modello più performante tra i tre, con i migliori risultati per tutte le metriche chiave. Questo modello offre un buon equilibrio tra precision e recall, riducendo sia i falsi positivi che i falsi negativi. L'accuratezza e l'AUC-ROC elevati indicano che l'SVM è robusto e può discriminare bene tra le classi, rendendolo ideale per contesti in cui sia importante identificare correttamente entrambe le classi con alta affidabilità. Random Forest e XGBoost mostrano prestazioni quasi identiche e costituiscono comunque valide alternative. Entrambi i modelli sono particolarmente utili in contesti dove si apprezza l'interpretabilità o si richiedono prestazioni computazionali ottimizzate. Tuttavia, rispetto all'SVM, questi modelli non riescono a eguagliarne l'efficacia complessiva nella classificazione su questo dataset specifico. In conclusione, considerando che l'obiettivo primario è massimizzare la generalizzazione e ottenere un modello in grado di bilanciare efficacemente l'identificazione dei casi positivi con la riduzione dei falsi positivi, l'SVM rappresenta la scelta più indicata.

CAPITOLO 5

Explainability e discussione

In questo capitolo, verranno presentati i principali metodi utilizzati per garantire l'explainability del modello, evidenziando l'importanza di interpretare le predizioni dei modelli di machine learning, soprattutto in un contesto sensibile come quello della predizione del T1D. Inizieremo con una spiegazione delle motivazioni che hanno portato alla scelta di esplorare l'explainability dei modelli, sottolineando l'esigenza di comprendere come le feature influenzino le decisioni del modello per ottenere previsioni affidabili e trasparenti. Successivamente, verranno introdotte le due tecnologie principali utilizzate per l'analisi interpretativa: Anchor e SHAP. Anchor si concentra sulla generazione di regole locali, facilitando la comprensione delle decisioni del modello per singoli casi tramite condizioni semplici, mentre SHAP offre una prospettiva globale e quantifica il contributo di ogni feature, rendendo evidenti i pattern generali che influenzano le predizioni del modello. Infine, verranno discussi i risultati delle analisi per ciascun modello. Attraverso i summary plot di SHAP e le regole generate con Anchor, verrà mostrato come le feature principali, come PTPRN2, emergano come variabili chiave nella classificazione. Questo capitolo fornirà una visione dettagliata su come i diversi metodi di explainability si completino a vicenda, offrendo una comprensione sia globale che locale delle dinamiche del modello.

Explainability e la sua importanza. L'explainability, o capacità di spiegare, si riferisce alla possibilità di comprendere e interpretare le decisioni prese da un modello di machine learning. In un contesto in cui i modelli di apprendimento automatico trovano applicazione in ambiti critici è fondamentale che le previsioni e le decisioni di questi sistemi non siano solo accurate, ma anche trasparenti e comprensibili. Un modello spiegabile consente di ottenere una comprensione dettagliata del perché e del come delle sue decisioni, favorendo la fiducia tra gli utenti e i decisori. L'importanza dell'explainability è particolarmente evidente nei modelli complessi o black-box, come le reti neurali profonde o i modelli di ensemble, che, pur raggiungendo elevate performance, risultano spesso opachi e di difficile interpretazione. Senza una spiegazione chiara, le decisioni di questi modelli possono apparire arbitrarie e non facilmente giustificabili, limitando la loro applicabilità in settori regolamentati o ad alto impatto. L'explainability fornisce uno strumento essenziale per identificare e mitigare potenziali bias o comportamenti inaspettati. Fornendo spiegazioni dettagliate, gli utenti possono valutare se il modello si basa su variabili rilevanti e se le sue decisioni sono coerenti con i principi del dominio di applicazione.

È stato scelto, quindi, di includere l'explainability per migliorare la trasparenza e promuovere un'adozione consapevole dei modelli di machine learning, specialmente in ambito medico. In questo contesto, dove le decisioni hanno un impatto diretto sulla salute dei pazienti, è essenziale che i modelli non siano solo accurati, ma anche interpretabili. La capacità di spiegare le scelte del modello consente agli esperti di confrontare i risultati con le loro conoscenze cliniche e di validare le decisioni. L'explainability favorisce la fiducia degli utenti finali, poiché fornisce le informazioni necessarie per comprendere le risposte del modello e valutare eventuali discrepanze. Questo approccio non solo facilita l'integrazione della tecnologia nella pratica clinica, ma crea una base per una collaborazione più solida tra l'intelligenza artificiale e gli specialisti, promuovendo un utilizzo etico e affidabile in settori critici come quello sanitario.

