



UNIVERSITA' POLITECNICA DELLE MARCHE
FACOLTA' DI INGEGNERIA

Corso di Laurea triennale in Ingegneria Biomedica

Sviluppo di un metodo di machine learning per la selezione di metriche da monitoraggio continuo del glucosio in grado di discriminare la presenza di ipotiroidismo in individui con diabete di tipo 1

Development of a machine learning method for selecting continuous glucose monitoring metrics capable of discriminating the presence of hypothyroidism in individuals with type 1 diabetes

Relatore:

Dott. Micaela Morettini

Tesi di Laurea di:

Agnese Callegari

Correlatore:

Dott. Piersanti Agnese

Dott. Tura Andrea

A.A. 2022/ 2023

Indice

Introduzione	pg. V
Capitolo 1: Accenni di fisiologia	
1.1) <i>Metabolismo del glucosio</i>	pg.VII
1.1.1) <i>Meccanismi di regolazione e insulina</i>	pg X
1.2) <i>Il diabete</i>	pg XII
1.2.1) <i>Complicanze associate</i>	pg XVI
1.2.2) <i>Ipotiroidismo e Diabete Tipo 1</i>	pg XVII
Capitolo 2: Terapia e monitoraggio del diabete	
2.1) <i>Panoramica strumenti utilizzati</i>	pg XX
2.1.1) <i>Monitoraggio Continuo del Glucosio</i>	pg XXIII
Capitolo 3: Sistemi di supporto alla decisione clinica	
3.1) <i>Machine learning</i>	pg XXIX
Capitolo 4: Sviluppo di un metodo di machine learning per la selezione di metriche da monitoraggio continuo del glucosio discriminanti la presenza di ipotiroidismo	
4.1) <i>Acquisizione dati</i>	
4.1.1) <i>Campagna Replace BG</i>	pg XXXIV
4.1.2) <i>Creazione del dataset di studio a partire dal dataset di partenza</i>	pg XXXV
4.2) <i>Elaborazione dati</i>	
4.2.1) <i>Calcolo delle metriche</i>	pg XXXVII
4.2.2) <i>Scelta delle metriche</i>	pg XL
4.2.3) <i>Creazione del modello per classificazione</i>	pg XLII
4.3) <i>Risultati e Discussioni</i>	pg XLVII
Conclusione	pg LVIII
Bibliografia	pg LX

Abstract

Il diabete mellito e le malattie croniche della tiroide sono tra le malattie endocrine più diffuse al mondo e spesso sono presenti come morbosità nello stesso soggetto. In particolar modo, l'ipotiroidismo è la forma più comune di disfunzione tiroidea nei pazienti diabetici, con una prevalenza del 5.7%. Recenti studi hanno inoltre dimostrato che il diabete mellito di tipo 1 è uno dei fattori di rischio per lo sviluppo dell'ipotiroidismo (IPTD).

Questo studio è stato effettuato con l'obiettivo di portare allo sviluppo di un modello in grado di classificare correttamente, partendo dai tracciati di glicemia dei pazienti con Diabete Mellito di Tipo 1 (DMT1), i soggetti con ipotiroidismo.

Per fare ciò sono stati sfruttati gli algoritmi di machine learning supervisionato, a cui sono stati forniti in ingresso i dati relativi a 79 soggetti con DMT1 prelevati dal database della campagna REPLACE_BG. Di questi 79 soggetti, 28 presentavano solo DMT1 e 51 invece anche ipotiroidismo. Nello specifico, i dati utilizzati per creare il modello erano costituiti dai tracciati glicemici acquisiti ogni cinque minuti da Monitoraggio in Continuo del Glucosio (CGM) Dexcom durante un monitoraggio continuo di 26 settimane.

Per ottimizzare il funzionamento del machine learning i tracciati glicemici sono stati precedentemente pre-elaborati e da essi sono state selezionate 41 features grazie all'utilizzo delle funzioni apposite fornite dal software AGATA.

Sfruttando poi il toolkit ORANGE, sono state ulteriormente selezionate, attraverso l'applicazione dell'Albero Decisionale, le otto features determinate fondamentali per la definizione del modello: `sdwIndex`, `aucGlucose`, `jIndex`, `mageIndex`, `igc`, `cogi`, `conga` e `stdGlucoseROC`. Tutte queste features, fatta eccezione per `cogi`, descrivono la variabilità del tracciato glicemico dei soggetti: questo risultato suggerisce quindi un possibile collegamento con la patologia dell'ipotiroidismo.

Grazie all'utilizzo di un approccio al Machine Learning quale l'Albero Decisionale si è creato un modello moderatamente accurato in grado di svolgere quanto richiesto, tale modello mostra prestazioni notevolmente migliori rispetto a quelli ottenuti utilizzando altre tecniche come Random Forest, Adaboost, Gradient Boosting, Naive Bayes e Logistic Regression.

In particolar modo il modello con Albero Decisionale mostra valori di AUC pari a 73.3%, CA 0,736, una precisione di 0,818, sensibilità pari a 0,766 e specificità invece 0,68: il modello, quindi,

individua in modo abbastanza accurato i soggetti con ipotiroidismo ma risulta più carente invece nella classificazione dei soggetti con solo DMT1.

In conclusione, quindi, è stato creato un modello moderatamente accurato per l'individuazione tra i soggetti con DMT1 dei pazienti con ipotiroidismo che può essere considerato una buona base per studi futuri.

Introduzione

Nel corso degli scorsi anni, il veloce sviluppo delle nuove tecnologie ha permesso le creazioni di precisi e innovativi strumenti che svolgono il ruolo di potenti ausili per i medici nel loro processo decisionale e che possono anche essere di supporto per i pazienti durante il trattamento delle varie patologie.

Un ambito particolarmente fecondo a tale riguardo è quello della diabetologia. Il diabete è una patologia ad alta incidenza che porta con sé importanti cambiamenti nello stile di vita dell'individuo affetto: sebbene infatti la mortalità associata a tale malattia negli anni sia notevolmente diminuita, è di vitale importanza che il soggetto sia fedele alla terapia affinché questa abbia effetto ed il monitoraggio dell'evolvere della patologia assume quindi notevole rilevanza.

In particolar modo, il diabete mellito è una malattia cronica caratterizzata da un eccesso di glucosio nel sangue, condizione nota come iperglicemia. Si divide in due forme principali: il diabete di tipo 1 e il diabete di tipo 2. Il Diabete Mellito di Tipo 1 (DMT1) colpisce i soggetti fin dalla nascita ed è una forma cronica di assenza di secrezione insulinica: la terapia, quindi, richiede una precisa somministrazione di insulina giornaliera e un frequente controllo dei valori di insulina al fine di evitare eventi pericolosi per la salute come crisi iperglicemiche o ipoglicemiche.

Per svolgere tutti questi, e anche tanti altri, compiti le nuove tecnologie sono di grande aiuto permettendo, ad esempio, la creazione di strumenti come il Continuous Glucose Monitoring (CGM), dispositivo che consente un monitoraggio continuo della glicemia nei liquidi interstiziali, con una misura ogni 1-5 minuti. Le numerose acquisizioni rese possibili dal CGM sono importanti risorse sia per il personale sanitario, poiché permettono una ricostruzione precisa della curva giornaliera ipoglicemica, sia per chi svolge attività di ricerca, fornendo una sostanziosa banca di dati a cui avere accesso per svolgere specifici studi scientifici.

L'elaborato che sarà presentato nelle successive pagine sfrutta infatti proprio i dati acquisiti tramite CGM dalla campagna scientifica REPLACE BG per poi rielaborarli e trarre da essi le informazioni necessarie alla creazione di un Sistema di Supporto alla Decisione Clinica (CDSS) in grado di riconoscere e classificare correttamente i soggetti che oltre ad essere affetti a DMT1 presentano anche l'ipotiroidismo.

Sebbene infatti, con gli anni la patologia del DMT1 sia stata ampiamente studiata e si siano ormai individuate le più adatte terapie, uno dei principali rischi medici a cui il soggetto diabetico va incontro è tutt'ora rappresentato dalle molte patologie a cui i soggetti con DMT1 sono suscettibili.

Questo studio si concentra su una, molto incidente e forse non altrettanto studiata, complicanza del DMT1, ovvero l'ipotiroidismo. L'ipotiroidismo è una condizione di ipo-attività della tiroide che determina una ridotta produzione di ormoni tiroidei e un conseguente rallentamento delle funzioni vitali. Diversi studi prospettici a lungo termine hanno confermato che la prevalenza dell'ipotiroidismo nei pazienti con diabete di tipo 1 è più elevata rispetto a quella nella popolazione generale, ciò suggerisce che il diabete di tipo 1 favorisce lo sviluppo dell'ipotiroidismo.

Uno degli obiettivi dell'elaborato è quindi individuare, grazie all'utilizzo delle nuove tecniche di machine learning, le specifiche features dei soggetti con diabete di tipo 1 che rappresentano un fattore di rischio per l'ipotiroidismo. Il modello creato vuole quindi essere uno strumento capace di permettere ai soggetti a rischio di sottoporsi preventivamente a monitoraggi della tiroide così da prevenire o limitare lo sviluppo successivo dell'ipotiroidismo.

Capitolo 1: Accenni di fisiologia

1.1) Metabolismo del glucosio

Per comprendere a pieno l'importanza dello studio svolto è necessario prima indagare in che modo il corpo umano riesce a fornire continuamente energia ai vari tessuti e organi così da garantirne il continuo funzionamento ottimale nonostante i periodi di digiuno tra un pasto e l'altro. Quanto appena illustrato è reso possibile grazie al modo in cui le vie metaboliche dei carboidrati, dei lipidi e delle proteine vengono integrate e controllate durante l'arco della giornata e sarà questo, infatti, l'argomento del capitolo a seguire.

Le due fasi funzionali nelle quali l'organismo fornisce energia per le attività cellulari sono la fase di assorbimento, durante la quale i nutrienti ingeriti entrano nel sangue attraverso il tratto gastrointestinale e la fase di post-assorbimento, durante la quale il tratto gastrointestinale non contiene sostanze nutritive e l'energia deriva dalle riserve proprie dell'organismo. La fase di post-assorbimento quindi si verifica in corrispondenza della tarda mattinata, del tardo pomeriggio e della maggior parte della notte [1].

Durante la fase di assorbimento alcune delle sostanze nutritive ingerite forniscono il fabbisogno energetico immediato per l'organismo mentre la restante parte viene aggiunta alle riserve di energia dell'organismo per essere poi utilizzata durante la successiva fase di post-assorbimento. Le riserve di energia dell'intero organismo sono sufficienti a una persona media per tollerare un digiuno di molte settimane, a condizione che l'acqua sia disponibile. Un pasto tipico contiene tutte le principali categorie alimentari che forniscono energia come carboidrati, lipidi e proteine ma tra questi è importante ricordare che i carboidrati costituiscono la maggior parte del contenuto energetico, quindi calorie, del pasto per cui successivamente l'attenzione verrà concentrata prevalentemente sull'assorbimento di questi.

In generale i carboidrati e le proteine sono assorbiti nel sangue che esce dal tratto gastrointestinale principalmente come monosaccaridi e aminoacidi rispettivamente, i lipidi invece vengono assorbiti nella linfa come chilomicroni. La linfa giunge poi direttamente al sistema venoso sistemico.

Entrando nel dettaglio alcuni dei carboidrati assorbiti dal tratto gastrointestinale sono il galattosio e il fruttosio, i quali sono convertiti in glucosio dal fegato: il glucosio è quindi la fonte principale di energia dell'organismo durante la fase di assorbimento. Gran parte del glucosio assorbito entra

nelle cellule e viene catabolizzato in anidride carbonica e acqua in un processo che libera energia la quale viene utilizzata per la formazione di ATP.

Il muscolo scheletrico costituisce la maggior parte della massa corporea; pertanto, è il principale utilizzatore di glucosio, anche quando si trova a riposo. Il muscolo scheletrico non solo catabolizza il glucosio durante la fase di assorbimento ma converte anche parte del glucosio nel polisaccaride glicogeno, che viene poi immagazzinato nelle cellule muscolari per un utilizzo futuro. Anche le cellule del tessuto adiposo, dette adipociti, catabolizzano il glucosio per produrre energia, ma soprattutto lo trasformano in grasso sotto forma di trigliceridi. Nel fegato si verifica un assorbimento netto di glucosio, che viene immagazzinato in forma di glicogeno come accade nel muscolo scheletrico, o viene trasformato in glicerolo-3-fosfato e acidi grassi utilizzati poi per la sintesi dei trigliceridi, come accade nel tessuto adiposo. La maggior parte dei lipidi sintetizzati nel fegato a partire dal glucosio viene assemblata, insieme a proteine specifiche, in aggregati molecolari di lipidi e proteine, le lipoproteine, che sono poi secrete dagli epatociti ed entrano nel sangue.

Quando termina la fase di assorbimento, la sintesi netta di glicogeno, trigliceridi e proteine cessa e inizia il catabolismo netto di tutte queste sostanze per questo motivo tale fase viene anche definita catabolica mentre quella precedente anabolizzante. Le sostanze catabolizzate da questa seconda fase costituiranno, quindi, la principale fonte di energia dell'organismo.

Il problema fondamentale della fase di post-assorbimento è che nel tratto gastrointestinale non viene assorbito glucosio; tuttavia, la concentrazione di glucosio plasmatico deve essere mantenuta in maniera omeostatica, poiché generalmente il sistema nervoso centrale utilizza solo glucosio come energia. Se la concentrazione plasmatica di glucosio diminuisce troppo, si possono verificare alterazioni dell'attività nervosa, che vanno da una lieve compromissione delle funzioni mentali a convulsioni, coma e anche morte. Sono quindi necessarie ulteriori fonti di glucosio plasmatico capaci di agire anche in condizione di digiuno.

Prima fra tutte vi è la glicogenolisi ovvero l'idrolisi delle riserve di glicogeno a monomeri di glucosio 6-fosfato che avviene nel fegato ed inizia pochi secondi dopo uno stimolo adeguato, come l'attivazione del sistema nervoso simpatico durante il digiuno. Tuttavia, la quantità di glucosio resa disponibile da questa fonte è in grado di far fronte alle necessità dell'organismo solo per alcune ore, prima che il glicogeno epatico si esaurisca quasi completamente.

Una seconda importante fonte di glucosio viene dal catabolismo dei trigliceridi nel tessuto adiposo che produce glicerolo e acidi grassi, tale processo è chiamato lipolisi. Il glicerolo e gli acidi grassi

entrano poi nel sangue per diffusione e il glicerolo che giunge al fegato viene utilizzato per la sintesi di glucosio: inoltre dopo poche ore dall'inizio della fase di post-assorbimento, le proteine diventano un'altra fonte di glucosio plasmatico. Elevate quantità di proteine sia nei muscoli sia in altri tessuti possono essere catabolizzate senza generare gravi malfunzionamenti a livello cellulare. Naturalmente ci sono dei limiti a tale processo; la continua perdita di proteine, che si verifica durante un digiuno prolungato, può portare, infine, a grave compromissione funzionale delle cellule, malessere e morte.

La sintesi di glucosio da uno di questi precursori appena elencati è conosciuta come gluconeogenesi. Durante un digiuno di 24 ore, la gluconeogenesi fornisce circa 180 g di glucosio, pertanto, non può fornire all'organismo tutta l'energia di cui ha bisogno durante il digiuno. Durante la transizione tra la fase di assorbimento e quella di post-assorbimento deve quindi verificarsi un processo di adattamento. La maggior parte degli organi e dei tessuti, a esclusione del sistema nervoso, riduce significativamente il catabolismo del glucosio e aumenta l'utilizzo dei lipidi, in modo tale che essi diventino la principale fonte di energia. Questo adattamento metabolico, chiamato risparmio di glucosio, conserva il glucosio prodotto dal fegato affinché esso venga utilizzato dal sistema nervoso.

Il passaggio essenziale di questo adattamento è la lipolisi la quale porta anche alla liberazione degli acidi grassi che circolano e sono captati e metabolizzati da quasi tutti i tessuti, a esclusione del sistema nervoso. Essi forniscono energia nel fegato venendo trasformati in composti chiamati chetoni o corpi chetonici. Questi sono rilasciati nel sangue e durante un digiuno prolungato costituiscono una importante fonte energetica per diversi tessuti, incluso quello nervoso.

Il risultato netto dell'utilizzo di acidi grassi e chetoni durante il digiuno è la produzione di energia per l'organismo e allo stesso tempo il risparmio del glucosio in favore del cervello e del sistema nervoso. Inoltre, come appena sottolineato, il cervello può usare i chetoni come fonte energetica e tale utilizzo aumenta quando i chetoni si accumulano nel sangue durante i primi giorni di digiuno. Tale fenomeno è di notevole importanza per la sopravvivenza; quando la richiesta di glucosio da parte del cervello diminuisce, poiché esso utilizza i chetoni come fonte energetica, la degradazione di proteine necessaria per fornire aminoacidi per la gluconeogenesi sarà molto minore. Di conseguenza, la capacità dell'organismo di resistere a un lungo digiuno, senza che i tessuti subiscano gravi danni, aumenta.

1.1.1) Meccanismi di regolazione e insulina

Concentrando ora l'attenzione sui fattori endocrini e nervosi che controllano e integrano le precedentemente descritte vie metaboliche i più importanti sono senza dubbio due ormoni pancreatici: l'insulina e il glucagone.

L'insulina e il glucagone sono ormoni polipeptidici secreti dalle isole di Langerhans, dette anche isole pancreatiche, ammassi di cellule endocrine nel pancreas. Esistono diversi tipi distinti di cellule delle isole pancreatiche, ognuna delle quali secerne un ormone diverso: le cellule beta infatti sono la fonte di insulina e le cellule alfa producono invece il glucagone [1].

Indagando più approfonditamente il primo ormone fondamentale, l'insulina, questa viene prodotta inizialmente sotto la forma inattiva di proinsulina a livello del reticolo plasmatico e viene poi attivata a livello dell'apparato di Golgi diventando insulina. La secrezione dell' insulina, e quindi la sua concentrazione plasmatica, aumenta durante la fase di assorbimento e diminuisce durante la fase di post-assorbimento questo perché un aumento della concentrazione plasmatica di insulina è la causa principale degli eventi della fase di assorbimento, mentre una ridotta concentrazione plasmatica di insulina è la principale causa degli eventi della fase di post-assorbimento.

Gli effetti metabolici dell'insulina sono principalmente esercitati sulle cellule muscolari, sugli adipociti e sugli epatociti. Nelle cellule muscolari scheletriche e negli adipociti, ad esempio, un aumento della concentrazione di insulina stimola le vescicole citoplasmatiche, che contengono nella loro membrana un particolare tipo di trasportatore del glucosio (GLUT-4), a fondersi con la membrana plasmatica. L'aumentato numero dei trasportatori del glucosio sulla membrana plasmatica, che si osserva in seguito a tale fusione, determina poi un aumento della velocità di movimento del glucosio dal liquido extracellulare all'interno delle cellule, tramite diffusione facilitata.

È importante sottolineare che le cellule del cervello presentano un diverso sottotipo di GLUT che ha un'elevata affinità per il glucosio e la cui attività non è insulino-dipendente. Questo garantisce che, anche se la concentrazione di insulina plasmatica è molto bassa, come in caso di digiuno prolungato, le cellule del cervello possono continuare a captare il glucosio dal sangue e mantenere così la loro funzione.

