UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN BIOINGEGNERIA

"METODI DI MACHINE LEARNING APPLICATI A POPOLAZIONI VIRTUALI DI SOGGETTI AFFETTI DA DIABETE DI TIPO 2"

Relatore: Prof. / Dott. Pedersen Morten Gram

Laureando/a: Califano Benito

Correlatore: Prof. /Dott. Visentin Roberto

ANNO ACCADEMICO 2022 – 2023

Data di laurea: 17/04/2023

Chi non ha niente in fin dei conti non perde nulla
Tanto vale provare a vivere, perché no?

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Il diabete mellito di tipo 2 è una malattia metabolica caratterizzata da glicemia alta in un contesto di insulino-resistenza ed insulino-deficienza relativa. Rappresenta circa il 90% dei casi di diabete nel mondo e lo sviluppo della patologia è causato da una combinazione tra lo stile di vita, endocrinopatie, e fattori genetici. L'Organizzazione Mondiale della Sanità riconosce la condizione di diabete dopo una rilevazione di elevati valori di glucosio nel sangue con la presenza di sintomi tipici, ed in questo contesto risulta di fondamentale importanza la somministrazione di un test chiamato "Oral Glucose Tolerance Test", comunemente OGTT, tramite il quale è possibile avere una diagnosi precisa della patologia. Grazie allo sviluppo e all'utilizzo di un software, il "T2D Simulator", di un gruppo di ricerca dell'Università di Padova, che fra le tante cose ci ha permesso di simulare il suddetto test su soggetti virtuali, è stato possibile ricostruire delle vere e proprie condizioni sperimentali, che hanno consentito l'ottenimento di un certo numero di datasets successivamente analizzati ed utilizzati per l'addestramento di specifici algoritmi di Machine Learning. L'applicazione degli algoritmi di apprendimento automatico risulta motivata in primis dalla costante esigenza di approfondire le nostre conoscenze riguardo le patologie in generale, fattibile grazie alla possibilità di esplorare grandi quantità dati e di informazioni tramite l'utilizzo di strumenti ad hoc che semplificano enormemente questo processo, ed in secondo luogo per lo sviluppo di strumenti informatici potenzialmente utilizzabili a supporto di attività cliniche, che come nel caso del test OGTT, risultano essere leggermente invasive per i pazienti a causa dell'elevato numero di prelievi ematici, ed economicamente dispendiose per le aziende ospedaliere che le conducono. Sono dunque state condotte preliminarmente delle analisi di regressione sui dataset accuratamente organizzati, con l'obiettivo di sviluppare dei modelli capaci di poter stimare in maniera accurata specifici parametri funzionali per le cinetiche di secrezione ed assorbimento di glucosio ed insulina. I modelli testati sono di comune utilizzo nell'ambito delle analisi regressive se si parla di supervised learning, ed in particolare sono: Linear Regression, Regression Tree, Random Forest e Boosting. Uno step successivo ha riguardato lo sviluppo di modelli per la classificazione di soggetti diabetici, preliminarmente in relazione all'assunzione o meno di Metformina, e solo successivamente in base allo stadio di avanzamento della patologia. Si menzionano a tal proposito la Logistic Regression, il k-NN Classifier, il Naive-Bayes Classifier, i Classification Tree, le Random Forest e le Support Vector Machine. Da precisare che per tutti gli algoritmi elencati si è seguita la tipica pipeline di operazioni di preprocessing dei dataset e di taratura degli iperparametri tramite k-Fold Cross Validation, caratteristica degli studi nell'ambito del Machine Learning.

Capitolo 1

Diabete Mellito e T2D Simulator

1.1 Premessa

Per diabete mellito si intende un gruppo eterogeneo di endocrinopatie caratterizzate da una alterata tolleranza glucidica cronica, conseguente a un difetto assoluto o relativo di insulina [10]. La percentuale di popolazione mondiale affetta viene stimata intorno al 5%, con una maggiore prevalenza nel sesso femminile [1]. L'OMS stima che ci sarà un fortissimo incremento di prevalenza di diabete mellito negli USA, in Medio Oriente e nel Sud-Est asiatico, con una previsione di più di 360 milioni di persone malate entro il 2030 [12]. L'applicazione di strumenti informatici quali simulatori e modelli di Machine Learning, possono risultare di fondamentale importanza per una maggiore comprensione della patologia sotto vari aspetti, ed essere di ausilio nelle fasi di ricerca e di sperimentazione di nuove tipologie di trattamento.