Tecnologie. Anchor [31] e SHAP [32] sono due tecniche di interpretazione del machine learning, ciascuna con caratteristiche uniche che le rendono complementari per ottenere una comprensione più completa delle previsioni di un modello. Anchor è

progettato per fornire spiegazioni locali e intuitive delle previsioni, generando "regole di ancoraggio" che spiegano le decisioni del modello tramite semplici condizioni del tipo "se... allora...". Queste regole identificano le caratteristiche chiave che portano il modello a fare una determinata previsione, offrendo così un'interpretazione chiara e facilmente comprensibile anche per utenti non esperti. Anchor si concentra su condizioni specifiche che, se soddisfatte, portano a una previsione con un elevato grado di precisione. SHAP (SHapley Additive exPlanations), d'altro canto, è basato sui valori di Shapley della teoria dei giochi e assegna a ciascuna feature un valore di importanza che rappresenta il suo contributo alla previsione finale del modello. SHAP fornisce sia interpretazioni globali (per tutto il modello) sia locali (per singole previsioni), permettendo di comprendere non solo l'impatto di una feature specifica su una singola previsione, ma anche come le feature influenzano le previsioni in modo complessivo e sistematico.

Sono stati scelti, quindi, Anchor e SHAP rispetto ad altri strumenti per le loro capacità complementari di fornire spiegazioni locali e globali, essenziali in ambito medico, dove è fondamentale che le decisioni dei modelli siano trasparenti e facilmente comprensibili. Anchor, con le sue regole, permette di spiegare in modo chiaro e intuitivo le singole predizioni, identificando le condizioni specifiche che portano a una decisione clinica. A differenza di altri metodi, Anchor garantisce precisione e coerenza nelle spiegazioni locali, riducendo la possibilità di variabilità nei risultati. SHAP, invece, offre una visione globale del modello e, grazie ai valori di Shapley, quantifica con rigore il contributo di ciascuna feature. Questo è particolarmente utile per validare le spiegazioni locali di Anchor, assicurandosi che siano in linea con l'andamento complessivo del modello e aiutando gli esperti a confrontare le decisioni del modello con le loro conoscenze cliniche. La combinazione di Anchor e SHAP, quindi, permette di ottenere spiegazioni affidabili e trasparenti, sempre allo scopo di migliorare la fiducia degli esperti nelle decisioni del modello e supportando l'adozione responsabile della tecnologia nel settore sanitario.

5.1 Explainability con Anchor.

Di seguito verranno presentati i risultati ottenuti utilizzando Anchor, con un'attenzione particolare ai campioni che hanno mostrato le migliori performance in termini di coverage. Le principali metriche utilizzate da Anchor sono la precision e la coverage. Tutti i risultati sono stati generati impostando un threshold di 0.8 su Anchor, il che significa che sono state selezionate le regole che garantiscono una precisione di almeno l'80% per la previsione. Questo threshold è stato scelto per bilanciare l'affidabilità delle spiegazioni con la loro copertura; un valore inferiore avrebbe potuto includere regole meno precise, rischiando di ridurre la qualità delle interpretazioni fornite. D'altra parte, un threshold troppo elevato avrebbe ridotto ulteriormente la coverage, limitando il numero di campioni per cui è possibile ottenere spiegazioni interpretabili.

XGBoost. Nell'analisi dei risultati di XGBoost applicato al dataset, emergono alcune feature chiave che il modello considera particolarmente informative per le predizioni. Tra queste, il gene PTPRN2 risulta essere uno dei principali indicatori. Il modello associa frequentemente livelli elevati di espressione di PTPRN2 alla classe negativa, mentre valori inferiori tendono a rappresentare un marker per la classe positiva. Questo comportamento suggerisce che PTPRN2 potrebbe giocare un ruolo significativo nel distinguere i campioni. Oltre a PTPRN2, altre feature come HLA-DOB, IL2RA, TNF, e CTSH sono spesso presenti nei pattern di anchor, confermando la loro importanza. In particolare, HLA-DOB e IL2RA sono utilizzati dal modello in combinazione con PTPRN2 per perfezionare ulteriormente la classificazione. Ad esempio, quando IL2RA assume valori bassi e HLA-DOB valori alti, la probabilità di classificare un campione come negativo aumenta considerevolmente. Al contrario, bassi livelli di TNF e alti livelli di CTSH sono indicatori forti per la classe positiva, riflettendo l'importanza di queste variabili nella differenziazione tra le classi. Per illustrare ulteriormente l'importanza di queste feature, consideriamo alcuni esempi specifici di regole che mostrano elevate prestazioni sia in termini di precisione che di copertura. In un caso di predizione positiva un anchor che combina le condizioni `['PTPRN2 <= 6.68', 'TNF <= 7.46', 'CTSH > 9.96']` raggiunge una