Considerando ancora una volta gli effetti dell'insulina sulle cellule muscolari scheletriche complessivamente in queste cellule, l'insulina favorisce la formazione e l'accumulo del glicogeno aumentando il trasporto del glucosio nella cellula, stimolando l'enzima glicogeno sintasi che catalizza la reazione limitante nella sintesi del glicogeno e inibendo l'enzima glicogeno fosforilasi

che catalizza il catabolismo del glicogeno. Pertanto, nei muscoli scheletrici l'insulina favorisce la trasformazione e l'accumulo del glucosio in glicogeno attraverso i tre meccanismi sopra elencati.

La secrezione di insulina è principalmente, ma non limitatamente, controllata dalla concentrazione del glucosio plasmatico che è detta glicemia. Un aumento della glicemia, come avviene dopo un pasto che contiene carboidrati, agisce sulle cellule beta delle isole di Langerhans stimolando la secrezione di insulina mentre una diminuzione del glucosio elimina lo stimolo per la secrezione di insulina.

Si viene quindi a creare un processo omeostatico regolato da un meccanismo a feedback negativo: dopo un pasto, l'aumento della concentrazione plasmatica di glucosio stimola la secrezione di insulina, la quale stimola l'ingresso del glucosio nel tessuto muscolare e nel tessuto adiposo e l'assorbimento netto del glucosio, invece della produzione netta da parte del fegato. Questi effetti riducono successivamente la concentrazione di glucosio plasmatico fino a valori simili a quelli precedenti il pasto, eliminando così lo stimolo per la secrezione di insulina in modo tale che essa ritorni ai livelli precedenti.

Oltre alla concentrazione del glucosio plasmatico, molti altri fattori controllano la secrezione di insulina, come per esempio, una famiglia di ormoni noti come incretine: le azioni delle incretine costituiscono un sistema di controllo a feedforward per la regolazione del glucosio durante l'ingestione di un pasto.

Questo meccanismo minimizza la possibilità di avere grossi aumenti di glucosio nel plasma dopo un pasto, condizione che potrebbe eccedere la capacità dei reni di riassorbire completamente tutto il glucosio che compare nel filtrato dei nefroni. Un analogo delle incretine viene attualmente utilizzato per il trattamento del diabete mellito di tipo 2, patologia in cui il pancreas spesso produce una quantità di insulina insufficiente e le cellule dell'organismo sono meno responsive all'insulina.

Infine, anche i neuroni del sistema autonomo diretti alle isole di Langerhans influenzano la secrezione di insulina: l'attivazione dei neuroni parasimpatici, che si verifica durante l'ingestione di un pasto, stimola la secrezione dell'insulina e rappresenta un secondo tipo di regolazione a feedforward. Al contrario, la secrezione è inibita dall'attivazione dei neuroni simpatici diretti alle isole pancreatiche o da un aumento della concentrazione plasmatica di adrenalina.

Secondo ormone fondamentale per il controllo del metabolismo è il glucagone, l'ormone polipeptidico prodotto dalle cellule alfa delle isole pancreatiche. I principali effetti fisiologici del glucagone avvengono a livello del fegato e si oppongono a quelli dell'insulina: stimola la glicogenolisi, stimola la gluconeogenesi e stimola la sintesi dei chetoni. I risultati complessivi sono

quelli di aumentare le concentrazioni plasmatiche di glucosio e di chetoni, importanti per la fase di post-assorbimento e per prevenire l'ipoglicemia. Il principale stimolo per la secrezione di glucagone è costituito dalla diminuzione della concentrazione di glucosio circolante che, a sua volta, causa una riduzione dell'insulina plasmatica. L'insulina e il glucagone lavorano quindi insieme con azioni opposte per garantire nell'organismo l'omeostasi e il mantenimento entro range fisiologici dei valori della glicemia.

1.2) Il diabete

Quanto descritto precedentemente rappresenta il normale funzionamento del metabolismo di glucosio nei soggetti sani che sono quindi capaci di mantenere i valori della glicemia entro un range che va dai 70 mg/dL ai 100 mg/dL a digiuno. Quando i soggetti presentano valori di glicemia cronicamente fuori da questo intervallo, e in particolar modo mostrano valori di glicemia plasmatica a digiuno superiori a 126 mg/dL o glicemia maggiore di 200mg/dl due ore dopo carico orale di glucosio 75g, viene loro diagnosticata la malattia autoimmune diabete [2].

Con la parola diabete comunemente si intende il diabete mellito, ben diverso dal diabete insipido e quello renale, ed è una malattia cronica caratterizzata dalla presenza di elevati livelli di glucosio nel sangue e dovuta a un'alterata quantità o funzione dell'insulina.

I soggetti affetti dal diabete mellito sono quindi individui che, non curati, presentano normalmente valori molto alti di glicemia: tale condizione è detta iperglicemia, se invece i pazienti sono curati con terapie insuliniche non adeguate può accadere, al contrario, che la glicemia a digiuno scende sotto ai 60mg/dL e si parla allora di una condizione piuttosto grave chiamata ipoglicemia.

Il diabete mellito si distingue in due malattie differenti a seconda della causa: nel Diabete Mellito di Tipo 1 (DMT1), comunemente chiamato mellito insulino-dipendente o diabete giovanile, l'insulina è completamente assente o quasi dalle isole di Langerhans e dal plasma, di conseguenza la terapia con somministrazione di insulina risulta fondamentale e imprescindibile.

Nel Diabete Mellito di Tipo 2 (DMT2), comunemente chiamato diabete mellito non insulino-dipendente o diabete mellito a insorgenza in età adulta, l'insulina è ancora presente nel plasma ma la sensibilità cellulare all'insulina è inferiore alla norma poiché le cellule bersaglio presentano una certa insulino-resistenza. In molti pazienti con DMT2 la risposta delle cellule beta pancreatiche è anch'essa compromessa quindi la terapia deve essere costituita da un insieme di farmaci che aumentino la sensibilità cellulare all'insulina e la secrezione dalle stesse cellule beta o che, in altro

modo, riducano la produzione epatica di glucosio. Nei casi più gravi la terapia può richiedere anche la somministrazione stessa di insulina.

Caso a parte è rappresentato dal diabete gestazionale che presenta le stesse caratteristiche del DMT2 ma che non è cronico in quanto si sviluppa solo durante la gravidanza e generalmente tende a scomparire al termine di questa. Tuttavia, le donne che ne hanno sofferto presentano un rischio più elevato di sviluppare DMT2 in età avanzata [3]. Pur essendo una condizione transitoria, se non viene diagnosticato ed adeguatamente curato, può portare a conseguenze, anche gravi, sia per la madre che per il bambino.

Il DMT1 riguarda circa il 10% delle persone con diabete e in genere insorge nell'infanzia o nell'adolescenza. Nel diabete tipo 1, il pancreas non produce insulina a causa della distruzione delle cellule β che producono questo ormone. La velocità di distruzione delle cellule β è, comunque, piuttosto variabile, per cui l'insorgenza della malattia può avvenire rapidamente in alcune persone, solitamente nei bambini e negli adolescenti, e più lentamente negli adulti (in questi rari casi si parla di una forma particolare, detta LADA: Late Autoimmune Diabetes in Adults) [4].

La causa del DMT1 è sconosciuta, ma caratteristica è la presenza nel sangue di anticorpi diretti contro antigeni presenti a livello delle cellule che producono insulina, detti ICA, GAD, IA-2, IA-2 β . Questo danno, che il sistema immunitario induce nei confronti delle cellule che producono insulina, potrebbe essere legato a fattori ambientali (tra i quali, sono stati chiamati in causa fattori dietetici) oppure a fattori genetici, individuati in una generica predisposizione a reagire contro fenomeni esterni, tra cui virus e batteri.

Quest'ultima ipotesi si basa su studi condotti nei gemelli monozigoti che hanno permesso di dimostrare che il rischio che entrambi sviluppino diabete tipo 1 è del 30-40%, mentre scende al 5-10% nei fratelli non gemelli e del 2-5% nei figli. Si potrebbe, quindi, trasmettere una "predisposizione alla malattia" attraverso la trasmissione di geni che interessano la risposta immunitaria e che, in corso di una banale risposta del sistema immunitario a comuni agenti infettivi, causano una reazione anche verso le beta cellule del pancreas, con la produzione di anticorpi diretti contro di esse.

Questa alterata risposta immunitaria causa una progressiva distruzione delle cellule beta, per cui l'insulina non può più essere prodotta e si scatena così la malattia diabetica.

Per questo motivo, il diabete di tipo 1 viene classificato tra le malattie cosiddette "autoimmuni", cioè dovute a una reazione immunitaria diretta contro l'organismo stesso.

La terapia del DMT1 comprende la somministrazione di insulina tramite iniezione sottocutanea, è impossibile infatti somministrare l'insulina per via orale poiché verrebbe distrutta dall'acido e dagli enzimi gastrointestinali.

Di fondamentale importanza è poi il monitoraggio dei valori di glicemia a cui i pazienti devono sottoporsi più volte al giorno al fine di controllare l'efficacia della terapia insulinica e per limitare l'occorrenza di emergenze cliniche come crisi ipoglicemiche e chetoacidosi. Gli strumenti utilizzati per questo scopo saranno studiati più nel dettaglio nel capitolo successivo.

A causa del deficit di insulina, infatti, i pazienti con DMT1 non trattati hanno sempre elevate concentrazioni di glucosio nel sangue, ovvero alti livelli di glicemia: ciò accade perché il glucosio non è in grado di entrare normalmente nelle cellule bersaglio dell'insulina e di conseguenza il fegato produce continuamente glucosio attraverso la glicogenolisi e la gluconeogenesi rilasciandolo nel sangue.

È importante ricordare, inoltre, che l'insulina normalmente sopprime la lipolisi e la conseguente formazione di chetoni; quindi, un altro effetto di una carenza di insulina è un marcato aumento del glicerolo e degli acidi grassi plasmatici poi trasformati in chetoni dal fegato. Se portate all'estremo, queste variazioni metaboliche culminano nell'emergenza acuta con pericolo di vita chiamata chetoacidosi diabetica [1].

Uno dei problemi medici più impattanti sulla vita dei soggetti diabetici è costituito dagli effetti che una concentrazione di glucosio plasmatico estremamente elevata ha sulla funzione renale: l'aumentato livello di glucosio plasmatico nel diabete mellito aumenta il carico filtrato di glucosio oltre la capacità di riassorbimento tubulare massima e, quindi, vengono secrete elevate quantità di glucosio. Per la stessa ragione, nell'urina possono comparire elevate quantità di chetoni. Queste perdite urinarie impoveriscono il corpo di nutrienti e portano alla perdita di peso che è quindi uno dei sintomi di questa patologia.

Una condizione molto peggiore, tuttavia, è il fatto che questi soluti non riassorbiti causano una diuresi osmotica: un'aumentata escrezione urinaria di Na^* e di acqua, che può portare a ipotensione, danno cerebrale e morte. Tuttavia, ad eccezione di questo esempio estremo, i diabetici spesso sono maggiormente soggetti a ipertensione piuttosto che a ipotensione e ciò è dovuto a diverse cause, tra cui vi è il danno vascolare e renale dovuto a valori cronici alti di glicemia. L'altra anomalia grave nella chetoacidosi diabetica è l'aumento della concentrazione plasmatica di

H⁺ causata dall'accumulo di chetoni. Questa aumentata concentrazione causa disfunzioni cerebrali che possono portare se non curate al coma e alla morte.

Non da sottovalutare inoltre è il rischio di una possibile crisi ipoglicemica a seguito alla somministrazione di una quantità troppo elevata di insulina, questa infatti può portare a svenimento, convulsioni e, se non curata tempestivamente, a danni importanti cerebrali.

Come già illustrato precedentemente il diabete mellito di tipo 1 colpisce una percentuale piccola della popolazione, molto più comune è invece quello di tipo 2 che riguarda circa il 90% della popolazione diabetica.

Il DMT2 è una sindrome a insorgenza in età media che interessa prevalentemente adulti in sovrappeso; tuttavia, non è una sindrome correlata all'età.

Precedentemente si è potuto notare come la progressiva perdita di peso che si osserva nei pazienti con DMT1 è un sintomo del diabete mentre accade il contrario nel caso di diabete mellito di tipo 2: il motivo è che i soggetti con DMT2, al contrario di quelli con DMT1, non secernono una quantità di glucosio nell'urina sufficiente a causare una riduzione di peso. Inoltre, nel DMT2 è proprio l'aumento eccessivo di peso che contribuisce allo sviluppo della resistenza all'insulina e alla compromissione nella secrezione di insulina nei pazienti diabetici.

L'obesità, infatti, induce un certo grado di insulino-resistenza in particolare nel muscolo e nelle cellule del tessuto adiposo. Un'ipotesi per spiegare tale fenomeno è che il tessuto adiposo in eccesso produca in quantità eccessive alcuni messaggeri che causano una down-regolazione dei trasportatori del glucosio sensibili all'insulina. Un'ulteriore ipotesi è che un eccesso di deposizione di grasso in un tessuto non adiposo (come, ad esempio, il muscolo) causi una riduzione della sensibilità all'insulina.

Altro importante fattore di rischio per lo sviluppo della malattia è la familiarità, infatti, circa il 40% dei diabetici di tipo 2 ha parenti di primo grado affetti dalla stessa malattia, mentre nei gemelli monozigoti la concordanza della malattia si avvicina al 100%, suggerendo una forte componente ereditaria per questo tipo di diabete [4].

Come affermato in precedenza, molti soggetti con DMT2 non solo hanno insulino-resistenza, ma presentano difetti nella capacità delle cellule beta pancreatiche di secernere adeguatamente insulina in risposta a un aumento della concentrazione del glucosio plasmatico. In altre parole, sebbene la resistenza all'insulina sia il fattore principale che induce iperglicemia nel DMT2, un'alterazione della funzione delle cellule beta, non ancora identificata, impedisce a queste cellule

di rispondere in modo massimale all'iperglicemia stessa. Ad oggi si ritiene che i mediatori della diminuita sensibilità all'insulina descritti precedentemente possano interferire anche con la normale secrezione di insulina in risposta all'iperglicemia.

La terapia più efficace per i soggetti obesi con DMT2 è rappresentata dalla riduzione del peso corporeo. Anche l'esecuzione di esercizio fisico è molto importante, poiché dopo un'attività fisica duratura svolta frequentemente la sensibilità all'insulina aumenta, indipendentemente dalle variazioni di peso. Ciò avviene perché l'esercizio provoca un sostanziale aumento del numero di trasportatori di membrana per il glucosio sulle cellule del muscolo scheletrico, favorendone l'assorbimento.

Ai pazienti con DMT2 solitamente sono prescritti anche farmaci attivi per via orale che abbassano i valori di glucosio plasmatico attraverso vari meccanismi. Una forma sintetica di incretina recentemente approvata e un'altra classe di farmaci chiamati sulfaniluree, abbassano la concentrazione di glucosio plasmatico agendo sulle cellule beta pancreatiche e stimolandole a secernere insulina [1].

1.2.1) Complicanze associate

Sfortunatamente, i soggetti affetti da una forma di diabete mellito tendono a sviluppare molteplici malattie croniche correlate. Una concentrazione di glucosio plasmatico cronicamente aumentata contribuisce infatti alla maggior parte di queste alterazioni, sia causando l'accumulo intracellulare di alcuni metaboliti del glucosio che hanno effetti dannosi sulle cellule se presenti in concentrazioni elevate, sia legandosi alle proteine e, quindi, compromettendone il corretto funzionamento.

Il diabete può dare complicanze a livello di diversi organi e distretti del corpo nel corso degli anni e ciò è spesso la causa della ridotta qualità di vita dei soggetti diabetici. Il diabete è infatti una patologia rilevante anche per le numerose complicanze micro e macro-vascolari a cui le persone che ne sono affette vanno incontro [5].

Fra le complicanze micro-vascolari, le più rilevanti sono la retinopatia diabetica, che rappresenta la maggiore causa di cecità fra gli adulti, e la nefropatia diabetica, che è la principale causa di insufficienza renale cronica e dialisi.

Infatti, il diabete può danneggiare i piccoli vasi della retina e può quindi provocare la perdita progressiva della vista, fino alla cecità. La retinopatia rappresenta la maggiore causa di cecità in

soggetti in età lavorativa nei Paesi industrializzati. Le persone con diabete sono inoltre più esposte anche al rischio di sviluppare cataratta e glaucoma, questo poiché le elevate concentrazioni di glucosio plasmatico possono portare a un accumulo di metaboliti nel cristallino, causandone il rigonfiamento osmotico e questo, a sua volta, riduce la capacità degli occhi di focalizzare correttamente la luce sulla retina.

Nel caso della nefropatia invece il diabete danneggia i vasi sanguigni presenti nei reni che hanno la funzione di filtrare l'urina per depurare l'organismo dalle sostanze di scarto. Questo provoca la perdita progressiva della funzionalità renale, inducendo un grado crescente di insufficienza renale, fino alla perdita completa di funzione. Il diabete è tra le principali cause di insufficienza renale terminale che rende necessario il ricorso alla dialisi.

Le persone affette da diabete presentano, inoltre, un rischio più elevato di complicanze macrovascolari, ovvero malattie cerebro e cardio-vascolari, rispetto alla popolazione.

E' importante notare come tali patologie cardio-vascolari siano la causa di oltre il 50% delle morti per diabete. Infatti, il diabete aumenta molto il rischio di angina, infarto, ictus, arteriosclerosi a livello delle arterie delle gambe e può danneggiare i nervi e i vasi degli arti inferiori; questo espone al rischio di gravi infezioni che possono partire anche da piccole ferite, indotte da scarpe strette o dall'uso di forbici infette per la pedicure. Se non trattati adeguatamente e tempestivamente questi piccoli focolai di infezione possono espandersi, portare alla cancrena e alla necessità di amputare le dita o il piede o la gamba.

Altra importante complicanza macro-vascolare è la neuropatia: il diabete danneggiando i piccoli vasi che portano nutrimento ai nervi, in particolare a quelli degli arti inferiori, provoca un danno a carico dei nervi stessi che si manifesta con formicolii, dolori a carattere urente, riduzione della sensibilità alle dita dei piedi, poi a tutto il piede e alla gamba.

In ogni caso le complicanze croniche del diabete possono essere prevenute o se ne può rallentare la progressione attraverso uno stretto controllo di tutti i fattori di rischio correlati per cui è di fondamentale importanza il monitoraggio dei valori della glicemia attraverso l'esecuzione di opportuni test e utilizzando i giusti strumenti.

1.2.2) Ipotiroidismo e diabete mellito tipo 1

Un interessante argomento di ricerca degli ultimi anni riguarda lo studio della relazione tra due patologie autoimmuni molto diffuse: il diabete, in particolar modo quello di tipo 1, e l'ipotiroidismo [6].

In generale le malattie della tiroide e il diabete mellito rappresentano le endocrinopatie di più frequente riscontro nella pratica clinica, e sono spesso associate nello stesso soggetto. Un'alterata funzione tiroidea può influenzare la tolleranza al glucosio e peggiorare il controllo metabolico in pazienti con diabete: l'ipertiroidismo aumenta il rischio di emergenze iperglicemiche, mentre l'ipotiroidismo può determinare un peggioramento del compenso della glicemia.