1.2 Diabete Mellito

1.2.1 Eziologia

Il diabete mellito può essere causato da una serie di fattori. Alcuni possono essere derivati da difetti dell'azione insulinica, come l'insulinoresistenza di tipo A, il leprecaunismo, la sindrome di Rabson-Mendenhall e le sindromi lipodistrofiche, inoltre, anche alcune malattie del pancreas possono essere cause del diabete, come nel caso della pancreatite, della fibrosi cistica, nell'emocromatosi e nel tumore del pancreas [27]. Anche patologie genetiche come sindrome di Down, sindrome di Turner e sindrome di Klinefelter sono responsabili dello sviluppo del diabete. Tra i principali fattori di rischio si riscontrano [14]:

- Obesità: BMI $\geq 25\ kg/m^2$ per il T2D
- Inattività fisica
- Ipertensione: pressione arteriosa sistolica $\geq 140\ mmHg$ e/o pressione arteriosa diastolica $\geq 90\ mmHg$
- Colesterolo HDL $\leq 35\ mg/dL$
- Trigliceridi $\geq 250\ mg/dL$
- Ipogonadismo

- Disturbi del sonno

Anche l'età favorisce la comparsa del diabete, poiché essa si accompagna ad una diminuita sensibilità dei tessuti periferici all'insulina. Di seguito si riporta una distinzione fra le due principali tipologie di diabete mellito.

Diabete Mellito di Tipo 1

Il diabete mellito di tipo 1 è una forma di diabete che si configura come malattia autoimmune [15], caratterizzata dalla distruzione delle cellule β pancreatiche, e che comporta solitamente associazione all'insulino-deficienza. Le cause sono un insieme di fattori che riguardano la genetica, l'ambiente e l'immunologia: ad una predisposizione genetica di base si unisce uno stimolo immunologico che, con il passare del tempo, porta alla distruzione delle cellule β. Quando la percentuale di cellule β perse arriva all'80%, ci si ritrova di fronte al diabete mellito di forma 1.

Diabete Mellito di Tipo 2

Il diabete mellito di tipo 2, è una malattia metabolica, caratterizzata da glicemia alta in un contesto di inusulino-resistenza e insulino-deficienza relativa [28]. Rappresenta circa il 90% dei casi di diabete. Lo sviluppo della patologia è causato da una combinazione tra lo stile di vita, endocrinopatie e fattori genetici. Anche la mancanza di sonno è stata associata allo sviluppo di diabete di tipo 2. Si ritiene che ciò agisca attraverso il suo effetto sul metabolismo.

L'oggetto di studio del presente elaborato di tesi è il diabete di tipo 2. In seguito si riportano, di conseguenza, informazioni riguardanti esclusivamente questa specifica tipologia di diabete.

1.2.2 Patogenesi del Diabete di tipo 2

Il diabete di tipo 2 è dovuto sia all'insufficiente produzione d'insulina dalle cellule β del pancreas che all'insulino-resistenza, cioè alla scarsa sensibilità delle cellule all'azione dell'insulina [5]. Nel fegato, l'insulina sopprime normalmente il rilascio di glucosio, tuttavia, nella condizione di insulino-resistenza, il fegato rilascia impropriamente glucosio nel sangue. Inizialmente il fisico reagisce all'insulino-resistenza e tiene sotto controllo la glicemia aumentando la sintesi di insulina, dopo un certo tempo però questo meccanismo cede e anche la sintesi insulinica diminuisce, ponendo le basi all'insorgenza del diabete mellito. Altri meccanismi, potenzialmente importanti, associati al diabete di tipo 2 e all'insulino-resistenza, sono una maggiore ripartizione dei lipidi nelle cellule adipose, la mancanza di incretine, bassi livelli di ormoni che aumentano la sensibilità all'insulina (es. testosterone, estrogeni, fattori di crescita insulino-simili etc.), elevati livelli di ormoni che inibiscono l'azione insulinica (es. glucocorticoidi, glucagone, mineralcorticoidi, adrenalina etc.), aumento della ritenzione di acqua e sale dai reni e regolamentazione inadeguata di metabolismo da parte del sistema nervoso centrale.