precisione di circa 93.6% e una copertura dell'8%. Questo indica che, quando PTPRN2 è basso, TNF è contenuto e CTSH è elevato, il modello ha un'elevata sicurezza nel predire la classe positiva. Un altro esempio significativo riguarda la predizione negativa con l'anchor `['PTPRN2 > 6.68']`, che copre una larga fetta dei campioni (fino al 25%) e mantiene una precisione stabile intorno all'85%. Questo suggerisce che un valore elevato di PTPRN2 è un forte indicatore della classe negativa, un trend che il modello sfrutta frequentemente. Un ulteriore esempio è un anchor per la classe positiva `['PTPRN2 <= 6.68', 'HLA-DQB1 <= 3.42', 'IL21 > 4.05']`, che raggiunge una precisione perfetta (1.0) con un coverage del 4%. Questa regola, molto specifica e accurata, identifica campioni positivi con elevata sicurezza, suggerendo che valori bassi di PTPRN2, HLA-DQB1 e valori alti di IL21 sono caratteristiche distintive dei campioni appartenenti alla classe positiva.

SVM. L'analisi delle regole generate dal modello SVM mostra che il gene PTPRN2 gioca un ruolo centrale localmente, specialmente nelle predizioni per la classe negativa, in cui si combina frequentemente con altri marcatori come IL2RA e HLA-DOA. Ad esempio, la regola `['PTPRN2 > 6.68', 'IL2RA <= 6.12']` ha una copertura del 7.95%, con una precisione dell'80.1%, suggerendo che elevati livelli di PTPRN2 con bassi valori di IL2RA contribuiscono fortemente alla classificazione negativa. Alcune regole per la classe positiva presentano coperture ancora maggiori. Ad esempio, la regola `['HNF1A <= 5.06', 'HLA-F <= 10.67', 'HLA-DQA2 > 4.99']` copre una considerevole percentuale del campione, pari al 15.83%, e ha una precisione dell'86.8%. Questa regola sottolinea l'importanza dei livelli contenuti di HNF1A e HLA-F insieme a un'elevata espressione di HLA-DQA2 per indicare un forte segnale per la classe positiva. Un'altra combinazione che si distingue per la copertura è `['IL1A <= 4.45', 'HLA-F <= 10.67', 'HLA-DQA2 > 4.99']`, con una copertura del 13.09% e una precisione del 90.2%, evidenziando che IL1A a livelli bassi, insieme a HLA-F e HLA-DQA2 in un determinato range, è un pattern caratteristico della classe positiva. Regole come `['CCR5 <= 7.72', 'IL2RA <= 6.12', 'HLA-DOB > 8.87']`, con una copertura del 12.25% e una precisione del 95.2%, sono invece predominanti per la classe negativa, mostrando che bassi livelli di CCR5 e IL2RA con HLA-DOB elevato sono predittivi per questa classe. Questo schema, ricor-

rente in diverse regole, sottolinea l'importanza di CCR5 e IL2RA nella classificazione negativa.

Random Forest. L'analisi delle regole della Random Forest evidenzia PTPRN2 come una delle feature più rilevanti per la classificazione. Molte regole che portano a una predizione negativa sono legate ad alti livelli di PTPRN2 in combinazione con altre feature come HLA-DOB e CYP27B1. Ad esempio, una regola associa PTPRN2 sopra 6.54 con livelli elevati di HLA-DOB e CYP27B1, ottenendo una precisione del 91.9% e una copertura del 2%, indicando che l'espressione elevata di questi geni è un forte indicatore per la classe negativa. Un'altra regola interessante per la classe negativa è `['PTPRN2 > 6.68', 'IL2RA <= 6.12']`, che ha una copertura del 7.8% con una precisione dell'84.8%, confermando che livelli elevati di PTPRN2 in presenza di IL2RA basso contribuiscono in modo significativo alla predizione negativa. Per quanto riguarda la classe positiva , il modello si basa su pattern con livelli inferiori di PTPRN2, associati a feature come HLA-DQA2 e PRF1. Ad esempio, la regola `['PTPRN2 <= 6.68', 'HLA-DQA2 > 5.06', 'PRF1 > 10.99']` presenta una precisione del 99.5% e una copertura del 2.6%, suggerendo che la ridotta espressione di PTPRN2, in combinazione con alti livelli di HLA-DQA2 e PRF1, è distintiva per la classe positiva. Un'altra regola con buona copertura è `['PTPRN2 <= 6.68', 'HLA-F <= 10.67', 'HLA-DQA2 > 5.06']`, che copre il 10.8% dei campioni con una precisione dell'89.4%. Questa combinazione evidenzia come bassi valori di PTPRN2, insieme a HLA-F e HLA-DQA2, siano marcatori potenti per i campioni positivi. Alcune regole riescono a coprire anche percentuali più ampie dei campioni. Ad esempio, una regola per la classe positiva, `['HLA-DOB <= 8.98', 'HLA-F <= 10.67', 'HLA-DQB1 <= 3.68']`, ottiene una copertura del 12.3% con una precisione dell'86.5%, indicando che valori bassi di HLA-DOB, HLA-F, e HLA-DQB1 sono fortemente associati alla classe positiva.