L'ipotiroidismo è una comune malattia endocrina cronica. Si manifesta principalmente nelle donne, negli anziani e negli individui di razza bianca. Purtroppo, l'ipotiroidismo è spesso relativamente asintomatico, i sintomi sono aspecifici e si sovrappongono a quelli tipici delle persone sane. L'ipotiroidismo è associato a morbilità e mortalità cardiovascolare e se non viene trattato può portare a gravi effetti negativi sulla salute e infine alla morte.

La carenza di iodio è tradizionalmente la principale causa di ipotiroidismo in tutto il mondo. Oltre alla carenza di iodio, esistono molti fattori di rischio per l'ipotiroidismo, tra cui il sesso femminile, l'età >65 anni e l'uso di farmaci, l'ipotiroidismo è anche più comune nei pazienti con malattie autoimmuni, come il diabete di tipo 1, e può verificarsi come parte di molteplici endocrinopatie autoimmuni.

In particolar modo uno studio prospettico di 18 anni ha riportato un'importante associazione tra ipotiroidismo autoimmune e DMT1: le probabilità di ipotiroidismo nei DMT1 sono più alte rispetto a quelle della popolazione generale; inoltre, il DMT1 di solito si verifica circa 10 anni prima rispetto alla diagnosi di disfunzione tiroidea. Ciò suggerisce che il DMT1 e l'ipotiroidismo siano correlati: recenti studi hanno infatti dimostrato, attraverso tecniche con randomizzazione mendeliana, che il DMT1 è uno dei fattori di rischio decisivi per lo sviluppo dell'ipotiroidismo [7].

Tuttavia, se la relazione tra DMT1 e ipotiroidismo è chiara il meccanismo dietro al DMT1 e ipotiroidismo non lo è altrettanto. Diversi meccanismi biologici possono forse spiegare l'effetto causale di questa relazione.

Gli studi dimostrano che entrambe le malattie condividono la stessa suscettibilità genetica. Per quanto riguarda i geni condivisi coinvolti in questa predisposizione immunitaria, sono stati chiamati in causa i geni CTLA-4, HLA di classe II e FOXP3. È stato riportato che lo sviluppo di tiroidite autoimmune è associato a specifici marcatori di rischio genetici; ad esempio, l'ipotiroidismo è associato a HLAHLA-DQA1*0501. Negli adulti con DM1, l'essere di sesso femminile, l'età e la presenza di anticorpi contro la glutammato decarbossilasi (GAD) sono stati riportati come associati allo sviluppo di autoimmunità tiroidea.

Una spiegazione della correlazione potrebbe essere l'esistenza di una suscettibilità comune, con la distruzione autoimmune innescata da fattori ambientali in specifici individui geneticamente suscettibili.

In ogni caso, è dimostrato che il DM1 è un importante fattore di rischio per l'ipotiroidismo e che esiste una ragionevole spiegazione biologica alla base di tale fenomeno. Considerando, inoltre, che lo sviluppo dell'ipotiroidismo in soggetti diabetici porta, come già osservato, a un peggioramento nel controllo della glicemia con particolare pericolo di crisi ipoglicemiche e, nei pazienti pediatrici, a ritardi dello sviluppo e crescita, diviene di vitale importanza l'esecuzione con regolarità di test di funzionalità tiroidea, per ridurre al minimo il rischio di ipotiroidismo non diagnosticato nei pazienti giovani con DM1.

Ecco quindi l'importanza dello studio svolto in questo elaborato: individuare attraverso tecniche di machine learning i potenziali fattori di rischio per lo sviluppo dell'ipotiroidismo nei soggetti con DM1 e riuscire quindi a classificare correttamente i soggetti diabetici tra predisposti e non, permetterebbe di allertare gli individui interessati così che questi possano monitorare con maggiore attenzione la loro funzionalità tiroidea al fine di prevenire, o ritardare, l'eventuale sviluppo dell'ipotiroidismo.

Capitolo 2: Terapia e monitoraggio del diabete

2.1) Panoramica degli strumenti utilizzati

Come già precedentemente accennato, la terapia per il DMT1, e in alcuni casi anche per il DMT2, include la somministrazione di insulina nei momenti e con le modalità più opportune. Normalmente le cellule del pancreas rilasciano insulina, secondo dopo secondo, in rapporto al variare della glicemia. L'insulina prodotta dal pancreas raggiunge il fegato dove almeno metà dell'ormone, dopo aver agito, viene distrutta mentre l'altra metà raggiunge tutti gli altri tessuti, che ricevono quindi insulina in quantità molto minore rispetto al fegato [1].

Con la terapia insulinica l'ormone entra nel corpo dopo l'iniezione sottocutanea e questo fa sì che il fegato ne riceva meno rispetto al soggetto sano mentre gli altri tessuti ne ricevano in proporzione maggiore della norma. Inoltre, l'insulina viene assorbita in maniera differente da una sede di iniezione all'altra: più rapidamente dall'addome, più lentamente dalla coscia o dal braccio. È importante inoltre ricordare che, nella modalità tradizionale di somministrazione di insulina, la dose somministrata tiene conto di una stima del fabbisogno e non di una certezza sullo stesso.

Questo spiega perché si osservino molto spesso variazioni notevoli della glicemia da un giorno all'altro nello stesso orario nonostante la dose somministrata di insulina sia stata la stessa e l'alimentazione sia stata quasi identica. Le ipoglicemie e anche gli eccessivi rialzi glicemici sono infatti in parte il risultato di una stima approssimativa del fabbisogno insulinico in quel certo momento della giornata, oltre che di un assorbimento dell'insulina imprevisto e imprevedibile.

Le insuline disponibili si distinguono, a seconda della velocità e della durata di azione, in rapidissime (agiscono immediatamente), rapide (agiscono dopo 10/15 minuti), intermedie e a lunga durata. Solitamente quelle rapidissime e rapide sono utilizzate per le iniezioni di bolo rapido ovvero un bolo di insulina somministrato subito prima del pasto e regolato dal paziente in base alle esigenze del momento, come per esempio differenti tipologie di pasto o di valori glicemici. L'insulina intermedia e quella a lunga durata sono usate per le infusioni basali, la cui funzione è quella di mantenere stabili i valori della glicemia nel periodo di digiuno tra un pasto e l'altro. Normalmente i malati di DMT1 si autosomministrano prima di ogni pasto un bolo rapido e una o due volte al giorno una infusione basale [8].

Inoltre, anche le modalità di somministrazione dell'insulina sono varie: ciò può essere fatto attraverso l'utilizzo di siringhe e pennini o anche attraverso un microinfusore, chiamato pompa di insulina.

In particolar modo il microinfusore è un dispositivo elettromeccanico che consente un'infusione continua di insulina a rapida azione nel tessuto sottocutaneo, il microinfusore può essere controllato attraverso l'utilizzo di applicazioni apposite che forniscono un'interfaccia intuitiva con l'utente. Ne esistono 2 tipologie: con il filo e senza filo. Nel primo caso, il microinfusore è dotato di un catetere di lunghezza variabile che collega il microinfusore al sottocute del paziente, attraverso un cerotto dotato di agocannula; nel secondo invece il microinfusore è adeso, sempre attraverso un cerotto dotato di ago-cannula, direttamente alla cute del paziente. In entrambi i casi tale strumento permette la somministrazione di insulina in corpo in modo più accurato a seconda delle necessità del soggetto e riduce tutti gli effetti imprevedibili dell'insulina intermedia e a lunga durata.

Il più recente metodo di somministrazione d'insulina è quello aereo, attraverso l'utilizzo di un inalatore apposito. Tale strumento permette però la somministrazione solo dell'insulina ultrarapida e quindi deve essere integrato con l'iniezione di un'insulina a lunga durata, inoltre non permette un dosaggio preciso e presenta alcuni effetti indesiderati come la tosse sebbene elimini la necessità dell'utilizzo di aghi o canule, spesso considerati fastidiosi per i pazienti.

Oltre alla somministrazione di insulina una seconda imprescindibile fase della terapia per i malati di DMT1 e anche per alcuni di DMT2, è costituita dal monitoraggio dei valori di glicemia plasmatica, così da poter effettivamente valutare l'efficacia della somministrazione insulinica o l'eventuale necessità di ulteriori aggiustamenti nelle dosi.

Il monitoraggio della glicemia deve avvenire sia giornalmente a cura del singolo diabetico tramite l'utilizzo di vari strumenti dedicati sia in maniera meno frequente nei laboratori di analisi tramite analisi apposite, entrambi questi dati vengono utilizzati dal medico curante per creare un quadro completo dello sviluppo della malattia e un piano terapeutico ottimale.

Tra questi ultimi uno dei più importanti esami, utilizzato sia per effettuare la diagnosi di diabete mellito che per monitorare soggetti già in cura, è il test emoglobina glicata o glicosilata o più semplicemente chiamato test A1c (HbA1c) [9].

Il test A1C è un esame del sangue che misura i livelli medi di glucosio nel sangue del soggetto negli ultimi 3 mesi grazie al calcolo della quantità di emoglobina glicosilata nel campione. Infatti, l'emoglobina diventa emoglobina glicosilata attraverso la glicosilazione che porta la molecola di

zucchero a legarsi all'emoglobina all'interno dei globuli rossi. Le persone con diabete presentano quindi valori più elevati di HbA1c nei globuli rossi poiché presentano maggior quantità di zucchero nel sangue, nel caso di un soggetto curato e in terapia insulinica i valori di HbA1c ematica devono essere mantenuti inferiori a 53 mmol/mol (7 %) affinché la terapia possa essere considerata efficace e così da limitare le complicanze della patologia.

Altrettanto importanti sono i test effettuati dai singoli individui giornalmente: lo strumento più comune per controllare il livello di glucosio nel sangue anche a casa è il glucometro (Blood Glucose Monitoring o brevemente BGM).

Il glucometro è infatti un apparecchio portatile per la determinazione rapida della glicemia su sangue capillare e costituisce il mezzo indispensabile per l'autocontrollo glicemico da parte del paziente. Tale modalità di monitoraggio viene indicata con SMBG, dall'acronimo inglese di Self Monitoring of Blood Glucose [10].

Consentono di misurare la glicemia su sangue intero, di solito sangue capillare ottenuto dalla puntura del polpastrello di un dito. Questo valore è quindi diverso, ma non troppo discostato, dalla glicemia ottenuta analizzando il sangue plasmatico misurata nei durante le analisi del sangue ma lo strumento consente, tramite apposite impostazioni, di calcolare in automatico il valore plasmatico corrispondente così da non creare confusione nell'utente.

Lo strumento complessivamente è costituito da tre elementi separati: un pennino contenente una lancetta, ovvero un piccolo ago molto sottile che pungendo il lato del polpastrello causa la fuoriuscita di una goccia di sangue, il blocco elaboratore con display che permette la visualizzazione della glicemia istantanea e la parte sensibile dello strumento, ovvero delle strisce reattive che vengono messe a contatto con la goccia di sangue e successivamente inserite nel blocco con display. Alcuni apparecchi memorizzano i valori glicemici determinati e possono essere collegati con PC mediante apposito accessorio, altri possono essere inseriti in sistemi di collegamento per il tele-autocontrollo del diabete.

Il principio su cui si basa la misura dei più comuni glucometri consiste nella quantificazione di una reazione enzimatica, l'ossidazione del glucosio, che risulta proporzionale alla sua concentrazione nel sangue. La reazione avviene quando viene posto un campione di sangue su una striscia reattiva alla glucosio-ossidasi. La misura può essere ottenuta con metodo riflettometrico, viene misurata l'intensità di colore determinata dal cromogeno che si sviluppa dall'ossidazione del glucosio, o impedenziometrico, si misura la conducibilità elettrica del sangue indotta dalla corrente

elettrica generata dall'ossidazione del glucosio. In tutti i casi, poiché la misura si basa su una reazione enzimatica, il tempo di reazione costituisce un requisito critico per la misura.

Tutti i modelli dei glucometri possiedono una sensibilità e riproducibilità dei risultati ritenute soddisfacenti in rapporto allo scopo, che è quello di fornire valori glicemici utilizzabili per una decisione terapeutica immediata o una rivalutazione della terapia.

La concordanza tra la glicemia misurata dal glucometro e quella determinata con metodiche di laboratorio convenzionali è mediamente buona, ma risulta variabile in rapporto al livello glicemico: la concordanza è elevata nel range 50-250 mg/dl, mentre diminuisce, pur risultando accettabile, nel range dei valori glicemici superiori o inferiori. Sebbene il valore ideale raccomandato per l'accuratezza sia di un errore analitico inferiore al 5%, l'imprecisione analitica dei glucometri dipende dal tipo di strumento, con un coefficiente di variazione compreso tra 4% e 20% nel range glicemico tra 130 e 350 mg/dl. Nonostante tale imprecisione, l'affidabilità del glucometro risulta accettabile quando il suo impiego sia confinato all'automonitoraggio della glicemia.

Ulteriore strumento, il cui utilizzo è sempre più in aumento negli ultimi anni, è il Monitoraggio in Continuo del Glucosio, o brevemente CGM. Nell'ultimo decennio, infatti, il CGM si è evoluto come una parte essenziale della gestione del diabete per molte persone affette da diabete di tipo 1.

Questa tecnologia offre alcuni vantaggi rispetto al tradizionale automonitoraggio della glicemia, effettuato attraverso il glucometro, fornendo informazioni in tempo reale sull'andamento del glucosio, direzione e velocità delle variazioni di glucosio, e avvisi di ipo o iperglicemia. Diversi studi multicentrici randomizzati e controllati hanno dimostrato i benefici del CGM nel ridurre l'HbA1c e l'ipoglicemia, in particolare negli adulti con DMT1.

2.1.1) Monitoraggio Continuo del Glucosio

La terapia tradizionale e più comune nei pazienti diabetici, come già ampiamente spiegato, si basa sull'approccio farmacologico, sull'autocontrollo della glicemia e sulla dieta e l'esercizio fisico finalizzati a mantenere l'euglicemia, ovvero un livello di glucosio nel sangue compreso tra i 70 mg/dl e i 120 mg/dl.

La modalità di automonitoraggio più comune è l'SMBG, gli svantaggi principali di questi strumenti, però, non sono pochi e risiedono nell'invasività della procedura per la lettura del valore di glucosio e nel fatto che essi consentono solamente poche misurazioni in una giornata, non

essendo quindi in grado di cogliere eventuali episodi di iperglicemia o ipoglicemia durante i periodi tra una misurazione e l'altra.

Per questo motivo sono stati sviluppati, e comparsi sul mercato a partire dai primi anni 2000, dispositivi che consentono un monitoraggio continuo della glicemia nei liquidi interstiziali, con una misura ogni 1-5 minuti, detti dispositivi per il Monitoraggio in Continuo del Glucosio (CGM). Essi consentono infatti di individuare un numero maggiore di rilevazioni rispetto al convenzionale SMBG e di effettuare un'analisi retrospettiva dell'andamento glicemico stesso, in base alla quale si può aggiustare opportunamente la terapia [11].

Da più di 20 anni, il CGM è diventato sempre più affidabile, vantaggioso da un punto di vista clinico ed economico e ha dimostrato efficacia in termini di miglioramento dell'A1C, riduzione dell'ipoglicemia e miglioramento del tempo nell'intervallo target del glucosio. Questo perché vedere i livelli di glucosio nel sangue in tempo reale può aiutare a prendere decisioni più consapevoli sui cibi e le bevande che si possono consumare, sull'attività fisica che si deve svolgere e sui farmaci che si assumono. Mantenere il livello di glucosio nel range di riferimento può aiutare a prevenire altri problemi di salute causati dal diabete.

Un CGM, quindi, stima il livello di glucosio ogni pochi minuti e ne tiene traccia nel tempo, permettendo così di controllare in tempo reale la curva dell'andamento glicemico e di allertare con suoni e vibrazioni di tipo differente i pazienti diabetici prima che i livelli di glucosio giungano a valori eccessivamente elevati o rischiosamente ridotti, aiutando quindi a prevenire le crisi ipoglicemiche o la chetoacidosi.

Poiché l'utilizzo del CGM consente l'acquisizione di una grande mole di dati, questi dati risultano essere molto utili per effettuare studi di natura scientifica, prestandosi ad esempio all'utilizzo di tecniche machine learning che possano portare alla scoperta di nuove nozioni mediche sul diabete o che possano accostarsi al medico e aiutarlo nella compilazione del piano di cura dei pazienti più adatto. Lo studio svolto in questo elaborato, che è finalizzato allo sviluppo di un sistema di supporto alla decisione clinica per il riconoscimento dei soggetti con DMT1 a rischio di ipotiroidismo, utilizza come database di partenza proprio un set di dati costituiti dalle acquisizioni fatte da dispositivi CGM.

Un CGM è composto da tre parti: la prima è un piccolo sensore, costituito da un piccolo ago-elettrodo sottile, che può essere inserito sotto la pelle, spesso quella della pancia o del braccio, con un cerotto adesivo che impedisce il movimento del sensore garantendo misure più pulite. Un nuovo tipo di sensore CGM, chiamato sensore impiantabile, può essere inserito all'interno del corpo. I

sensori CGM stimano il livello di glucosio nel fluido tra le cellule, valori molto simili al livello di glucosio plasmatico. La parte sensibile del CGM è detta monouso poiché superata una certa scadenza questa deve essere sostituita, ad esempio la sostituzione è solitamente fatta ogni poche settimane, a seconda del tipo di sensore di cui si dispone.

La seconda parte del CGM è costituita da un trasmettitore: questo invia le informazioni, senza l'uso di cavi, alla terza parte, un programma software memorizzato su uno smartphone, su una pompa di insulina o su un dispositivo separato che viene generalmente chiamato ricevitore.

Per quanto riguarda il funzionamento dei ricevitori, nei modelli di tipo impiantabile questi sono costituiti da un ago-elettrodo posizionato nel sottocute dove avviene la misura della concentrazione di glucosio nel liquido interstiziale, oppure da una microfibra da microdialisi, impiantabile tramite ago-cannula. In questo caso, la microfibra posizionata nel sottocute preleva per osmosi il glucosio interstiziale, che viene poi convogliato tramite un liquido di perfusione ad un bio-sensore esterno.

Nel caso invece dei modelli di tipo non invasivo, questi sono invece dotati di elettrodo applicato alla superficie della cute: in questa sede viene misurata la quantità di glucosio estratta dal tessuto sottocutaneo con un processo di ionoforesi che consiste nell' applicazione di un potenziale sulla cute che determina il trasporto transcutaneo di glucosio per elettro-osmosi. Entrambi questi sistemi riducono o eliminano il dolore e gli inconvenienti connessi con il prelievo di sangue capillare e con la lettura della striscia mediante glucometro.

Oggi i pazienti affetti da diabete possono disporre, in particolare, di due tipologie di dispositivi per il monitoraggio continuo del glucosio: si tratta del CGM in tempo reale e del CGM a rilevazione intermittente (Flash glucose monitoring) [12].

La principale differenza tra i due dispositivi sta nel fatto che i CGM in tempo reale inviano continuamente e in modo automatico ogni valore misurato al ricevitore, che poi li memorizza e ne permette la visualizzazione rielaborando i dati. Si tratta quindi di una mole di dati molto grande che deve essere opportunamente gestita dal dispositivo e che può occupare molto spazio in memoria. Il vantaggio più significativo dei sistemi CGM real-time è la presenza di allarmi sonori che avvisano in tempo reale se la glicemia sta diventando troppo alta, in questo modo si ha il tempo di apportare modifiche che possono ridurre l'impatto di una glicemia alta o bassa o evitarla del tutto. Il CGM in tempo reale solitamente è notevolmente più costoso della sua alternativa FGM.

Il CGM a rilevazione intermittente invece richiede la scansione del dispositivo per ottenere i dati sul glucosio. Il sensore campiona e misura continuamente i livelli di glucosio, genera un nuovo valore di glucosio ogni minuto e registra la lettura ogni 15 minuti per un tempo di utilizzo del

sensores compreso tra 10 e 14 giorni, a seconda del modello ma non invia automaticamente i dati raccolti al ricevitore ed ha uno spazio in memoria piuttosto limitato. È fondamentale quindi, affinché l'utilizzo del FGM sia efficace la costanza e precisione del paziente nella scansione del dispositivo: il sensore può infatti fornire i valori del glucosio per 10-14 giorni se la scansione viene effettuata almeno ogni otto ore. In caso contrario, le informazioni sul glucosio più vecchie di otto ore verranno sovrascritte e non saranno disponibili per la decisione o il download.

Prima di dicembre 2016, i sistemi CGM disponibili in commercio negli Stati Uniti sono stati approvati dalla Food and Drug Administration (FDA) statunitense solo per l'uso come dispositivi aggiuntivi alle informazioni ottenute dal normale monitoraggio del glucosio nel sangue con BGM. Pertanto, era necessaria una misurazione BGM per confermare la concentrazione di glucosio misurata dal sensore CGM prima di prendere una decisione sul dosaggio dell'insulina. Questa decisione normativa è stata presa presumibilmente perché l'accuratezza dei sistemi CGM era, al tempo, considerata inadeguata per il dosaggio dell'insulina senza la conferma del BGM. Tuttavia, con ogni nuova generazione di sensori, l'accuratezza è notevolmente migliorata e numerosi studi hanno evidenziato come oggi il CGM sia sufficientemente accurato da poter essere implementato in modo sicuro come strumento autonomo per il monitoraggio del glucosio e per le decisioni terapeutiche [13].

Lo strumento ideale del futuro per la gestione del diabete è costituito da sistemi miniaturizzati "ad ansa chiusa" in grado di infondere la quantità di insulina necessaria a mantenere la glicemia a livelli quasi normali in base al valore glicemico istantaneo. Al giorno d'oggi, grazie all'utilizzo e all'integrazione reciproca degli strumenti precedentemente presentati, si sta arrivando molto vicini alla realizzazione di questo strumento ideale [10].

Si parla già infatti di pancreas artificiale, o sistema automatico di somministrazione dell'insulina (AID): un sistema che è in grado di imitare accuratamente il modo in cui un pancreas sano controlla la glicemia nell'organismo. Il pancreas artificiale è costituito da un CGM, una pompa di insulina e un programma software che condivide le informazioni tra il CGM e la pompa di insulina.

Il CGM stima i livelli di glucosio e invia in modalità wireless le informazioni al programma software su uno smartphone o su una pompa di insulina. Il programma calcola la quantità di insulina di cui il corpo ha bisogno e il microinfusore eroga l'insulina quando i livelli di glucosio superano l'intervallo di riferimento. D'altra parte, se i livelli di glucosio scendono al di sotto dell'intervallo di riferimento, il pancreas artificiale può ridurre o interrompere la quantità di insulina somministrata dal microinfusore di insulina.

L'ottimizzazione di questo sistema permetterebbe la regolazione della glicemia nei range sani in modo continuo e completamente automatico, consentendo così ai pazienti diabetici di effettuare una terapia efficace e ben monitorata.

Capitolo 3: Sistemi di supporto alla decisione clinica

Il Sistema di Supporto alle Decisioni Cliniche (CDSS) è un Sistema informatico di Supporto alle Decisioni (DSS) in ambito sanitario. In genere si basa su software per computer progettati per assistere medici e altri professionisti sanitari con compiti decisionali, come quando applicare una diagnosi particolare, ulteriori test o trattamenti specifici [14].

Lo scopo principale del moderno CDSS è assistere i medici nel punto di cura, affinché il medico possa combinare le proprie conoscenze con informazioni o suggerimenti forniti dal CDSS.

Un CDSS prende i dati e le informazioni immesse ed elabora queste informazioni con l'utilizzo di modelli organizzativi, algoritmi e calcoli al fine di ottenere una varietà di potenziali opzioni di azione basate sulle circostanze uniche del singolo paziente. Al giorno d'oggi sempre più CDSS in fase di sviluppo hanno la capacità di sfruttare dati e osservazioni altrimenti irraggiungibili o non interpretabili dall'uomo: è necessario però tenere sempre presente che i CDSS non sono pensati per sostituire i medici, ma per potenziarli.

I sistemi di supporto alla decisione clinica possono essere classificati in due categorie principali in base al loro grado di interazione e al ruolo svolto durante il processo decisionale. Le due categorie sono spesso denominate CDSS attivi e passivi.

I CDSS passivi forniscono informazioni o suggerimenti senza un'interazione diretta con l'utente, questi sistemi forniscono raccomandazioni o evidenze scientifiche basate su algoritmi predefiniti o regole senza richiedere un'azione immediata da parte del professionista sanitario. Questo è il caso, ad esempio, di un avviso di interazione farmacologica che appare quando viene prescritto un farmaco specifico.

I CDSS attivi, invece, interagiscono direttamente con l'utente, spesso richiedendo input o azioni specifiche. Questi sistemi possono fare domande, chiedere ulteriori dettagli o richiedere una conferma prima di fornire una raccomandazione finale: sono progettati per guidare attivamente il professionista sanitario nel processo decisionale. Ad esempio, un sistema di questo genere è un sistema che, durante la prescrizione di un farmaco, richiede dettagli sulla storia clinica del paziente o chiede conferme aggiuntive prima di finalizzare la prescrizione.

Una seconda classificazione dei CDSS può essere fatta sulla base del tipo di approccio utilizzato per generare raccomandazioni e supportare le decisioni dei professionisti sanitari.

I CDSS possono essere basati sulla conoscenza (knowledge-based) ed in tal caso utilizzano regole e conoscenze esplicite fornite da esperti umani o integrate da fonti scientifiche. Questi sistemi incorporano informazioni mediche, linee guida cliniche e protocolli per generare raccomandazioni. L'approccio spesso implica la creazione di regole "if-then", in cui il sistema segue uno schema logico di risposte a situazioni specifiche.

I CDSS non basati sulla conoscenza (non knowledge based) si affidano all'analisi dei dati e all'Intelligenza Artificiale (IA) per identificare pattern e relazioni nei dati clinici. Questi sistemi utilizzano algoritmi di machine learning per analizzare grandi quantità di dati, identificare pattern nascosti e formulare raccomandazioni basate su modelli predittivi. Possono adattarsi a nuove informazioni senza richiedere un intervento umano diretto e sono in grado di apprendere dai dati disponibili.

In sintesi, mentre i CDSS basati sulla conoscenza dipendono dalla formulazione di regole esplicite da parte di esperti umani, i CDSS non basati sulla conoscenza si affidano all'analisi dei dati e all'apprendimento automatico per identificare modelli e relazioni nei dati clinici. Entrambi gli approcci hanno vantaggi e limitazioni, e la scelta tra i due dipende spesso dalle esigenze specifiche del contesto clinico e dalle caratteristiche dei dati disponibili.

3.1) Machine learning

Il machine learning, o apprendimento automatico, è una branca dell'intelligenza artificiale che si occupa dello sviluppo di algoritmi e modelli che consentono ai computer di apprendere da dati e migliorare le proprie prestazioni senza essere esplicitamente programmati per compiti specifici. In altre parole, il machine learning permette ai sistemi informatici di riconoscere pattern nei dati, trarre conclusioni, e prendere decisioni basate sull'esperienza acquisita da insiemi di dati precedentemente analizzati [15].

Il processo di apprendimento automatico coinvolge tipicamente tre fasi principali:

- **Addestramento (Training):** Durante questa fase, il modello di machine learning viene esposto a un grande insieme di dati di input e alle rispettive risposte o etichette. Il modello impara a riconoscere pattern e relazioni nei dati.
- **Validazione e Test:** Dopo l'addestramento, il modello viene testato su nuovi dati per valutare la sua capacità di generalizzazione e di fare previsioni accurate su dati non visti precedentemente. La validazione aiuta a ottimizzare i parametri del modello per garantire prestazioni robuste.
- **Predizione o Inferenza:** Una volta addestrato e validato, il modello è pronto per fare previsioni o inferenze su nuovi dati di input senza la necessità di essere esplicitamente programmato per tali dati.

Il machine learning, quindi, insegna ai computer a fare ciò che riesce naturalmente agli esseri umani: imparare dall'esperienza. Viene utilizzato per problemi complessi che coinvolgono una grande quantità di dati con molte variabili, in cui non esiste formula o un'equazione in grado di descrivere il sistema.

Il machine learning utilizza due tipi di tecniche fondamentali: le tecniche di apprendimento non supervisionato e quelle di apprendimento supervisionato [16].

L'apprendimento non supervisionato trae inferenze da set di dati in input che non hanno associate le corrispondenti risposte etichettate, questa tecnica è utilizzata quando si desidera esplorare i dati ma non si ha ancora un compito specifico in mente o non si è sicuri di quali informazioni i dati contengano.

Tra le tecniche che sfruttano tale tipo di apprendimento la più utilizzata è il clustering: suddivide i dati in gruppi diversi in base alle caratteristiche condivise tra questi; il clustering viene infatti utilizzato per trovare raggruppamenti nascosti in applicazioni come l'analisi delle sequenze genetiche, la ricerca di mercato e il riconoscimento di oggetti.

L'apprendimento non supervisionato è anche un buon modo per ridurre la dimensione dei dati con cui lavorare, ciò risulta importante poiché la maggior parte delle tecniche di apprendimento non supervisionato sono una forma di analisi di clustering che separa i dati in gruppi in base a caratteristiche condivise. Quindi, l'apprendimento non supervisionato può essere utilizzato come un passaggio preliminare per ridurre la dimensione dell'input dati per l'apprendimento supervisionato.

D'altra parte, l'apprendimento supervisionato richiede invece che ognuno dei dati in input sia associato a un'uscita correttamente etichettata.

Tutte le tecniche di apprendimento supervisionato assumono la forma di classificazione o regressione: le tecniche di classificazione prevedono risposte discrete.

Questo tipo di tecniche sono utilizzate se le uscite che si vogliono prevedere possono essere separate in diversi gruppi; esempi di problemi di classificazione includono l'elaborazione di immagini mediche, il riconoscimento vocale e la valutazione del credito. Le tecniche di regressione, d'altra parte, prevedono risposte continue; un buon esempio di ciò è qualsiasi applicazione in cui l'output che si sta prevedendo può assumere qualsiasi valore in un certo intervallo, come i prezzi delle azioni e l'elaborazione dei segnali acustici.

Approfondendo le tecniche di classificazione, l'algoritmo di regressione logistica è uno dei più semplici algoritmi sfruttati da tali tecniche ed è utilizzato principalmente per problemi di classificazione binaria, ovvero problemi in cui ci sono solo due possibili risultati. Funziona meglio quando i dati possono essere ben separati da un singolo confine lineare; è possibile utilizzarlo anche come riferimento di base per confrontarlo con metodi di classificazione più complessi.

Altre molto usate tecniche di classificazione sono gli algoritmi degli alberi decisionali bagged e boosted combinano singoli alberi decisionali che hanno una minore capacità predittiva in un insieme di molti alberi che hanno complessivamente una maggiore capacità predittiva. Tali tecniche sono meglio utilizzate quando i predittori sono discreti o si comportano in modo non lineare e quando si dispone di più tempo per addestrare il modello.

Per quanto riguarda le tecniche di regressione, ci sono due algoritmi importanti che possono essere utilizzati: la regressione lineare e la regressione non lineare. La regressione lineare è una tecnica di modellizzazione statistica, viene utilizzata quando l'algoritmo deve essere facile da interpretare e veloce, o come punto di riferimento per valutare altri modelli di regressione più complessi. La regressione non lineare aiuta a descrivere relazioni e dati più complessi, è meglio utilizzata quando i dati presentano tendenze non lineari pronunciate e non possono essere facilmente trasformati in uno spazio lineare.

La differenza fondamentale tra l'apprendimento supervisionato e l'apprendimento non supervisionato è quindi che l'apprendimento supervisionato richiede esempi correttamente etichettati per addestrare il modello di ML e quindi utilizza tale modello nelle fasi successive per

etichettare nuovi dati. Le tecniche utilizzate e gli algoritmi selezionati dipendono dalla dimensione e dal tipo di dati dello studio, da quali sono i tipi di modelli che si vogliono ottenere dai dati e da come tali modelli verranno utilizzati successivamente.

Un altro aspetto importante riguardante l'uso delle tecniche di apprendimento automatico consiste in tutti i diversi modi per migliorare un modello. Approfondendo brevemente solo tre dei metodi più famosi: il primo è la selezione delle caratteristiche, dove si identificano gli input più rilevanti dai dati che forniscono la migliore capacità predittiva. Infatti, un modello può essere valido solo quanto le caratteristiche che si utilizzano per addestrarlo.

Il secondo metodo consiste nella trasformazione delle caratteristiche: una forma di riduzione della dimensione per ridurre la complessità dei dati, il che può renderli molto più facili da rappresentare e analizzare.

Il terzo modo per aumentare l'accuratezza dei modelli è l'ottimizzazione dei parametri, è un processo iterativo dove l'obiettivo è trovare le migliori impostazioni possibili per addestrare il modello: addestri il tuo modello molte volte utilizzando diverse impostazioni fino a scoprire la combinazione di impostazioni che produce il modello più accurato.

In ambito biomedico il machine learning sfrutta algoritmi e modelli di apprendimento automatico per analizzare, interpretare e estrarre informazioni significative dai dati biomedici. Questa area ha il potenziale di rivoluzionare la ricerca medica, la diagnosi, il trattamento e la gestione delle malattie. Alcune delle principali applicazioni in ambito biomedico sono infatti le seguenti:

- **Diagnosi medica:** Gli algoritmi di machine learning possono essere addestrati su grandi set di dati di casi clinici, consentendo loro di riconoscere pattern e correlazioni difficilmente individuabili dall'occhio umano. Questo può migliorare la precisione e la tempestività delle diagnosi mediche.
- **Predizione del rischio:** Utilizzando dati clinici e genetici, i modelli di machine learning possono stimare il rischio di sviluppare determinate condizioni mediche. Ciò è particolarmente utile per identificare precocemente i pazienti a rischio e adottare misure preventive.
- **Gestione dei pazienti:** I modelli di machine learning possono essere utilizzati per monitorare i pazienti in modo continuo, prevedere l'andamento delle malattie croniche e ottimizzare la gestione delle cure, riducendo i ricoveri ospedalieri non necessari.

- **Analisi di immagini mediche:** Il machine learning può essere impiegato per analizzare immagini mediche, come radiografie, risonanze magnetiche e tomografie computerizzate. Questo aiuta nella diagnosi precoce di patologie e nell'identificazione di dettagli che potrebbero sfuggire all'occhio umano.
- **Ricerca scientifica:** Il machine learning facilita l'analisi di grandi dataset biomedici, accelerando la ricerca scientifica e aprendo nuove prospettive per la comprensione delle malattie e lo sviluppo di terapie innovative.

Il machine learning rappresenta, infatti, uno strumento potente nel campo della ricerca legato a tutti i tipi di diabete (DMT1, DMT2 e diabete gestazionale). Le sue applicazioni sono state infatti dedicate alla classificazione di individui affetti da DMT2 o prediabete rispetto a soggetti sani, alla previsione della progressione dal diabete gestazionale al DMT2, alla rilevazione precoce del rischio di sviluppo del diabete a partire dai dati dei registri elettronici della salute e allo sviluppo di nuovi indici surrogati che descrivono specifici processi fisiologici [17].

Nel caso dell'elaborato in esame, le tecniche di ML sono state utilizzate per generare modelli affidabili che permettessero di distinguere i soggetti con solo DMT1 da quelli con anche ipotiroidismo.

E' indicato quindi l'utilizzo del machine learning supervisionato, si conoscono già infatti dal dataset di partenza i soggetti affetti da ipotiroidismo e quelli non, ed è allora possibile etichettare correttamente i dati forniti in ingresso al modello per addestrarlo; inoltre tra le tecniche di regressione e quelle di classificazione, queste ultime sono le più adatte al problema in esame poiché la risposta in uscita può assumere solo due valori discreti: malato (di ipotiroidismo) o sano (quindi avente solo DMT1).

Un tale modello, se trovato, permetterebbe ai soggetti suscettibili ad ipotiroidismo di essere preventivamente monitorati così da ritardare o evitare, qualora possibile, lo sviluppo della patologia.

Capitolo 4: Sviluppo di un metodo di machine learning per la selezione di metriche da monitoraggio continuo del glucosio discriminanti la presenza di ipotiroidismo

4.1) Acquisizione dati

4.1.1) Campagna Replace BG

I dati con i quali è stato realizzato il progetto sono stati opportunamente selezionati da un grande database open source messo a disposizione dal Jaeb Center for Health Research (JCHR), che sarà di qui in avanti denominato “dataset Aleppo”. In particolare i dati utilizzati provengono tutti dallo studio REPLACE-BG: A Randomized Trial Comparing Continuous Glucose Monitoring With and Without Routine Blood Glucose Monitoring in Adults With Well-Controlled Type 1 Diabetes [13].

Il lavoro consisteva in uno studio clinico randomizzato di non inferiorità condotto in 14 siti della T1D Exchange Clinic Network su 226 soggetti i quali dovettero automonitorarsi per 26 settimane tramite l'utilizzo (esclusivo o no) di un CGM Dexcom in grado di rilevare la glicemia sanguigna con un periodo di campionamento di 5 minuti.

I partecipanti avevano età ≥ 18 anni (media 44 ± 14 anni), erano affetti da T1D da ≥ 1 anno (durata media 24 ± 12 anni), utilizzavano un microinfusore di insulina e avevano un'HbA1c pari a 9,0% (75 mmol/mol) (media $7,0 \pm 0,7\%$ [$53,7 \pm 6,7$ mmol/mol]).

L'obiettivo dello studio era determinare se l'utilizzo del CGM senza la conferma del BGM fosse tanto sicuro ed efficace quanto l'uso del CGM in aggiunta al BGM, come da raccomandazione medica, in adulti con diabete di tipo 1 ben controllato.

I risultati ottenuti dallo studio confermano l'ipotesi appena espressa: l'uso del CGM è stato in media di $6,7 \pm 0,5$ e $6,8 \pm 0,4$ giorni/settimana nei gruppi solo CGM e CGM+BGM rispettivamente nei gruppi CGM+BGM, nel corso delle 26 settimane di studio. I test del CGM al giorno (compresi i due necessari per la calibrazione del CGM) sono stati in media $2,8 \pm 0,9$ e $5,4 \pm 1,4$ nei due gruppi. Il tempo medio nel range 70-180 mg/dL è stato del $63 \pm 13\%$ sia al basale che a 26 settimane nel gruppo con solo CGM e $65 \pm 13\%$ e $65 \pm 11\%$ nel gruppo CGM+BGM. Non si sono verificati eventi ipoglicemici gravi nel gruppo CGM e uno nel gruppo CGM+BGM.