5.2 Explainability con SHAP.

In questa sezione, analizziamo l'interpretabilità dei modelli di machine learning utilizzando SHAP. Attraverso SHAP, è possibile generare diversi tipi di plot che

visualizzano il contributo delle feature, rendendo più intuitiva la comprensione delle variabili più influenti nel modello. Sono stati scelti approcci SHAP specifici per i vari modelli utilizzati:

- Kernel Explainer per l'SVM con kernel non lineare: questo approccio è particolarmente adatto ai modelli complessi, poiché permette di stimare i valori SHAP anche quando la relazione tra feature e output non è lineare.
- Tree Explainer per la Random Forest e XGBoost: ideale per modelli basati su strutture ad albero, sfrutta la natura del modello per calcolare i valori SHAP in modo efficiente e accurato.

Grazie a questi metodi, SHAP permette la creazione di summary plot, che offrono una panoramica visiva chiara dell'importanza di ciascuna feature. In questi grafici, ogni punto rappresenta il contributo di una feature a una singola predizione, con il colore che ne indica il valore: dall'azzurro per i valori bassi al rosso per quelli alti. Le feature sono ordinate sull'asse verticale per importanza decrescente, mentre sull'asse orizzontale è rappresentato il valore SHAP, che mostra quanto ciascuna feature influisce sulla predizione. In questo modo, il summary plot rende immediatamente visibili le variabili più determinanti nel processo di predizione, facilitando l'interpretazione dei risultati ottenuti.

XGBoost. Il summary plot riportato in Figura 5.1 evidenzia le caratteristiche più impattanti, in linea con quanto emerso dall'analisi delle regole di Anchor. Tra le feature di maggiore importanza si distinguono PTPRN2, IL2RA, TNF e CTSH. Dal grafico si osserva che elevati valori di PTPRN2 sono associati alla classe 0 ("Healthy"), mentre valori ridotti di questa feature sono indicativi della classe 1 ("Type 1 Diabetes"). Al contrario, IL2RA presenta un comportamento opposto: alti livelli di espressione sono correlati alla classe 1, mentre valori più bassi caratterizzano la classe 0. Per quanto riguarda TNF, si nota che valori elevati si associano alla classe 0, mentre valori bassi tendono a essere un marker per la classe 1. Infine, CTSH mostra una correlazione in cui bassi valori indicano la classe 0, mentre livelli elevati di espressione sono predittivi per la classe 1. Questi risultati sottolineano il ruolo differenziale di ciascuna feature

nella classificazione delle due classi e forniscono ulteriori indicazioni sull'importanza delle specifiche caratteristiche nell'output del modello.

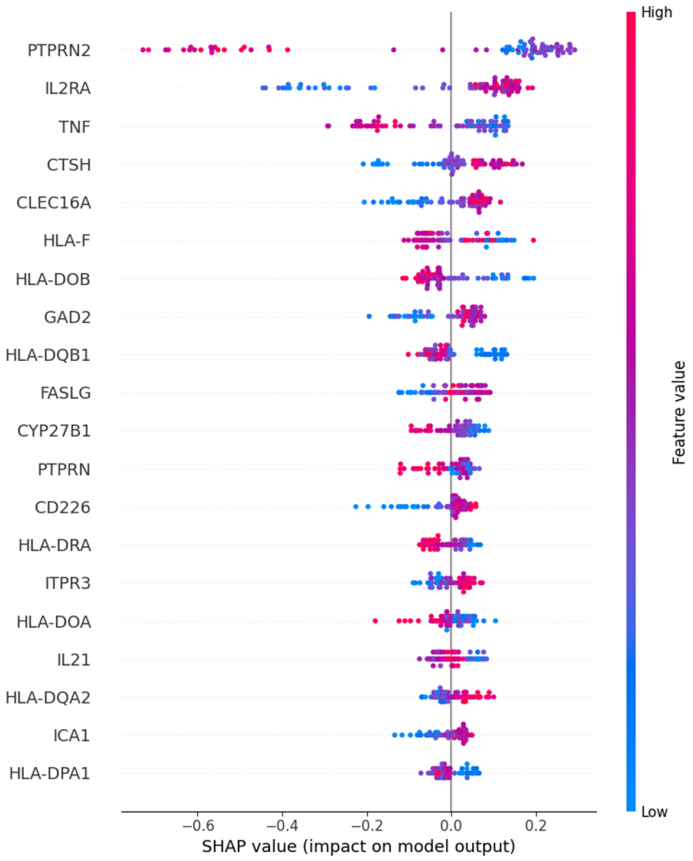

Figura 5.1: Summary plot per XGBoost.