4.1.2) Creazione del dataset di studio dal dataset di partenza

Di tutti i dati relativi ai 226 soggetti acquisiti grazie allo studio appena illustrato, solo alcuni sono stati ritenuti di rilievo per il presente elaborato: in particolare modo solo il 9% dei soggetti presentava unicamente T1D come morbosità, il restante 91% dei soggetti presentava infatti altre patologie in associazione al T1D. Ciò è evidenziato dal grafico sottostante (Fig. 1).

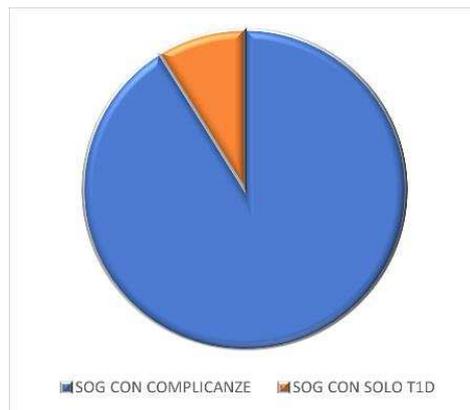


Fig. 1: Il grafico a torta mostra in percentuale il numero di soggetti presenti nel Dataset Aleppo con solo DM1, in arancione, e il numero di soggetti presentanti qualche altra morbosità, in blu

Questo dato evidenzia ancora una volta quanto le complicanze legate al diabete rappresentino uno degli aspetti più pericolosi di questa malattia: ecco perché risulta di vitale importanza effettuare studi come il presente che permettano di individuare i fattori di rischio di tali patologie con opportuno anticipo.

Le patologie più frequenti e il numero di soggetti che ve ne sono affetti sono illustrati nella sottostante figura (Fig 2): si può notare come la complicanza associata più presente sia senza dubbio costituita dalla ipertensione [18]. Come già spiegato infatti vi è una stretta correlazione tra malattie cardiovascolari e diabete: circa il 60% in media di persone affette da diabete sviluppa anche ipertensione, ed anche i soggetti che soffrono di alta pressione arteriosa hanno maggior probabilità di sviluppare una certa insulino-resistenza. Questa importante relazione è, però, già stata argomento di numerose ricerche e studi per cui per questo studio si ha preferito rivolgere l'attenzione sulla seconda patologia più incidente nella popolazione di riferimento: l'ipotiroidismo.

Infatti, dei totali 226 individui inclusi nello studio ben 51 di questi presentavano, o hanno sviluppato nel corso dei 6 mesi di monitoraggio, ipotiroidismo, numero sensibilmente maggiore rispetto a tutte le altre complicanze presenti, come mostra il grafico sottostante.

Anche questo risultato, come quello precedente, è supportato dalle evidenze scientifiche: i disturbi autoimmuni della tiroide sono le malattie immunologiche più diffuse nei pazienti con diabete di tipo 1. Studi trasversali hanno riportato una prevalenza di ipotiroidismo nel 12-24% delle donne e nel 6% dei maschi con diabete di tipo 1 [7]. La maggior parte dei pazienti presenta una malattia subclinica e lo sviluppo del diabete di solito precede la diagnosi di ipotiroidismo.

La relazione diretta tra le due patologie è quindi una evidenza scientifica comprovata tuttavia mancano studi prospettici a lungo termine per valutare l'incidenza e la storia naturale dei disturbi tiroidei nei pazienti con diabete di tipo 1 in grado di spiegare profondamente il perché di questa connessione. A questo proposito, è uno degli obiettivi di questo studio indagare quali caratteristiche dei pazienti diabetici possano essere indici di rischio per la ricorrenza dell'ipotiroidismo e cercare di comprendere la ragione dietro ai risultati raggiunti così da poter fornire adeguati sistemi di supporto in grado di allertare preventivamente gli individui a rischio.

Altre complicanze del T1D scientificamente provate e presenti nella popolazione in modo significativo, sebbene notevolmente meno frequenti delle due patologie già elencate sono la retinopatia e la neuropatia, la cui relazione con il diabete è già stata approfondita nei paragrafi precedenti.

Al fine quindi di poter sviluppare un sistema di supporto alla decisione clinica in grado di classificare correttamente quali pazienti siano a rischio o meno di ipotiroidismo, si sono prelevati dal dataset Aleppo tutti i tracciati di glicemia dei 51 soggetti con ipotiroidismo, considerati in questo elaborato come complessivamente il gruppo "IPTD". Oltre a questi anche 28 tracciati appartenenti a soggetti con solo T1D al fine di fungere come tracciati di riferimento, questo gruppo sarà chiamato gruppo "controlli". Dei 51 soggetti "IPTD", 38 sono femmine e 13 maschi, inoltre ben 49 pazienti sono bianchi e i restanti due appartengono invece a più di un gruppo etnico; nel gruppo "controlli" invece tutti i soggetti sono bianchi e di questi 18 sono maschi e 10 femmine. Questa distribuzione di valori è concorde con quanto già precedentemente detto: l'ipotiroidismo tende infatti a colpire con maggiore frequenza soggetti di sesso femminile e razza bianca.

Il confronto fra i valori assunti dalle stesse features calcolate in questi due diversi gruppi sarà alla base della progettazione del modello desiderato.

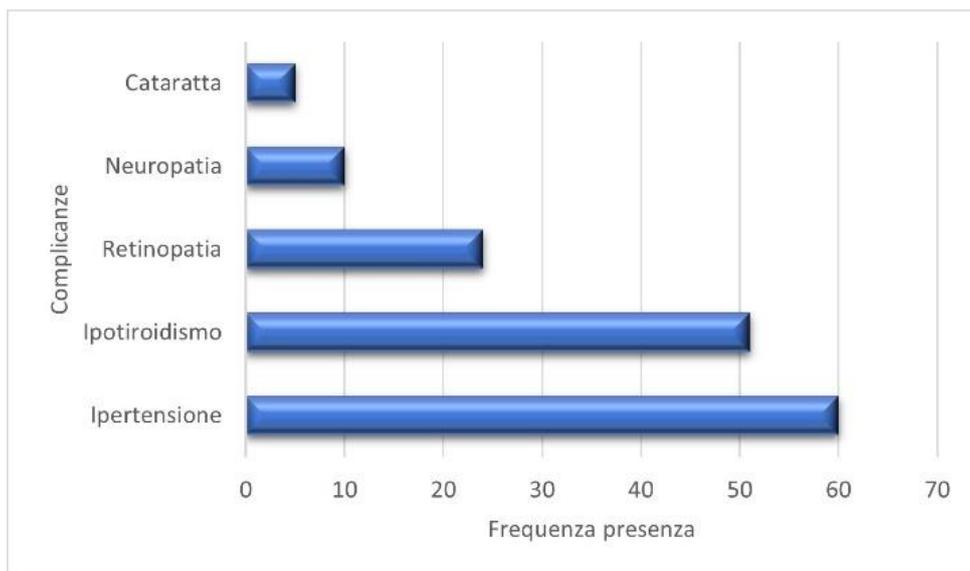


Fig. 2: il grafico a colonne rappresenta con quanta frequenza sono presentati nel dataset Aleppo soggetti con Cataratta, Neuropatia, Retinopatia, Ipotiroidismo e Ipertensione.

4.2) Elaborazione dati

4.2.1) Calcolo delle metriche

Di fondamentale importanza per la realizzazione del CDSS è quindi il calcolo delle metriche relative ai valori di glicemia in ognuno dei tracciati dei controlli e dei IPTD.

L'implementazione di queste features costituisce molto spesso un task delicato, ripetitivo e che richiede molto tempo. È inoltre di fondamentale importanza che tali algoritmi siano standardizzati così da rendere i risultati ottenuti confrontabili con quelli di altri lavori.

Vengono allora utilizzati dei software che rendono automatizzati tali dispendiose implementazioni delle metriche e che, più in generale, aiutano chiunque usufruisca dei CGM ad interpretare i dati. Tra questi, il software con cui si ha scelto di lavorare è AGATA-Automated Glucose dATA Analysis: un toolbox opensource implementato in ambiente Matlab con cui estrarre fino a 42 delle caratteristiche più utilizzate in letteratura [19] [20]. AGATA mette a disposizione anche una interfaccia grafica per agevolare chi non pratico con tali strumenti nel loro utilizzo.

AGATA può essere utilizzato per implementare diverse pipeline di pre-elaborazione e analisi, tra cui la lettura e la preparazione dei dati, l'elaborazione dei dati, l'analisi dei dati e la visualizzazione dei dati.

Indipendentemente dall'utilizzo che vuole essere fatto di AGATA, affinché una qualsiasi delle funzioni messe a disposizione dal software possa essere eseguita fino al termine è necessario che il tracciato in ingresso soddisfi alcuni fondamentali requisiti. Inoltre, osservando i dati selezionati dal database Aleppo, i tracciati presentano complessivamente numerosi buchi di acquisizioni mancanti, come mostra la tabella sottostante (Tab. 1): l'utilizzo effettivo del CGM durante lo studio REPLACE-BG era stato lasciato a discrezione del soggetto quindi durante i sei mesi dello studio molti dei partecipanti non sono stati diligenti nell'uso dello strumento.

Tab. 1: La tabella mostra per ognuno dei due gruppi in esame il numero, in percentuale, di valori mancanti nella traccia glicemica fornita e il numero di giorni di osservazione della traccia glicemica

#	missingGlucosePercentage	numberDaysOfObservation
controlli	15,58	269,82
IPTD	13,66	277,08

Quindi i dati CGM originali e grezzi devono essere prima pre-processati in modo opportuno, gli outliers devono essere eliminati e i dati mancanti riempiti. A questo proposito AGATA mette a disposizione una funzione con cui imputare i dati mancanti usando l'interpolazione lineare, nel caso dello studio in questione la funzione è stata utilizzata solo per riempire buchi temporali fino ad un massimo di 30 minuti: riempiendo anche buchi maggiori si rischia di falsificare l'andamento complessivo della glicemia rendendo poi nullo tutto lo studio successivo. L'intervallo di 30 minuti è stato scelto perché un buon compromesso tra il riempimento e la corruzione del tracciato ed è per questo molto usato negli studi sul controllo della glicemia presenti già in letteratura [20].

Un'ulteriore pre-elaborazione che è stata effettuata sui dati riguarda la forma in cui questi tracciati possono essere dati in ingresso alle funzioni di AGATA: il software infatti lavora solo con ingressi nella forma timetables con un intervallo temporale omogeneo e glicemia espressa in mg/dL. È stato necessario allora trasformare i dati e ricampionare i tracciati per far sì che tra ogni acquisizione vi sia esattamente un intervallo temporale di 5 minuti, anche in questo caso si è usufruito di due funzioni apposite messe a disposizione dal software stesso. Le due funzioni risultano essere molto utili e facilmente applicabili con, però, l'importante difetto di eliminare dalle timetables dei singoli soggetti l'informazione inerente all'ID: permettono infatti di ottenere in uscita delle timetables con solo le due colonne inerenti alla data e ora dell'acquisizione e al valore glicemico corrispondente.

Una volta che i tracciati sono stati opportunamente preparati è infine possibile calcolare le metriche d'interesse: come precedentemente anticipato inizialmente sono state calcolate tutte e solo poi confrontando i risultati ottenuti nei due gruppi sono state preservate solo quelle che presentavano differenze statistiche significative tra i controlli e i IPTD.

Nella tabella di seguito (Tab. 2) sono riportate le definizioni di alcune delle features.

Tab. 2: La seguente tabella mostra elencati alcuni dei più importanti parametri ricavati dal tracciato di CGM e riporta le corrispondenti definizioni

Metric	Description
aucGlucose	<i>Area under the glyceimic curve</i>
CVGA	<i>Control variability grid analysis</i>
cogi	<i>Continuous glucose monitoring index</i>
cvGlucose	<i>Coefficient of variation of glucose trace</i>
efIndex	<i>Excursion frequency, corresponding to the sum of all excursions</i>
gmi	<i>Glucose management indicator</i>
iqrGlucose	<i>Interquartile range of glucose</i>
jIndex	<i>Measure of both the mean level and variability of glyceimic</i>
mageIndex	<i>Mean amplitude of glyceimic excursions</i>
meanGlucose	<i>Mean glucose over all days</i>
medianGlucose	<i>Median glucose over all days</i>
rangeGlucose	<i>Difference between max and min glucose over all days</i>
sddmIndex	<i>Standard deviation of within-day means index</i>
sdwIndex	<i>Mean of within-day standard deviation index</i>
stdGlucose	<i>Standard deviation of glucose trace</i>
conga	<i>Continuous overall net glyceimic action</i>
modd	<i>Mean of daily differences in glucose</i>
stdGlucoseROC	<i>Standard deviation of the rate of change of glucose</i>
adrr	<i>Average daily risk range, assessment of total daily glucose variations within risk space</i>
bgri	<i>Blood glucose risk index of the glucose concentration</i>
hbgi	<i>High blood glucose index</i>
lbgi	<i>Low blood glucose index</i>
gri	<i>Glyceimic Risk Index proposed by Klonoff et al.</i>
timeInHyperglycemia	<i>Percentage of time spent in hyperglycemia</i>
timeInHypoglycemia	<i>Percentage of time spent in hypoglycemia</i>
timeInTarget	<i>Percentage of time spent in the target range</i>
gradeScore	<i>Glyceimic risk assessment in diabetes equation</i>
hypoIndex	<i>Hypoglycemia index</i>
hyperIndex	<i>Hyperglycemia index</i>
igc	<i>Index of glyceimic control, sum of hyper index and hypo index</i>
mrIndex	<i>Measure of variation of glucose values around a reference value</i>

Come anticipato, il software implementa una serie di funzioni per pre-elaborare i dati sul glucosio, calcolare le metriche di controllo, rilevare gli eventi avversi e visualizzare i risultati. Inoltre, AGATA include diverse funzioni per calcolare le metriche relative all'errore di predizione tra i dati originali e i profili del glucosio previsti, eseguire l'analisi automatica di un dato set di dati e confrontare due bracci di uno studio. In questo elaborato, ad esempio, la funzione usata per il calcolo delle metriche per ognuno dei due gruppi controlli e IPTD è la funzione AnalyzeOneArm.

In particolar modo le features ricavabili dai tracciati ottenuti con CGM possono essere divise in quattro classi principali a seconda dell'aspetto del tracciato glicemico che vanno ad analizzare. Le features possono quindi essere relative alla variabilità dei tracciati, appartengono infatti a questa categoria le ben note media, mediana, deviazione standard etc; oppure i parametri possono essere time-related ovvero descrivere, ad esempio, per quanto tempo il tracciato si trova all'interno di un determinato range di valori glicemici target. Le caratteristiche possono anche essere di tipo risk, metriche che permettono di quantificare i fattori di rischio per il soggetto con DMT1 come, ad esempio, l'indice di bassa glicemia lbgi; o, infine, possono appartenere alla classe delle Glycemic Transformation: ovvero richiedere una trasformazione, solitamente logaritmica, del dato affinché possa esservi estrapolato il contenuto informativo.

Nonostante le descrizioni delle metriche calcolate con AGATA coincidano con quelle fornite da anche tutti gli altri software per l'analisi dei dati provenienti da CGM, come si può notare dalla Tabella 2, nella pratica i singoli valori con i quali vengono definite le soglie per, ad esempio, individuare le zone di iperglicemia o ipoglicemia differiscono da quelli individuati dai toolbox di maggior rilevanza nella comunità scientifica come iglu.

Le features che presentavano nel loro algoritmo di definizione queste incongruenze con iglu sono state quindi corrette e ricalcolate così da garantire uniformità con altre pubblicazioni scientifiche del campo.

4.2.2) Scelta delle metriche

Al fine dello sviluppo di un efficiente CDSS, come già accennato nell'apposito paragrafo, è di fondamentale importanza inserire in ingresso alla macchina un numero adeguato, e non eccessivo, di features così da non confondere e portare fuori strada lo strumento con informazioni superflue o ridondanti.

Ne deriva quindi che sia sempre buona pratica studiare attraverso l'utilizzo di tecniche statistiche la distribuzione dei valori di features ottenuti e le varie relazioni tra di esse, così da poter individuare le caratteristiche significative per lo studio in esame: quest'ultime, infatti, conterranno al proprio interno maggiore potere informativo. Può essere, allora, rilevante clinicamente soffermarsi sui possibili legami tra la patologia in esame, l'ipotiroidismo, e il significato delle features selezionate alla fine dell'indagine statistica così da riuscire a trovare legami profondi e innovativi tra le due.

L'analisi delle features avrà quindi l'obiettivo di individuare e, se ritenuto opportuno, ignorare tutte le metriche ritenute ridondanti o superflue. In particolar modo la loro ridondanza viene valutata attraverso il calcolo dei coefficienti di regressione tra le metriche in esame mentre il poco contenuto informativo viene valutato attraverso test statistici come il t-test che permettono di determinare quali features presentino valori significativamente diversi nel caso del gruppo controlli e quello IPTD.

E' di fondamentale importanza infatti che vi sia un numero apprezzabile di features che assumono valori significativamente differenti tra i due gruppi perché se così non fosse statisticamente i due gruppi, i malati di solo DMT1 e quelli malati di DMT1 e ipotiroidismo, non sarebbero distinguibili, rendendo vano lo sviluppo di un sistema CDSS per l'individuazione dei soggetti DMT1 a rischio di ipotiroidismo.

Per avere una idea dell'entità delle differenze tra i due gruppi il software Agata mette a disposizione delle utili funzioni che permettono di visualizzare graficamente alcuni tracciati. Il primo tracciato, ottenuto grazie alla funzione `plotGlucoseArmAsOneDayComparison`, genera un singolo grafico contenente i due profili glicemici, relativi ai due gruppi IPTD e controlli, dove ciascun profilo glicemico giornaliero è stato ottenuto sovrapponendo quello di ogni soggetto del relativo gruppo. Sono stati poi graficati anche i valori di GRI e CVGA dei soggetti dei gruppi controlli e IPTD, sfruttando ancora una volta le apposite funzioni di Agata `plotCVGAComparison` e `plotGRIComparison`.

Conclusa l'analisi visiva dei tracciati, si è provveduto alla determinazione rigorosa delle differenze statistiche tra i due gruppi così da accertarsi in modo definitivo della effettiva distinzione, dal punto di vista glicemico, dei due gruppi.

Per poter scegliere il test statistico più adatto alle singole features, è stato prima necessario distinguere le metriche tra normali, o gaussiane, e non normali; ciò è stato fatto grazie all'applicazione del test Lilliefors su ogni caratteristica per ognuno dei due gruppi.

Le features degli IPTD risultate normali sono state confrontate con le corrispondenti dei controlli attraverso il t-test, mentre a quelle non normali è stato applicato il Wilcoxon rank sum test, l'intervallo di confidenza che si ha deciso di utilizzare è del 95%.

E' buona norma, soprattutto in un caso complesso come quello in esame, dare in ingresso all'algoritmo di machine learning anche le restanti features non significative: ciò viene fatto poiché la macchina potrebbe riuscire a trovare, tramite i propri algoritmi di apprendimento, delle

combinazioni di features significative per lo studio o scovare legami nascosti tra le features e i soggetti con ipotiroidismo che con i soli test statistici non erano evidenti.