SVM. Il summary plot in Figura 5.2 per il modello SVM evidenzia le feature chiave utilizzate per la classificazione. Si osserva una coerenza generale con i risultati ottenuti dalle regole di Anchor, con PTPRN2 che emerge come una delle principali feature discriminanti tra le due classi: valori bassi sono associati alla classe 0, mentre valori elevati sono indicativi della classe 1. Tuttavia, emerge una discrepanza rispetto alle regole di Anchor, in quanto HLA-DQB1 risulta essere la feature più discriminante a livello globale secondo l'SVM, un'informazione che l'approccio Anchor non è riuscito a identificare. Questo risultato sottolinea l'importanza di considerare metodi di interpretazione multipli per ottenere una visione più completa del comportamento del modello.

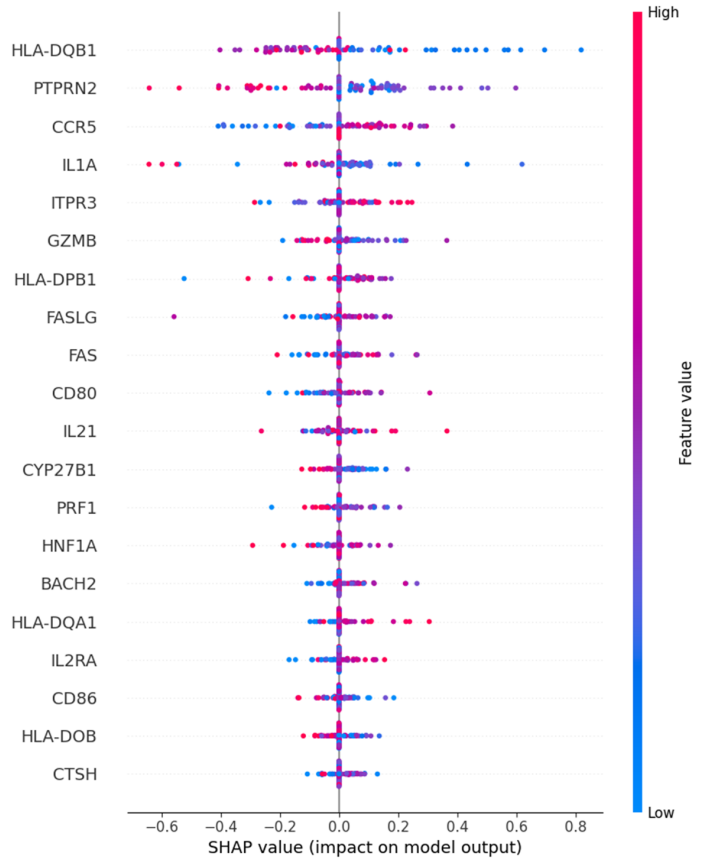

Figura 5.2: Summary plot per SVM.

Random Forest. Il summary plot in Figura 5.3 mostra le feature più rilevanti per la classificazione utilizzando Random Forest. Si conferma la predominanza di PTPRN2 tra le caratteristiche principali: valori bassi sono associati alla classe 1 ("Type 1 Diabetes"), mentre valori elevati tendono a indicare la classe 0 ("Healthy"). Inoltre, il grafico supporta le regole dedotte attraverso l'analisi di Anchor, come quella che associa livelli elevati di PTPRN2, HLA-DOB e CYP27B1 a una predizione della classe 0. Questi risultati sottolineano la coerenza tra i metodi di interpretazione e l'importanza di queste feature nella differenziazione tra le classi.

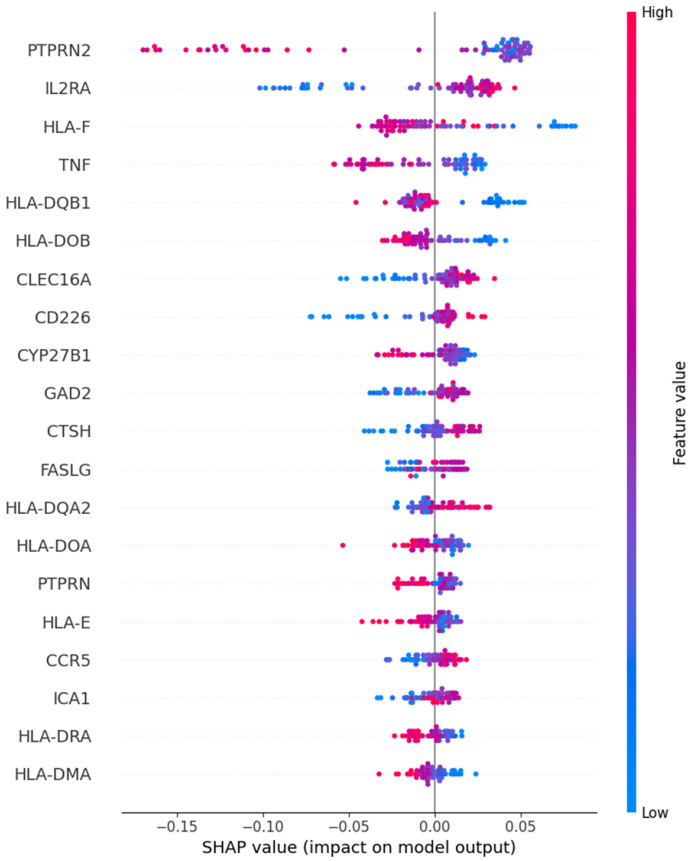

Figura 5.3: Summary plot per Random Forest.

I summary plot confermano molte delle intuizioni ottenute dalle regole di Anchor, mostrando l'importanza di feature come PTPRN2 nella classificazione delle due classi. La coerenza tra i due metodi sottolinea il ruolo chiave di queste variabili, fornendo una visione completa e dettagliata che arricchisce l'interpretabilità dei modelli utilizzati.

CAPITOLO 6

Conclusioni

Riepilogo dei risultati. Questa ricerca ha affrontato lo studio del diabete di tipo 1 (T1D) utilizzando un'analisi dell'espressioni geniche e approcci di machine learning. Partendo dalla comprensione del dominio biologico, sono stati selezionati geni chiave associati al T1D per creare un dataset completo. Attraverso l'applicazione di tecniche di preprocessing è stato possibile ottenere un dataset uniforme per l'analisi predittiva. Sono stati addestrati tre modelli: XGBoost, Support Vector Machine (SVM) e Random Forest, utilizzando parametri ottimizzati. Le prestazioni di questi modelli sono state valutate tramite metriche come accuracy, precision, recall, F1-score e AUC-ROC. I risultati mostrano che l'SVM ha raggiunto le migliori prestazioni, con un'accuracy dell'80.8%, precision dell'81.3%, e un recall pari al 80.8%, suggerendo un'alta capacità di bilanciare tra la riduzione dei falsi positivi e l'identificazione accurata dei positivi reali. La curva AUC-ROC conferma il primato dell'SVM con un valore di 87.6%, leggermente superiore a quelli di Random Forest (87.4%) e XGBoost (87.2%). Questi risultati indicano che l'SVM riesce a separare efficacemente le classi positive e negative, rendendolo particolarmente adatto per contesti in cui è essenziale ridurre al minimo sia i falsi negativi che i falsi positivi. XGBoost e Random Forest hanno mostrato prestazioni molto simili, con una lieve superiorità di XGBoost per quanto riguarda l'interpretabilità delle decisioni grazie all'uso del Tree Explainer per

l'analisi SHAP. L'explainability, implementata con tecniche come SHAP e Anchor, ha svolto un ruolo centrale, fornendo una comprensione dettagliata delle decisioni dei modelli e facilitando la selezione delle feature più rilevanti. I summary plot di SHAP hanno rivelato feature cruciali come PTPRN2 e HLA-DQB1, consentendo di identificare i geni che influenzano maggiormente la classificazione della classe positiva all'interno del dataset. In particolare, il modello XGBoost ha associato livelli elevati di PTPRN2 alla classe "Healthy", mentre bassi livelli erano indicativi della classe "Type 1 Diabetes".

Limitazioni. Il lavoro ha incontrato diverse limitazioni, principalmente legate alla disponibilità e qualità dei dati. La disponibilità di dati pubblici di espressione genica nell'età di insorgenza della malattia è altamente limitata. Sebbene fossero presenti alcuni dati longitudinali, questi erano pochi, costringendo a trattare i campioni come singoli pazienti. Questo aspetto ha ridotto la possibilità di analizzare in modo approfondito la progressione della malattia e di fare inferenze robuste su una scala temporale più ampia. Inoltre, l'uso di dati di espressione genica provenienti da due piattaforme diverse ha introdotto un batch effect significativo, ridotto dall'utilizzo delle tecniche.