Si possono, invece, eliminare dall'ingresso del software di ML le metriche che portano con sé informazioni considerate ridondanti. Per far ciò, come già accennato, è stata studiata la correlazione tra ogni metrica e tra le metriche e la variabile di label (di valore "controllo" o "IPTD"). Le features con coefficiente di correlazione in modulo maggiore o uguale a 0,98 sono state considerate come coincidenti, infatti hanno lo stesso contenuto informativo poiché al variare di una varia in modo proporzionale anche l'altra, per decidere quindi quale mantenere e quale trascurare può essere utile osservare se una delle due presenti una correlazione significativa con la label e, in caso affermativo, conservare quella metrica e tralasciare l'altra.

4.2.3) Creazione modello per la classificazione

Una volta stabilite le features con cui lavorare, le successive elaborazioni dei dati, precedentemente svolte in ambiente Matlab, sono d'ora in poi state eseguite in Orange 3.30.2, un toolkit interattivo open-source basato su Python per machine learning, data mining e visualizzazione dei dati che presenta un front-end di programmazione grafico e intuitivo.

Alla fine del lavoro svolto in questo ambiente si ha ottenuto il workspace mostrato nella figura sottostante (Fig. 6): ogni widget grafico presente rappresenta una operazione di elaborazione svolta sui dati in ingresso.

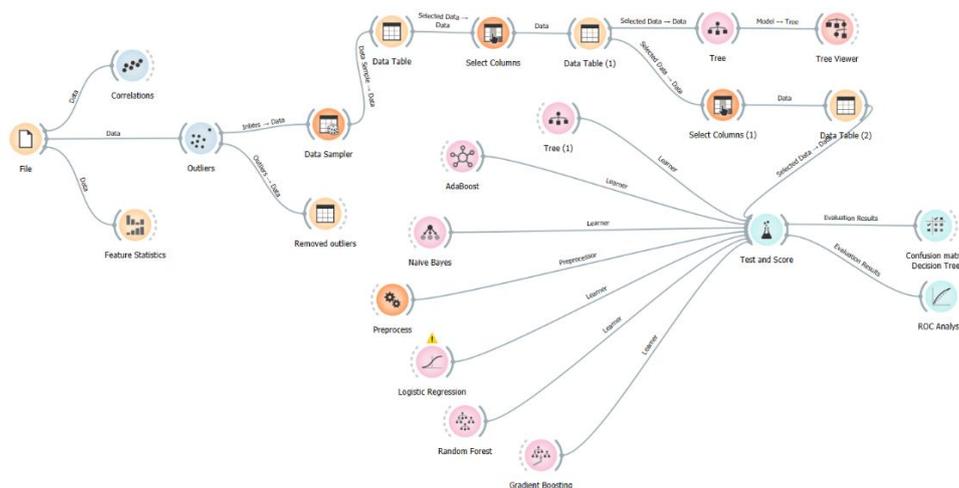


Fig. 6: la figura mostra il workflow ottenuto su Orange alla fine della elaborazione

Come è possibile notare dalla immagine sovrastante (Fig. 6), in un primo momento i dati caricati sono stati ulteriormente ispezionati in modo da individuare i soggetti considerati non rappresentativi del loro gruppo (ovvero gli outliers) così da eliminarli dal dataset in esame: dei 79 soggetti coinvolti nello studio 7 sono stati considerati outliers, e di questi 4 appartenenti al gruppo malati e i restanti 3 ai controlli. Il modello è stato quindi poi allenato e testato con un database totale di 72 soggetti.

Un altro passo fondamentale e preliminare all'utilizzo dei modelli per la classificazione è costituito dalla selezione delle caratteristiche basata sull'albero decisionale, così facendo sono state determinate le caratteristiche predittive più rilevanti suddividendo i dati in partizioni più omogenee rispetto alla classe.

Un albero decisionale è un modello di apprendimento automatico supervisionato non parametrico e rappresenta una struttura gerarchica a forma di albero; infatti, ogni albero decisionale inizia con un nodo radice, che non ha rami in entrata ma solo in uscita, questi ultimi alimentano i nodi interni, noti anche come nodi decisionali. Sia il nodo radice che quelli interni conducono valutazioni per formare sottoinsiemi omogenei, che sono rappresentati da nodi foglia o nodi terminali: questi, infatti, costituiscono tutti i possibili risultati all'interno del set di dati [21].

L'apprendimento dell'albero decisionale prevede l'utilizzo di una strategia "dividi et impera" applicata conducendo una ricerca per identificare i punti di divisione ottimali all'interno di un albero. Questo processo di suddivisione viene quindi ripetuto in modo ricorsivo dall'alto verso il basso fino a quando tutti o la maggior parte dei record sono stati classificati in etichette di classe specifiche. Il fatto che tutti i punti dati siano classificati o meno come insiemi omogenei dipende in gran parte dalla complessità dell'albero decisionale: gli alberi più piccoli sono più facilmente in grado di raggiungere nodi foglia semplici.

Man mano che un albero cresce di dimensioni diventa sempre più difficile mantenere questa semplicità ciò di solito si traduce in una quantità insufficiente di dati all'interno di un determinato albero secondario: questo fenomeno è noto come frammentazione dei dati e spesso può portare a un overfitting. Di conseguenza, gli alberi decisionali prediligono strutture piccole, infatti è buona pratica per ridurre la complessità e prevenire l'overfitting, utilizzare la potatura: un processo che rimuove i rami che si dividono in caratteristiche di bassa importanza, ovvero con basso potere decisionale.

La bontà della suddivisione attuata dal modello viene solitamente valutata attraverso il rapporto di guadagno di informazione. Il rapporto di guadagno di informazione (Information Gain Ratio)

viene infatti utilizzato per valutare quale attributo o caratteristica di un insieme di dati è più informativo nella separazione delle classi. [22]

La formula del rapporto di guadagno di informazione {1} è data da:

$$Gain\ Ration = \frac{Information\ Gain}{Split\ Information} \quad \{1\}$$

dove l'Information Gain rappresenta la riduzione di entropia o impurità nei dati ottenuta dalla suddivisione rispetto a un particolare attributo. Maggiore è l'Information Gain, maggiore è la capacità dell'attributo di separare le classi. La Split Information invece misura la distribuzione uniforme delle istanze nei sottoinsiemi generati dall'attributo: minore è questo valore più sarà definita la separazione dei gruppi.

Il rapporto di guadagno di informazione aiuta a mitigare il problema legato all'utilizzo della sola Information Gain, questa infatti tende a favorire attributi con un gran numero di valori distinti. Dividendo tale valore per la Split Information, invece, sono penalizzati gli attributi con molti valori distinti o branch molto numerosi, contribuendo così a mantenere una maggiore generalizzazione nel calcolo della Gain Ration.

Nel caso in esame, per costituire l'albero decisionale sono stati imposti i seguenti parametri:

- "Numero minimo di istanze nel nodo terminale" = 2, che include almeno 2 campioni di allenamento; "Non suddividere sottoinsiemi più piccoli di" = 10, che vieta la suddivisione dei nodi terminali con meno di 10 istanze; "Limitare la profondità massima dell'albero a" = 10, che limita la distanza tra la radice e il nodo terminale più lontano. Inoltre, è stata selezionata l'opzione "Induci albero binario", che suddivide solo in due nodi, e "Arresta quando la maggioranza raggiunge il 90%" è stata impostata in modo che l'algoritmo si fermi quando raggiunge il 90% degli esempi classificati.

Il modello ad albero così definito è stato poi testato insieme ad altri cinque modelli per determinare dal confronto tra questi quello più adatto allo studio. Tutti i modelli che si ha scelto di analizzare e testare sono ampiamente usati in letteratura per problemi di classificazione come quello in esame.

Il set di caratteristiche finale così ottenuto è stato quindi ulteriormente valutato fornendolo come input ai seguenti algoritmi di classificazione: Random Forest, Adaboost e Gradient Boosting oltre che i più famosi, Naive Bayes e Logistic Regression.

Descrivendo brevemente ognuno di questi algoritmi, il Random Forest combina l'output di più strutture ad albero decisionali per raggiungere un unico risultato complessivamente migliore, si basa quindi sui concetti fondamentali del modello ad albero decisionale già illustrato [23].

L'idea invece dietro all'algoritmo di AdaBoost (da Adaptive Boosting) è quella di costruire una sequenza di modelli di apprendimento debole e combinare i loro risultati per creare un modello forte. AdaBoost, infatti, costruisce una sequenza di classificatori deboli, ovvero dei modelli che hanno prestazioni leggermente migliori di un classificatore casuale ma non alte; ognuno dei weak learners si concentra sui casi che sono stati classificati erroneamente dai modelli precedenti e viene quindi addestrato per correggere gli errori dei modelli precedenti.

Ad ogni classificatore viene assegnato un peso in base alla sua accuratezza, i più accurati ricevono pesi più alti, mentre classificatori meno accurati ricevono pesi più bassi. I classificatori vengono poi combinati per creare un modello forte, la combinazione tiene conto dei pesi assegnati ai classificatori dando maggior peso a quelli più accurati. Questo modello forte è quindi utilizzato per fare previsioni sulle nuove istanze.

Come Adaboost anche Gradient Boosting è una tecnica di apprendimento automatico che è progettata per migliorare le prestazioni dei modelli di apprendimento automatico normalmente più deboli, come l'albero decisionale o il modello lineare. Il suo principio di funzionamento è molto simile a quello precedentemente descritto ed è basato su algoritmi ricorsivi per diminuire, modello dopo modello, gli errori residui tra le previsioni del modello e i valori effettivi del target.

Naive Bayes, invece, è un modello di classificazione che si basa sul Teorema di Bayes, un teorema statistico che descrive come aggiornare le probabilità di una serie di ipotesi in luce di nuovi dati o evidenze acquisite. È chiamato "naive" (ingenuo) perché assume che le caratteristiche utilizzate per la classificazione siano indipendenti tra loro, anche se in realtà potrebbero non esserlo.

Il modello calcola, utilizzando i dati dell'addestramento, le probabilità condizionali di ciascuna caratteristica dato un determinato target di classe. Per classificare una nuova istanza, il modello moltiplica le probabilità condizionali associate a ciascuna caratteristica per ottenere la probabilità complessiva di quella combinazione di caratteristiche dato il target di classe. La classe che massimizza la probabilità condizionale, data la combinazione di caratteristiche, è la classe assegnata all'istanza.

Infine, il tipo modello più indicato per il problema di classificazione in esame, oltre all'Albero Decisionale, è l'algoritmo di Logistic Regression. La regressione logistica è infatti un modello

statistico utilizzato per la classificazione binaria, che permette di assegnare un'istanza, un soggetto nel caso di questo elaborato, a una di sole due classi possibili.

La regressione logistica utilizza la funzione sigmoide per modellare la probabilità di appartenenza a una classe di un soggetto. La funzione logistica {2} è definita come:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}} \quad \{2\}$$

dove $P(Y=1)$ è la probabilità appartenenza alla classe positiva (nello studio in questione IPTD) e X_1, X_2, \dots, X_n sono le variabili indipendenti.

Durante la fase di addestramento del modello vengono appresi i valori ottimali dei coefficienti β . Una volta addestrato, il modello produce una curva di decisione che separa le due classi nello spazio delle caratteristiche; la curva di decisione è definita dal valore predetto della probabilità di appartenenza a una classe rispetto alle variabili indipendenti. Ogni istanza viene quindi classificata assegnando una classe in base alla posizione rispetto alla curva di decisione; tipicamente, si utilizza 0,5 come soglia per determinare quale classe assegnare all'istanza.

L'ottimizzazione degli iperparametri dei modelli è stata eseguita seguendo quella utilizzata in altri studi inerenti all'ambito del diabete [17] e presenta i seguenti valori:

- Random Forest: "Numero di alberi" = 25, "Numero di attributi considerati ad ogni suddivisione" = 2, "Non suddividere sottoinsiemi più piccoli di" = 5, "Limitare la profondità degli alberi individuali" = 15;
- Adaboost: "Numero di stimatori" = 100, "Tasso di apprendimento" = 1, algoritmo di classificazione SAMME.R come metodo di boosting che aggiorna i pesi dell'estimatore di base con le stime di probabilità, e selezione di "Esponenziale" come funzione di perdita di regressione;
- Gradient Boosting: "Numero di stimatori" = 100, "Tasso di apprendimento" = 0.3, "Lambda" = 0.05, "Limitare la profondità degli alberi individuali" = 3, "Frazione di funzionalità per ogni albero" = 1 e catboost come metodo. Tutte le caratteristiche sono state normalizzate su una scala comune (media uguale a zero e varianza unitaria).
- Logistic Regression: Regularization type Ridge L2, Strength Strong, C=0,001.

Tutti gli algoritmi di classificazione sono poi stati convalidati mediante la procedura di cross-validation con 5 folds: la tecnica consiste nel dividere i dati in cinque fold; l'algoritmo viene testato trattenendo i dati di uno dei fold alla volta, il modello viene indotto dagli altri fold e i dati contenuti

nel fold trattenuto vengono classificati. Questo processo viene ripetuto per tutti i fold e i risultati ottenuti alla fine sono una media dei risultati di ogni prova di classificazione [24].

I risultati ottenuti dal testing and score dei dati selezionati in ingresso ai vari modelli precedentemente elencati e i risultati di tutte le precedenti elaborazioni descritte saranno mostrati nel dettaglio nel paragrafo a seguire.

4.3) Risultati e Discussioni

Per prima cosa, quindi, sono state calcolate le metriche descrittive dei tracciati glicemici ottenuti con CGM, i risultati ottenuti per entrambi i gruppi, controlli e IPTD, sono riportati nella tabella seguente (Tab. 3) così da facilitare il confronto dei valori ottenuti.

Tab. 3: La tabella mostra i valori assunti da tutte le metriche calcolate con AGATA per ognuno dei due gruppi. In blu sono rappresentate le features relative alla variability, in arancione quelle relative al risk, in verde quelle relative al time e in grigio le glycemieTransformation.

#	controlli	IPTD
<i>mageIndex (mmol/L)</i>	(125,84+27,71)	(140,05+42,08)
<i>stdGlucoseROC (mmol/(L *min))</i>	(1,03 +0,13)	(0,94+0,14)
<i>congA (mmol/L)</i>	(45,05+ 6,71)	(40,80+6,92)
<i>modd (mmol/L)</i>	(69,11+ 11,31)	(60,92+12,15)
<i>aucGlucose ((mmol/L)*min)</i>	(53373537,65 +- 1,23e+07)	(53849646,88+-1,17e+07)
<i>CVGA</i>	(8889,44 +- 2,75e+03)	(8674,34+-4,56e+03)
<i>cogi</i>	(60,81+- 9,62)	(66,98+-8,22)
<i>cvGlucose</i>	(39,45+-10,23)	(37,11+-5,11)
<i>efIndex (mmol/L)</i>	(3,87 +- 0,94)	(3,78+-0,80)
<i>gmi</i>	(7.35 +- 0,56)	(7,13+-0,47)
<i>iqrGlucose (mmol/L)</i>	(90,00+-25,00)	(79+-22,50)
<i>jlIndex</i>	(56,70 +- 17,44)	(48,80+-12,41)
<i>magePlusIndex (mmol/L)</i>	125,76 +- 24,39)	(117,20+-28,72)
<i>mageMinusIndex (mmol/L)</i>	(122,83 +- 21,97)	(115,74+-28,71)
<i>meanGlucose (mmol/L)</i>	(168,75+- 23,28)	(159,58+-19,56)
<i>medianGlucose (mmol/L)</i>	(155,50+-18,00)	(153+-22,00)
<i>rangeGlucose (mmol/L)</i>	(376,75+-54)	(366+-26,38)
<i>sddmIndex</i>	(29,96+-7,97)	(28,49+-7,27)
<i>sdwIndex</i>	(56,90+- 8,61)	(50,47+-9,59)
<i>stdGlucose</i>	(67,18 +- 10,83)	(59,50+- 10,88)
<i>addr</i>	(47,26+-8,01)	(41,27+-8,78)
<i>bgri</i>	(10,23+- 4,15)	(8,30+-2,76)
<i>hbgi</i>	(8,23+-2,99)	(7,45+-4,14)
<i>lbgi</i>	(0,98+-0,47)	(0,85+-0,91)
<i>gri</i>	(48,66+-16,54)	(42,44+-15,75)
<i>timeInHyperglycemia</i>	(37,84+-11,15)	(33,50+-18,55)
<i>timeInL1Hyperglycemia</i>	(24,58+-5,14)	(24,08+-7,78)
<i>timeInL2Hyperglycemia</i>	(13,22+- 9,84)	(9,10+-6,07)
<i>timeInHypoglycemia</i>	(3,71+-2,20)	(3,21+-4,00)
<i>timeInL1Hypoglycemia</i>	(2,80+-1,82)	(2,43+-2,37)
<i>timeInL2Hypoglycemia</i>	(1,26+- 1,26)	(0,88+-0,78)
<i>timeInTarget</i>	(59,01+-14,54)	(62,12+-16,67)
<i>timeInTightTarget</i>	(35,50+-8,22)	(39,98+-11,69)
<i>gradeScore</i>	(10,10+-2,78)	(8,77+-2,33)
<i>gradeEuScore</i>	(6,71+-3,92)	(7,65+-5,17)
<i>gradeHyperScore</i>	(88,68+-5,32)	(88,03+-7,83)
<i>gradeHypoScore</i>	(4,00+-3,02)	(3,12+-5,26)
<i>hypoIndex</i>	(0,71+-0,40)	(0,61+-0,66)
<i>hyperIndex</i>	(2,06+-0,81)	(1,84+-1,09)
<i>igc</i>	(2,90+-1,14)	(2,31+-0,82)
<i>mriIndex</i>	(32,69+-13,02)	(29,17+-16,47)

In seguito al calcolo delle features è stato plottato il tracciato del profilo giornaliero complessivo dei due gruppi, il grafico ottenuto è riportato nella figura sottostante (Fig. 3).

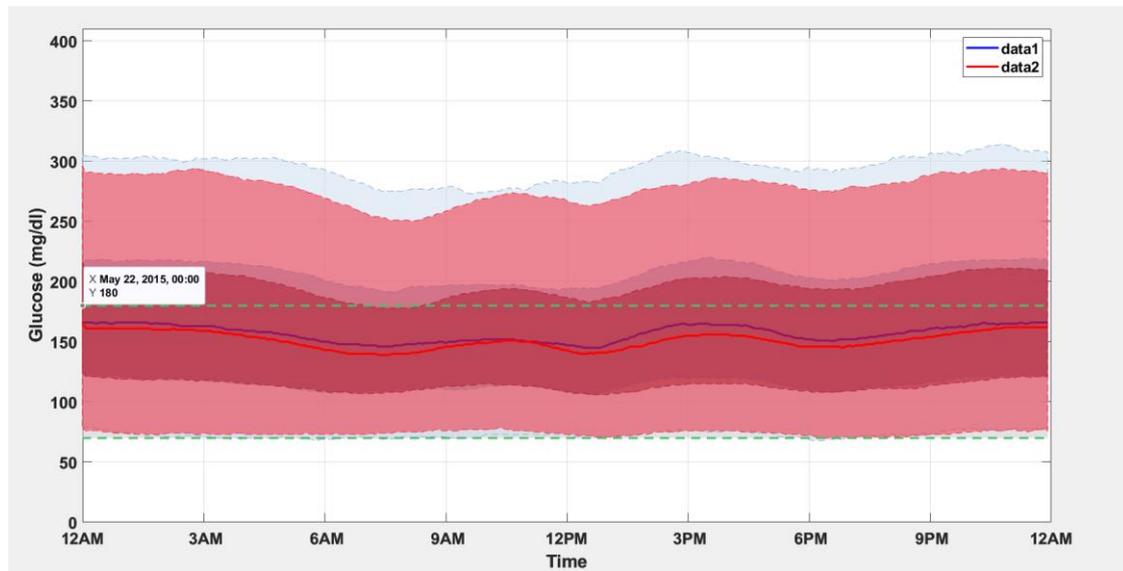


Fig. 3: Il profilo giornaliero mediano è rappresentato dalla linea in grassetto rispettivamente rossa e viola per i controlli e i IPTD, le aree blu scuro e rosa scuro evidenziano i percentili dal 25° al 75° dei rispettivi gruppi, mentre le aree rosa chiaro e blu chiaro i percentili dal 5° al 95°

Le curve di glicemia giornaliera presentano il loro tipico andamento: possono infatti essere individuati i punti di minima proprio poco prima del pasto e di massima poco dopo; infatti, a seguito dell'assunzione di cibo i livelli di glucosio nel sangue aumentano traducendosi, graficamente, in un importante aumento della pendenza della curva.

I soggetti in esame mostrano complessivamente un andamento della glicemia ben controllato e generalmente dentro il target range, si può apprezzare, allora, l'efficacia della terapia insulinica da loro effettuata. I picchi positivi, infatti, raramente superano i 200 mg/gL e quelli negativi non risultano mai in crisi ipoglicemiche (valori di glicemia più bassi di 75 mg/dL).

Per quanto riguarda il confronto tra i due gruppi, a prima vista i due tracciati risultano quasi coincidenti: è subito lampante quindi che la patologia dell'ipotiroidismo, per quanto sia una complicanza importante e debilitante per chi ne è affetto, non sembra influire in modo evidente sui soggetti. In casi come questi, quindi, risulta impossibile sviluppare con le sole tecniche statistiche degli algoritmi per la classificazione dei soggetti nei due gruppi, diventa quindi necessario l'intervento dell'intelligenza artificiale per cercare di trovare dei legami nascosti tra le features e i due gruppi.

Risultati simili si ottengono analizzando i grafici del confronto tra le distribuzioni dei valori delle features CVGA e GRI, sempre forniti da Agata, nei gruppi controlli e malati (Fig. 4 e Fig. 5).

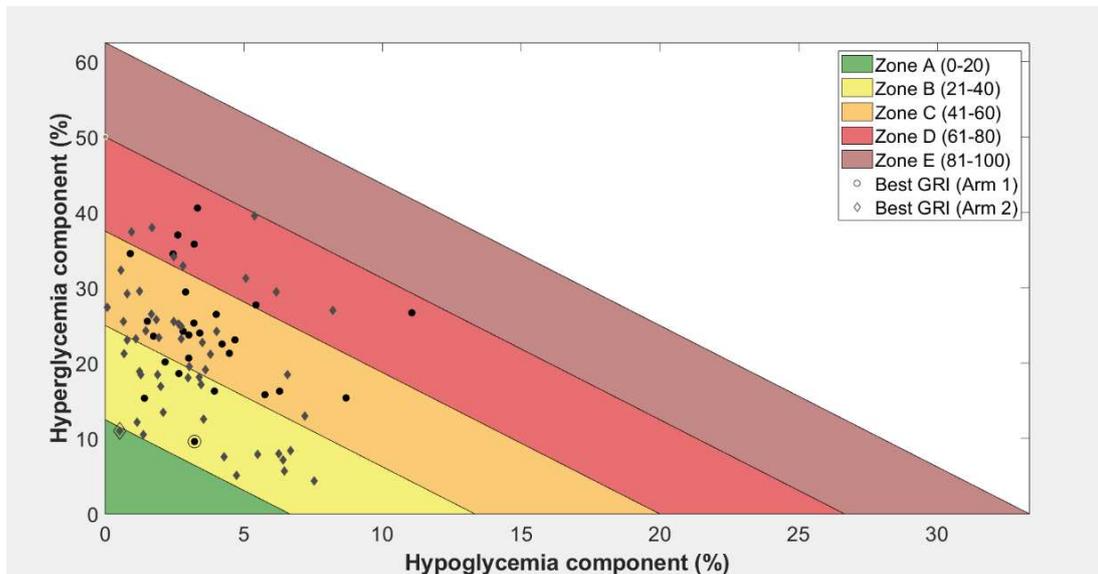


Fig. 4: La figura mostra il confronto tra i valori di GRI assunti dai soggetti del gruppo controlli, rappresentati dai cerchi neri, e i soggetti del gruppo IPTD, rappresentati dai diamanti neri

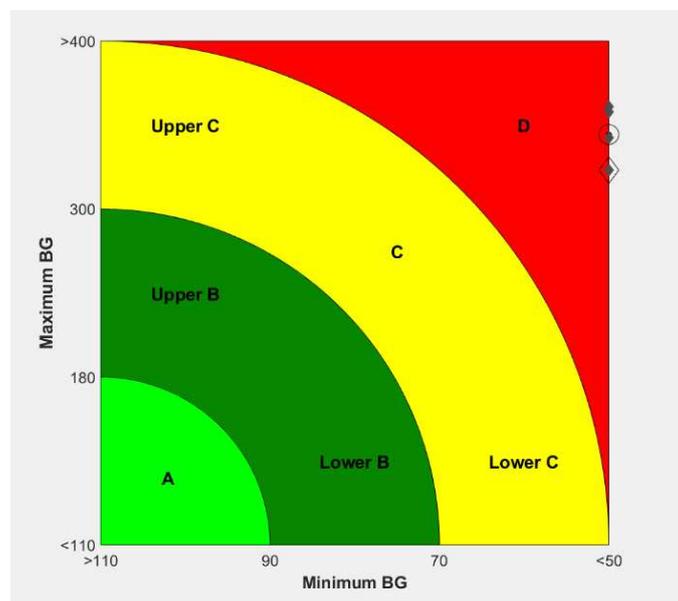


Fig. 5: La figura mostra il confronto tra i valori di CONGA assunti dai soggetti del gruppo controlli, rappresentati dai cerchi neri, e i soggetti del gruppo IPTD, rappresentati dai diamanti neri

Potrebbe quindi sorgere il dubbio che i soggetti con solo DMT1 e quelli con anche ipotiroidismo dal punto di vista glicemico siano identici e che, di conseguenza, lo studio che si sta effettuando sia inutile; fortunatamente continuando l'analisi e utilizzando tecniche statistiche rigorose per quantificare le differenze tra i due gruppi questa ipotesi viene smentita.

Applicando i test statistici t-test e Wilcoxon rank sum test, i risultati ottenuti indicano infatti che, delle 19 metriche con distribuzione normale in entrambi i gruppi, 10 sono diverse in modo statisticamente significativo nei due gruppi mentre delle restanti 22 metriche non normali, 7 sono statisticamente significative. (Tab. 4)

Tab. 4: La tabella mostra le features che assumono valori diversi in modo statisticamente significativo tra i due gruppi, controlli e IPTD, ed il corrispondente valore di p-value.

#	controlli	malati	p-value
<i>jIndex</i>	(56,70 +- 17,44)	(48,80+-12,41)	0,022022686
<i>stdGlucoseROC (mmol/(L*min))</i>	(1,03 +-0,13)	(0,94+-0,14)	0,01256567
<i>sdwIndex</i>	(56,90+- 8,61)	(50,47+-9,59)	0,004181387
<i>stdGlucose</i>	(67,18 +- 10,83)	(59,50+- 10,88)	0,003541582
<i>cong (mmol/L)</i>	(45,05+- 6,71)	(40,80+-6,92)	0,010009111
<i>modd (mmol/L)</i>	(69,11+- 11,31)	(60,92+-12,15)	0,004407817
<i>cogi</i>	(60,81+- 9,62)	(66,98+-8,22)	0,002295451
<i>cvGlucose</i>	(39,45+-10,23)	(37,11+-5,11)	0,004902815
<i>iqrGlucose</i>	(90,00+-25,00)	(79+-22,50)	0,001308191
<i>gri</i>	(48,66+-16,54)	(42,44+-15,75)	0,007359716
<i>timeInTarget</i>	(59,01+-14,54)	(62,12+-16,67)	0,044017025
<i>gradeEuScore</i>	(6,71+-3,92)	(7,65+-5,17)	0,02941254
<i>addr</i>	(47,26+-8,01)	(41,27+-8,78)	0,003756292
<i>bgri</i>	(10,23+- 4,15)	(8,30+-2,76)	0,015717693
<i>timeInL2Hyperglycemia</i>	(13,22+- 9,84)	(9,10+-6,07)	0,024019489
<i>gradeScore</i>	(10,10+-2,78)	(8,77+-2,33)	0,026703668
<i>igc</i>	(2,90+-1,14)	(2,31+-0,82)	0,002907664
<i>timeInTarget</i>	(59,01+-14,54)	(62,12+-16,67)	0,044017025
<i>gradeEuScore</i>	(6,71+-3,92)	(7,65+-5,17)	0,02941254
<i>igc</i>	(2,90+-1,14)	(2,31+-0,82)	0,002907664
<i>cong (mmol/L)</i>	(45,05+- 6,71)	(40,80+-6,92)	0,010009111
<i>modd (mmol/L)</i>	(69,11+- 11,31)	(60,92+-12,15)	0,004407817

Si può allora affermare che lo studio che si sta effettuando sia sensato: per quanto simili siano i due gruppi, questi non sono però del tutto coincidenti e vi sono infatti 17 features su cui indagare che potenzialmente potrebbero essere utili per la classificazione dei soggetti con DMT1 nei due gruppi.

Successivamente l'analisi statistica delle metriche è continuata con il calcolo del coefficiente di correlazione tra i parametri. Alla fine di questo studio sono state eliminate 16 features poiché fortemente correlate ad altre features ritenute più importanti, o di più facile interpretazione, che sono state conservate. La label, invece, non ha presentato un coefficiente di correlazione maggiore di 0,4 in modulo con nessuna features, sebbene si possa notare in generale un trend di valori in segno negativi del coefficiente. Questo risultato è concorde con quanto già osservato: non esiste una sola metrica che possa permettere di discernere con chiarezza i soggetti di controllo da quelli con IPTD.

Utilizzando invece gli algoritmi offerti dal toolkit ORANGE, è stata possibile una ulteriore selezione delle features ritenute fondamentali per la classificazione. Grazie all'utilizzo dell'albero decisionale delle 25 features individuate con lo studio sulla correlazione solo 8 sono state ritenute necessarie per permettere l'implementazione della classificazione, come mostrato nella figura sottostante (Fig. 7).

Le otto features selezionate sono: sdwIndex, aucGlucose, jIndex, mageIndex, igc, cogi, conga e stdGlucoseROC. È interessante notare come, ad eccezione di igc, tutte le altre metriche individuate dall'albero come significative per la classificazione appartengono al macrogruppo di features relative alla variabilità dei tracciati glicemici, sebbene dal grafico precedentemente visualizzato la variabilità dei due gruppi sembrasse quasi coincidente. Potrebbe allora essere utile, in uno studio più approfondito, indagare su questa possibile relazione tra variabilità del tracciato e ipotiroidismo nei soggetti con DMT1.

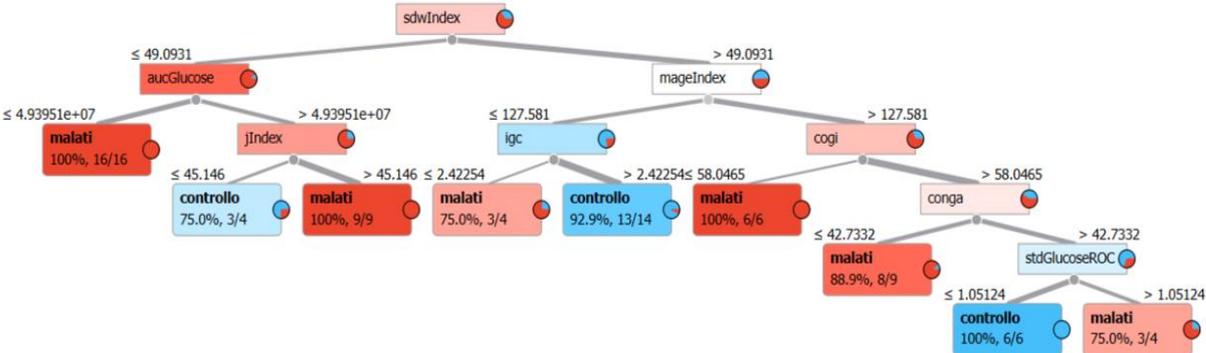


Fig. 7: La figura mostra il modello implementato dall'albero decisionale per la classificazione dei soggetti, partendo dal loro tracciato glicemico, nei gruppi di controllo e quello IPTD. Si possono individuare facilmente le 8 features necessarie per il modello.

Nonostante le features individuate dall'albero decisionale siano solo 8 è già possibile notare visivamente come il problema della classificazione nei due gruppi sia alquanto complesso: infatti non si può apprezzare alcuna ben definita suddivisione tra i soggetti di controllo e quelli malati. Questo risultato è coerente con quanto già notato dall'analisi dei grafici precedenti: sebbene i due gruppi non siano esattamente identici presentano comunque numerose somiglianze nei valori delle features e poche ben definibili distinzioni.

Successivamente questo modello appena mostrato e gli altri cinque modelli ottenuti applicando gli algoritmi di Random Forest, Adaboost e Gradient Boosting, Naive Bayes e Logistic Regression, sono stati addestrati e testati.

Considerando la classe IPTD come positiva e la classe controlli come negativa, i casi IPTD correttamente classificati dal modello di classificazione sono stati considerati veri positivi (TP); i casi controlli correttamente classificati sono stati considerati veri negativi (TN); i casi IPTD classificati come controlli sono stati considerati falsi negativi (FN); e i casi controlli classificati come IPTD sono stati considerati falsi positivi (FP). Tali distinzioni sono importanti poiché permettono di facilitare il confronto tra i vari modelli implementati per trovarne il migliore: il modello più robusto e accurato sarà infatti quello con il minore numero di FN e FP.

Oltre al calcolo dei FP e FN, per valutare le prestazioni fornite da ciascun modello di classificazione sono stati analizzati i seguenti parametri (forniti automaticamente da ORANGE alla fine della fase di testing di ciascun modello) [25]:

- l'area sotto la curva delle caratteristiche operative del ricevitore (AUC)
- l'accuratezza di classificazione (CA), che indica la proporzione di casi correttamente classificati
- la precisione (Prec), che indica la proporzione di TP tra tutti i casi classificati come positivi
- la sensibilità (Recall), che indica la proporzione di casi positivi che sono correttamente classificati
- la specificità (Spec), che indica la proporzione di casi negativi che sono correttamente classificati
- la media armonica ponderata di precisione e sensibilità (F1)

Inoltre, per una più efficace analisi dei risultati ottenuti, sono state calcolate le curve ROC relative ad ogni modello e sono poi state inserite in un unico grafico, così da poterne apprezzare meglio le differenze e similitudini.

La curva ROC è di fondamentale importanza statisticamente perché la sua area sottesa al grafico, chiamata AUC (Area Under the Curve) è una misura della accuratezza del modello. Infatti, se un ipotetico test discriminasse perfettamente i malati dai sani, cioè avesse il 100% di accuratezza, l'area della curva ROC avrebbe valore 1, se ne può quindi dedurre che tanto maggiore è l'area sotto la curva tanto maggiore è il potere discriminante del test [26].

L'interpretazione dei valori dell'AUC viene poi guidata prendendo come riferimento la classificazione proposta da Swets, secondo cui a vari intervalli di valori assunti da AUC possono essere associati diversi livelli di accuratezza, e quindi performance, del modello [27].

Nello specifico se $AUC=0.5$ il test non è informativo, se $0.5 < AUC \leq 0.7$ il test è poco accurato, se $0.7 < AUC \leq 0.9$ il test è moderatamente accurato e se $0.9 < AUC < 1.0$ il test è altamente accurato.

La figura sottostante (Fig. 8) mostra a confronto gli andamenti delle ROC ottenute adottando i vari modelli e la tabella invece elenca, per ogni algoritmo di machine learning, il preciso valore di AUC ottenuto.

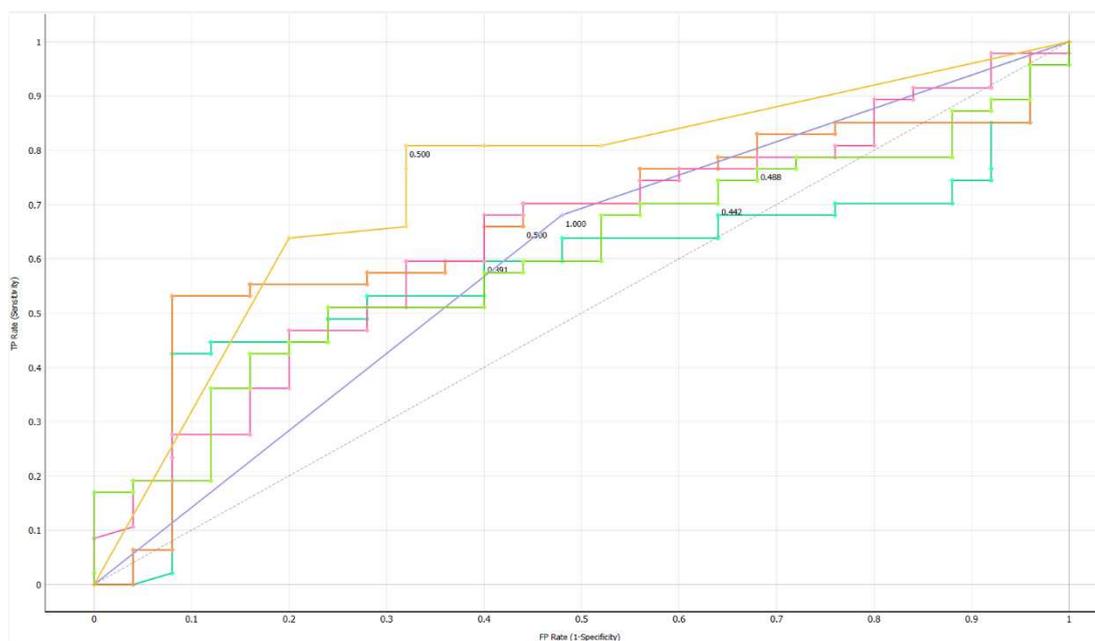


Fig.8: La figura rappresenta in un unico grafico le curve ROC ottenute dal testing dei sei modelli: in giallo è graficata la ROC del Decision Tree, in arancione quella della Logistic Regression, in azzurro Naive Bayes, in verde Gradient Boosting, in rosa Random Forest e in blu Adaboost.

Risulta evidente già dal grafico che nessuno dei modelli utilizzati presenti delle prestazioni particolarmente buone, il modello Naive Bayes, il Gradient Boosting e la Logistic Regression

addirittura presentano per alti valori di sensibilità un andamento che scende al di sotto della curva di casualità.