Prospettive future. Un'importante estensione di questo lavoro potrebbe essere l'integrazione di dati temporali per sviluppare un'analisi di serie temporali, esaminando i cambiamenti di espressione genica nel tempo. Dati longitudinali più ampi potrebbero consentire di costruire modelli che tracciano la progressione del T1D, migliorando così la capacità di prevedere l'insorgenza e la gravità della malattia in relazione a specifiche variazioni geniche. Ciò potrebbe anche facilitare lo sviluppo di modelli predittivi più accurati e personalizzati, che tengano conto delle variazioni dinamiche delle espressioni geniche tra gli individui.

Bibliografia

[1] E. Kawasaki, "Anti-islet autoantibodies in type 1 diabetes," *International Journal of Molecular Sciences*, vol. 24, no. 12, 2023. [Online]. Available: https://www.mdpi.com/1422-0067/24/12/10012 (Citato a pagina 5)

[2] A. Zajec, K. Trebušak Podkrajšek, T. Tesovnik, R. Šket, B. Čugalj Kern, B. Jenko Bizjan, D. Šmigoc Schweiger, T. Battelino, and J. Kovač, "Pathogenesis of type 1 diabetes: Established facts and new insights," *Genes (Basel)*, vol. 13, no. 4, p. 706, Apr. 2022. (Citato alle pagine 6, 13 e 17)

[3] L. Yu, M. Rewers, R. Gianani, E. Kawasaki, Y. Zhang, C. Verge, P. Chase, G. Klingensmith, H. Erlich, J. Norris, and G. S. Eisenbarth, "Antiislet autoantibodies usually develop sequentially rather than simultaneously," *J. Clin. Endocrinol. Metab.*, vol. 81, no. 12, pp. 4264–4267, Dec. 1996. (Citato a pagina 6)

[4] Machiarelli, Arcucci, Bianchi, Cappello, Castaldo, Continenza, David, DiMeglio, Guerra, Montagnani, Nottola, Nurzynska, Palmierini, Raspanti, and Spera, *Anatomia per lauree triennali e magistrali*. Italia: Idelson-Gnocchi, 2019. (Citato a pagina 7)

[5] P. J. Delves, "Componenti molecolari del sistema immunitario," 2024/02 2024. [Online]. Available: https://www.msdmanuals.com/it-it/professionale/

immunologia-malattie-allergiche/biologia-del-sistema-immunitario/componenti-molecolari-del-sistema-immunitario (Citato a pagina 11)

[6] A. K. Steck and M. J. Rewers, "Genetics of Type 1 Diabetes," *Clinical Chemistry*, vol. 57, no. 2, pp. 176–185, 02 2011. [Online]. Available: https://doi.org/10.1373/clinchem.2010.148221 (Citato a pagina 14)

[7] A. D. Association, "Genetics of diabetes," Available at https://diabetes.org/about-diabetes/genetics-diabetes. (Citato a pagina 14)

[8] B. Coico, Sunshine, *Immunologia*. Edises, 2005. (Citato a pagina 15)

[9] J. A. Noble and A. M. Valdes, "Genetics of the HLA region in the prediction of type 1 diabetes," *Curr. Diab. Rep.*, vol. 11, no. 6, pp. 533–542, Dec. 2011. (Citato a pagina 15)

[10] M. J. Redondo, A. K. Steck, and A. Pugliese, "Genetics of type 1 diabetes," *Pediatric Diabetes*, vol. 19, no. 3, pp. 346–353, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/pedi.12597 (Citato alle pagine 15 e 16)

[11] I. Minniakhmetov, B. Yalaev, R. Khusainova, E. Bondarenko, G. Melnichenko, I. Dedov, and N. Mokrysheva, "Genetic and epigenetic aspects of type 1 diabetes mellitus: Modern view on the problem," *Biomedicines*, vol. 12, no. 2, 2024. [Online]. Available: https://www.mdpi.com/2227-9059/12/2/399 (Citato a pagina 15)

[12] M. J. Redondo, A. K. Steck, and A. Pugliese, "Genetics of type 1 diabetes," *Pediatr. Diabetes*, vol. 19, no. 3, pp. 346–353, May 2018. (Citato a pagina 15)

[13] A. K. Steck and M. J. Rewers, "Genetics of Type 1 Diabetes," *Clinical Chemistry*, vol. 57, no. 2, pp. 176–185, 02 2011. [Online]. Available: https://doi.org/10.1373/clinchem.2010.148221 (Citato alle pagine 15, 16 e 17)