Sulla base della classificazione proposta da Swets il modello Naive Bayes è praticamente non informativo (AUC di poco maggiore di 0,5) mentre AdaBoost, Random Forest, Gradient Boosting e Logistic Regression sono tutti modelli poco accurati.

Ancora una volta il risultato ottenuto sembra confermare quanto già osservato nelle scorse analisi statistiche dei dati: le due distribuzioni sono molto simili e le features selezionate non hanno un forte potere di discernimento tra i due gruppi.

Osservando ancora il grafico sovrastante è possibile individuare chiaramente il modello più efficiente tra i sei studiati, questo è infatti l'Albero Decisionale. Se si studiano poi i valori dei parametri per la valutazione delle prestazioni dei modelli presenti nella tabella sottostante (Tab. 5), si può apprezzare come il Tree presenti per quasi tutti i valori più alti.

Tab. 5: La tabella mostra i valori dei parametri di accuratezza del modello ottenuti dal testing di tutte e sei i modelli realizzati

Model	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
Tree (1)	0.733	0.736	0.791	0.818	0.766	0.436	0.680	6.848
Logistic Regression	0.667	0.625	0.697	0.738	0.660	0.212	0.560	0.689
Gradient Boosting	0.633	0.597	0.695	0.688	0.702	0.103	0.400	1.495
Random Forest	0.600	0.625	0.727	0.692	0.766	0.134	0.360	0.656
AdaBoost	0.575	0.625	0.703	0.727	0.681	0.196	0.520	12.952
Naive Bayes	0.535	0.583	0.643	0.730	0.574	0.166	0.600	1.184

Questo modello presenta buoni valori di precisione ($Pr > 0,80$) e l'AUC più alto; secondo la classificazione di Swets il modello si troverebbe sulla soglia tra il poco accurato e il moderatamente accurato: non ha ancora una accuratezza soddisfacente ma rappresenta comunque un netto miglioramento rispetto alle prestazioni della maggior parte degli altri modelli.

Il modello ad albero presenta anche i valori migliori di accuratezza e di media armonica ponderata di precisione e sensibilità (F1). Entrando nello specifico, sebbene la media ponderata tra sensibilità e precisione sia complessivamente adeguata osservato il valore di ognuna rispetto al Tree si può

notare come la sensibilità (Recall) sia buona ma la specificità lasci a desiderare, raggiungendo un valore di solo 0,68.

Nonostante questo valore sia piuttosto basso anche in questo caso il Tree presenta la performance migliore, infatti, gli altri modelli presentano valori ancora più bassi di specificità: ricordando la definizione di quest'ultima si può quindi arrivare alla conclusione che questi modelli abbiano maggiore difficoltà nel riconoscere correttamente i soggetti di controllo piuttosto che quelli con ipotiroidismo.

Per approfondire meglio questo aspetto risultano utili le matrici di confusione, semplici tabelle che permettono di visualizzare chiaramente, tra le altre cose, anche i soggetti di controllo erroneamente classificati.

Una matrice di confusione è infatti una tabella utilizzata per valutare le prestazioni di un modello di classificazione su un set di dati di test in cui le etichette reali sono conosciute. Questa matrice mostra il numero di previsioni corrette e errate fatte dal modello, e può essere utilizzata per calcolare diverse metriche di valutazione, come precisione, recall, F1-score e l'accuratezza complessiva [26].

L'immagine (Fig. 9) che segue mostra le matrici di confusione ottenute per ognuno dei sei modelli testati.

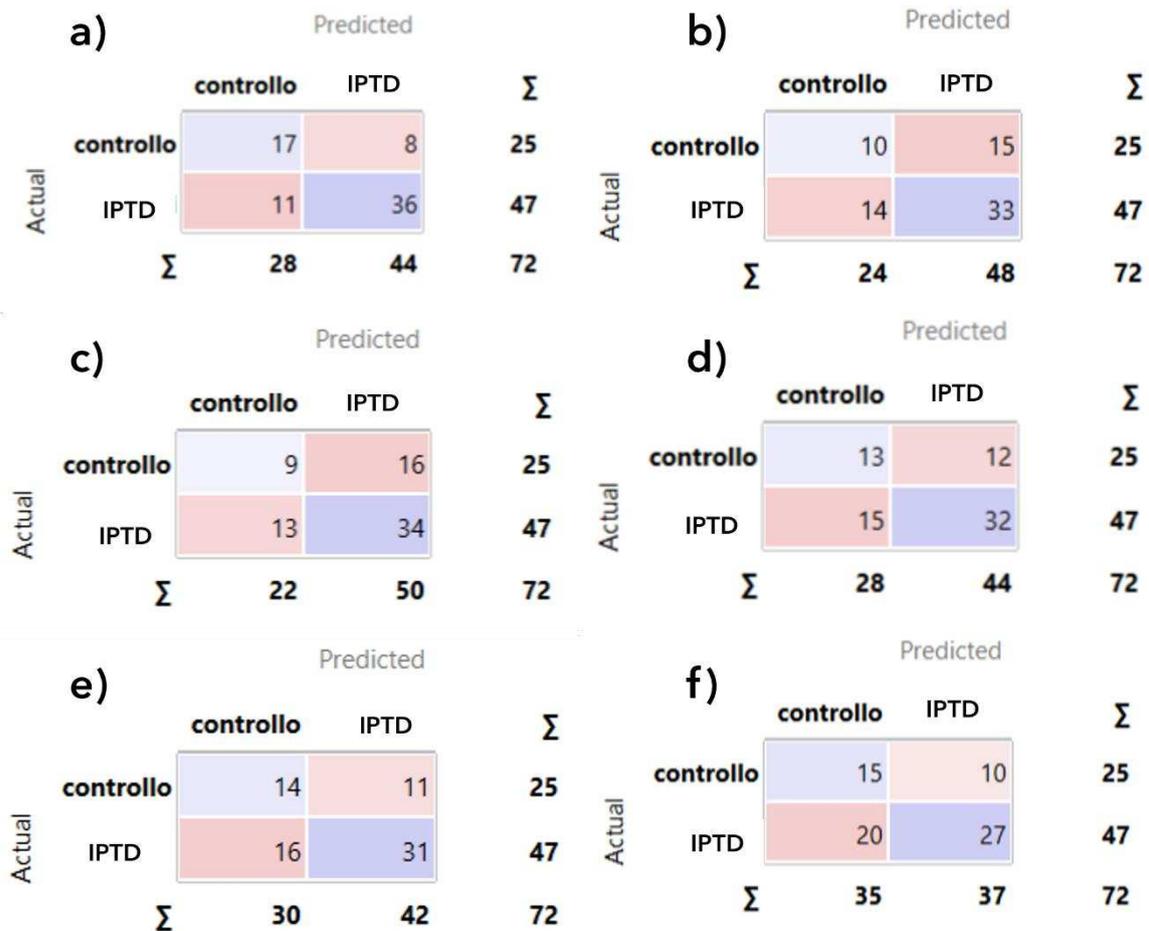


Fig. 9: l'immagine mostra le matrici di confusione ottenute testando i seguenti modelli: a) Albero Decisionale, b) Gradient Boosting, c) Random Forest, d) Ada Boost, e) Logistic Regression e f) Naive Bayes

Come già notato, tutte le matrici di confusione presentano numerosi soggetti di controllo erroneamente classificati come malati, l'utilizzo di tali modelli, quindi, comporta un importante rischio di incorrere in falsi positivi. I falsi negativi, soggetti con ipotiroidismo non riconosciuti, sono invece presenti in minore numero nei modelli, ciò è un bene perché in ambito medico si predilige favorire la sensibilità a discapito della specificità [28].

Se anche infatti un soggetto in realtà sano viene determinato a rischio di ipotiroidismo, tale soggetto dovrà al più sottoporsi a ulteriori test e venire monitorato con più attenzione, niente in ogni caso che possa realmente nuocere alla sua salute; al contrario invece un soggetto realmente malato non riconosciuto rischia di trascurare la propria condizione, poiché ve ne è all'oscuro, con un conseguente possibile peggioramento della sua condizione.

Ancora una volta, l'osservazione delle matrici di confusione conferma quanto già scoperto: l'albero decisionale è il modello che presenta la minor quantità di falsi negativi e falsi positivi, risultando quindi il più affidabile e performante tra tutti.

Un aspetto negativo legato all'utilizzo di questo modello è però legato al valore del parametro di LogLoss, esso infatti tiene conto dell'incertezza della previsione fatta dal modello in base a quanto varia rispetto all'etichetta effettiva: il Tree presenta infatti il secondo valore in assoluto più alto di tale parametro. E' quindi un modello molto soggetto a variabilità dei valori [29].

Dall'analisi condotta in ambiente Orange e dall'applicazione di tecniche di machine learning, si è arrivati a concludere che le features selezionate e calcolate dal database non siano da sole sufficienti per implementare un modello di classificazione abbastanza accurato da essere immediatamente sfruttabile, solitamente infatti nelle pubblicazioni scientifiche un modello viene considerato valido e utilizzabile quando presenta un valore di AUC maggiore di 0,8.

Nessuno dei modelli trovati è riuscito a raggiungere questo risultato ma quello che più ci si è avvicinato è il modello dell'Albero Decisionale; quest'ultimo presenta buoni valori di precisione e coefficiente di accuratezza, con un adeguato valore di sensibilità ma minore specificità.

Conclusione

Lo studio svolto aveva l'obiettivo di realizzare, attraverso l'utilizzo delle potenti tecniche di machine learning, un modello che potesse permettere di riconoscere e individuare correttamente i soggetti con ipotiroidismo in concomitanza al DMT1.

Le analisi effettuate hanno messo in evidenza la difficoltà del problema in esame: i soggetti con ipotiroidismo non presentano infatti numerose e sostanziali differenze rispetto a quelli con solo diabete mellito. Ciò può essere osservato in un primo momento dal grafico dell'andamento medio giornaliero della glicemia, il tracciato inerente ai soggetti unicamente diabetici e quello relativo agli ipotiroidici seguono quasi identicamente lo stesso andamento senza discostarsi affatto l'uno dall'altro.

I due gruppi però non sono del tutto identici, sarebbe infatti fisiologicamente improbabile che una patologia importante come l'ipotiroidismo non comporti alcuna alterazione glicemica; infatti, ben 17 metriche risultano significativamente differenti nei due gruppi.

L'utilizzo di tecniche di machine learning, in particolar modo l'applicazione dell'Albero Decisionale, ha permesso di restringere ulteriormente il campo delle features necessarie per la creazione del modello a 8 caratteristiche fondamentali: `sdwIndex`, `aucGlucose`, `jIndex`, `mageIndex`, `igc`, `cogi`, `conga` e `stdGlucoseROC`. Ancora una volta si può osservare la complessità del modello, analizzando infatti il grafico ad albero appare evidente il grande numero di foglie e rami necessari per collocare correttamente tutti i soggetti; infatti, maggiore è la dimensione del diagramma ad albero minore sarà l'efficacia del modello.

Sono stati valutati i modelli ottenuti tramite gli algoritmi di classificazione Random Forest, Adaboost e Gradient Boosting, Naive Bayes e Logistic Regression, oltretutto testando il già precedentemente creato modello ad Albero Decisionale, ed i risultati ottenuti mostrano un complessivo basso livello di accuratezza dei modelli.

Tra tutti i modelli, però, quello ad Albero Decisionale presenta prestazioni nettamente migliori: un coefficiente di precisione buono, AUC e sensibilità adeguate con però un minore valore di specificità. Tutti i modelli implementati, infatti, presentano particolari difficoltà nel riconoscere accuratamente i soggetti di controllo, ovvero coloro che non presentano ipotiroidismo. Ciò è reso evidente dall'alto numero di falsi positivi nelle varie matrici di confusione.

Il modello che si è quindi ottenuto alla fine della ricerca è un buon punto di partenza per ulteriori studi che vadano a migliorare delle prestazioni del modello ad albero. Sebbene, infatti, l'elaborato presentato non sia riuscito a progettare un modello ad alta accuratezza, è stato individuato distintamente l'algoritmo di ML più adatto per il problema, l'Albero Decisionale, e sono state selezionate 8 features significative per il riconoscimento dei soggetti con ipotiroidismo, riducendo molto il numero di metriche su cui concentrarsi negli studi futuri.

Inoltre, sulla base del diagramma ad albero ottenuto da questo elaborato, il quale presenta come caratteristiche decisionali quasi esclusivamente le metriche relative alla variabilità della glicemia, potrebbe essere interessante indagare il possibile legame tra la variabilità del tracciato glicemico e lo sviluppo di ipotiroidismo. Potrebbe anche risultare utile allargare le feature prese in esame, aggiungendo allo studio anche le caratteristiche antropometriche, l'ipotiroidismo infatti colpisce molto più frequentemente soggetti donne e anziani e di gruppo etnico bianco.

Infine, poiché le features analizzate in questo studio da sole non avevano abbastanza contenuto informativo da permettere la costruzione di un modello più semplice e accurato, potrebbe essere utile approcciarsi alla risoluzione del problema di classificazione attraverso tecniche di machine learning non supervisionato ed eventualmente effettuare un confronto con le performance ottenute da questo modello per determinare quale tra i due metodi sia il più indicato per la classificazione in esame.

In conclusione, questo studio può essere considerato come un utile primo approccio al problema, il quale per complessità e rilevanza si presta ad essere l'oggetto di future indagini.

Bibliografia e Sitografia

- [1] Widmaier E, Raff H, Strang K. *Vander Fisiologia*. Casa editrice Ambrosiana. Aprile 2018. Seconda edizione.
- [2] *Definition, diagnosis and classification of diabetes mellitus and its complications*. Consultation, W. H. O. :25. 1999.
- [3] Ilari L., Piersanti A., Gobl C., Burattini L., Kautzky-Willer A., Tura A., Morettini M. *Unraveling the Factors Determining Development of Type 2 Diabetes in Women with History of Gestational Diabetes Mellitus through Machine-Learning Techniques*. *Frontiers in Physiology*. 13:789212. 2022.
- [4] *Informazioni generali sul diabete*. Istituto superiore di sanità. 15 Gennaio 2021. <https://www.epicentro.iss.it/diabete/>
- [5] Yeram PB., Kulkarni YA. *Glycosides and Vascular Complications of Diabetes*. *Chemistry & Biodiversity*. 19.10: e202200067. 2022 October.
- [6] Guastamacchia E., Triggiani V., Aglialoro A., Aiello A., Ianni L., Maccario M., Zini M., Giorda C., Guglielmi R., Betterle C., Attanasio R., Borretta G., Papini E., Castello R., Ceriello A. *TIREOPATIE e DIABETE: Raccomandazioni per la pratica clinica*. Associazione Medici Endocrinologi & Associazione Medici Diabetologi. 2014.
- [7] Yingqi Z., Shucheng S., Kai Z., Jie Y., Jiqing L., Fuzhong X. *Causal relationship between type 1 diabetes and hypothyroidism: A Mendelian randomization study*. *Clinical Endocrinology*. Volume 97, Issue 6, Pages 740-746. August 2022.
- [8] *4 ways to take insulin*. Centers for disease control and prevention. 18 Aprile 2023. <https://www.cdc.gov/diabetes/basics/type-1-4-ways-to-take-insulin.html>
- [9] *All about your A1C*. Centers for disease control and prevention. 30 Settembre 2022. <https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html>
- [10] Perin P, Lostia S, Maldonato A, Musacchio N, Purrello F e Vespasiani G. *Raccomandazioni sull'uso dell'autocontrollo domiciliare della glicemia*. Associazione Medici Diabetologi, Società Italiana di Diabetologia. 2003.

- [11] Piersanti A., Giurato F., Gobl C., Burattini L., Tura A., Morettini M. *Software Packages and Tools for the Analysis of Continuous Glucose Monitoring Data*. DIABETES TECHNOLOGY & THERAPEUTICS. Volume 25, Number 1, 2023.
- [12] Didyuk O, Econom N, Guardia A, Livingston K, Klueh U. *Continuous Glucose Monitoring Devices: Past, Present, and Future Focus on the History and Evolution of Technological Innovation*. Journal of diabetes science and technology. 2021 May;15(3):676-683.
- [13] Aleppo G, Ruedy K, Riddlesworth D, Kruger D, Peters A, Hirsch I, Bergenstal R, Toschi E, Ahmann A, Shah V, Rickels M, Bode B, Philis-Tsimikas A, Pop-Busui R, Rodriguez H, Eyth E, Bhargava A, Kollman C, Beck R. *REPLACE-BG: A Randomized Trial Comparing Continuous Glucose Monitoring With and Without Routine Blood Glucose Monitoring in Adults With Well-Controlled Type 1 Diabetes*. Diabetes Care Volume 40. April 2017.
- [14] Berner, Eta S., ed. *Clinical Decision Support Systems*. New York, NY: Springer. 2007.
- [15] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. *Introduction to Machine Learning, Neural Networks, and Deep Learning*. Translational vision science & technology. 9(2):14. 2020 Feb 27.
- [16] *Machine Learning in MATLAB*. MathWorks. <https://it.mathworks.com/help/stats/machine-learning-in-matlab.html>
- [17] Piersanti A., Burattini L., Salvatori B., Tura A., Morettini M. *A Machine Learning Framework based on Continuous Glucose Monitoring to Prevent the Occurrence of Exercise-Induced Hypoglycemia in Children with Type 1 Diabetes*. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). 2023.
- [18] Petrie JR, Guzik TJ, Touyz RM. *Diabetes, Hypertension and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms*. Can J Cardiol. 34(5):575-584. 2018 May.
- [19] Fico G. *Exploring the frequency domain of continuous glucose monitoring signals to improve characterization of glucose variability and of diabetic profiles*. Diabetes Science and Technology. 11(4), 773-779. 2017.
- [20] Cappon G., Sparacino G., Facchinetti A. *AGATA: A Toolbox for Automated Glucose Data Analysis*. Journal of Diabetes Science and Technology, 1–13. 2023.

- [21] Maimon, Oded Z., and Lior Rokach. *Data mining with decision trees: theory and applications*. Vol. 81. World scientific, 2014.
- [22] Shouman, Mai. *Using decision tree for diagnosing heart disease patients*. Proceedings of the Ninth Australasian Data Mining Conference. 121: 23–30. 2011.
- [23] Caruana R., Niculescu-Mizil A. *An empirical comparison of supervised learning algorithms*. Proceedings of the 23rd International Conference on Machine Learning. 2006.
- [24] *Cross-Validation in MATLAB*. MathWorks. <https://it.mathworks.com/discovery/cross-validation.html>
- [25] Powers, David M. *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies. 2 (1): 37–63. 2011.
- [26] D'Arrigo G., Provenzano F., Torino C., Zoccali C., Tripepi G. *TEST DIAGNOSTICI E L'ANALISI DELLA CURVA ROC*. G Ital Nefrol. 28 (6): 642-647. 2011.
- [27] Swets J A., Dawes R M., Monahan J. *Psychological science can improve diagnostic decisions*. Psychological Science in the Public Interest. 1:1–26. 2000.
- [28] Kraemer H C. *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage. 1992.
- [29] *Test and Score*. Orange Data Mining. <https://orangedatamining.com/widget-catalog/evaluate/testandscore/>