[14] S. E. Regnell and Å. Lernmark, "Early prediction of autoimmune (type 1) diabetes," *Diabetologia*, vol. 60, no. 8, pp. 1370–1381, Aug. 2017. (Citato a pagina 16)

[15] F. Pociot, B. Akolkar, P. Concannon, H. A. Erlich, C. Julier, G. Morahan, C. R. Nierras, J. A. Todd, S. S. Rich, and J. Nerup, "Genetics of type 1 diabetes: what's next?" *Diabetes*, vol. 59, no. 7, pp. 1561–1571, Jul. 2010. (Citato a pagina 16)

[16] S. A. Paschou, N. Papadopoulou-Marketou, G. P. Chrousos, and C. Kanaka-Gantenbein, "On type 1 diabetes mellitus pathogenesis," *Endocr. Connect.*, vol. 7, no. 1, pp. R38–R46, Jan. 2018. (Citato a pagina 16)

[17] R. Mittal, N. Camick, J. R. N. Lemos, and K. Hirani, "Gene-environment interaction in the pathophysiology of type 1 diabetes," *Frontiers in Endocrinology*, vol. 15, 2024. [Online]. Available: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2024.1335435 (Citato alle pagine 16 e 17)

[18] A. K. Steck, F. Dong, R. Wong, A. Fouts, E. Liu, J. Romanos, C. Wijmenga, J. M. Norris, and M. J. Rewers, "Improving prediction of type 1 diabetes by testing non-hla genetic variants in addition to hla markers," *Pediatric Diabetes*, vol. 15, no. 5, pp. 355–362, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/pedi.12092 (Citato a pagina 16)

[19] S. E. Regnell and Å. Lernmark, "Early prediction of autoimmune (type 1) diabetes," *Diabetologia*, vol. 60, no. 8, pp. 1370–1381, Aug. 2017. (Citato a pagina 16)

[20] M. J. Dufort, C. J. Greenbaum, C. Speake, and P. S. Linsley, "Cell type-specific immune phenotypes predict loss of insulin secretion in new-onset type 1 diabetes," *JCI Insight*, vol. 4, no. 4, Feb. 2019. (Citato a pagina 17)

[21] A. M. Giwa, R. Ahmed, Z. Omidian, N. Majety, K. E. Karakus, S. M. Omer, T. Donner, and A. R. A. Hamad, "Current understandings of the pathogenesis of type 1 diabetes: Genetics to environment," *World J. Diabetes*, vol. 11, no. 1, pp. 13–25, Jan. 2020. (Citato a pagina 17)

[22] N. AlRefaai and S. Z. AlRashid, "Classification of gene expression dataset for type 1 diabetes using machine learning methods," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 5, pp. 2986–2992, 2023. [Online]. Available: https://beei.org/index.php/EEI/article/view/4322 (Citato a pagina 18)

[23] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, 2018. [Online]. Available: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2018.00515 (Citato a pagina 18)

[24] L. J. Muhammad, E. A. Algehyne, and S. S. Usman, "Predictive supervised machine learning models for diabetes mellitus," *SN Comput. Sci.*, vol. 1, no. 5, p. 240, Jul. 2020. (Citato a pagina 18)

[25] A. R. Patil, J. Schug, C. Liu, D. Lahori, H. C. Descamps, Human Pancreas Analysis Consortium, A. Naji, K. H. Kaestner, R. B. Faryabi, and G. Vahedi, "Modeling type 1 diabetes progression using machine learning and single-cell transcriptomic measurements in human islets," *Cell Rep. Med.*, vol. 5, no. 5, p. 101535, May 2024. (Citato a pagina 19)

[26] G. E. Omnibus, "Gene expression omnibus," https://www.ncbi.nlm.nih.gov/geo/. (Citato a pagina 21)

[27] J. A. Miller, C. Cai, P. Langfelder, D. H. Geschwind, S. M. Kurian, D. R. Salomon, and S. Horvath, "Strategies for aggregating gene expression data: the collapserows R function," *BMC Bioinformatics*, vol. 12, no. 1, p. 322, Aug. 2011. (Citato a pagina 23)

[28] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000. (Citato a pagina 27)

[29] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Sci.*, vol. 28, no. 11, pp. 1947–1951, Nov. 2019. (Citato a pagina 27)

[30] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe, "KEGG for taxonomy-based analysis of pathways and genomes," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D587–D592, Jan. 2023. (Citato a pagina 27)

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*,

vol. 32, no. 1, Apr. 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11491 (Citato a pagina 44)

[32] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: https://arxiv.org/abs/1705.07874 (Citato a pagina 44)

Ringraziamenti

Ringraziamenti qui...