1.2.3 Sintomi del Diabete di tipo 2

I sintomi classici del diabete sono poliuria (minzione frequente), polidipsia (aumento della sete), iperfagia e perdita di peso [24]. Altri disturbi comunemente associati a questa malattia sono astenia cronica, disfunzione erettile, ipogonadismo, infezioni alle vie urinarie, prurito, vista offuscata, neuropatia periferica, ricorrenti infezioni vaginali. Le

persone con diabete mellito tipo 2 possono raramente presentarsi con coma iperosmolare-iperglicemico non chetosico (una presenza di glucosio nel sangue molto elevata, associata ad una diminuzione del livello di coscienza e ipotensione) [5].

1.2.4 Diagnosi del Diabete di tipo 2

L'Organizzazione Mondiale della Sanità riconosce la condizione di diabete (tipo 1 e tipo 2) dopo una rilevazione di elevati valori di glucosio nel sangue con la presenza di sintomi tipici. I valori elevati di glicemia possono essere così rilevati [17]

- Glicemia plasmatica a digiuno $\geq 7\ mmol/l$ (126 mg/dl)
- Test orale di tolleranza al glucosio, o OGTT (Oral Glucose Tolerance Test).

OGTT - Oral Glucose Tolerance Test

Il test orale di tolleranza al glucosio [29], è un test clinico che permette di valutare come la concentrazione di glucosio (e di insulina nel caso in cui venga associato a dosaggio dell'insulina) cambia nel sangue dopo l'assunzione di una dose nota, quindi se il corpo ha un metabolismo glucidico normale o alterato. In condizioni normali, dopo un carico orale di glucosio, nel sangue aumenta la glicemia dopo qualche minuto. Le cellule β del pancreas vengono stimolate dall'alta concentrazione di glucosio a secernere insulina, riversando quest'ultima nel sangue attraverso il processo dell'esocitosi. In definitiva l'insulina abbassa la concentrazione di glucosio nel sangue, quindi in condizioni normali, nel giro di qualche ora, la glicemia scende a livelli simili a quelli basali. Se i valori di glicemia sono alterati significa che il metabolismo del glucosio non è normale, e questo porta alla diagnosi di diabete. Valori normali di glicemia sono:

- A digiuno: inferiori a 100 mg/dl
- Dopo due ore dall'ingestione di 75 g di glucosio: minori di 140 mg/dl

Per la diagnosi di diabete, in generale:

- La glicemia a digiuno è uguale o superiore a 126 mg/dl
- La glicemia dopo 2 ore è uguale o superiore a 200 mg/dl

1.2.5 Trattamento del Diabete di tipo 2

La gestione del diabete di tipo 2 si concentra su interventi sullo stile di vita [13], sulla riduzione degli altri fattori di rischio cardiovascolare e sul mantenimento di livelli di glicemia nell'intervallo di normalità. La gestione di altri fattori di rischio cardiovascolare, tra cui ipertensione, colesterolo alto e microalbuminuria, migliora l'aspettativa di vita di una persona. Una dieta corretta e l'esercizio fisico sono fondamentali per la cura del diabete. Se, nei pazienti con diabete lieve, i cambiamenti nello stile di vita non hanno portato ad un controllo migliore del livello di zuccheri nel sangue, il trattamento farmacologico deve essere preso in considerazione. Vi sono diverse classi di farmaci disponibili come anti-diabetici. La metformina è generalmente raccomandata come trattamento di prima linea, in quanto vi sono alcune prove che sia in grado di diminuire la mortalità cardiovascolare. La maggior parte dei diabetici, inoltre, necessita di apporti di insulina esterni.

1.3 T2D Simulator

Prove in silico su pazienti affetti da diabete di tipo 2, sono molto utili per il test di trattamenti e lo sviluppo di nuovi farmaci. A tal proposito è stato sviluppato in linguaggio MATLAB, da un gruppo di ricerca dell'Università di Padova, il T2D Simulator [25], un simulatore che permette di riprodurre la variabilità osservata in una popolazione di pazienti affetti da diabete di tipo 2, e consente inoltre di condurre esperimenti su soggetti virtuali, con varie condizioni patologiche. Il simulatore si basa su un modello composto da 15 equazioni differenziali contenenti 39 parametri accuratamente stimati. Esso descrive il transito di glucosio attraverso il tratto gastrointestinale, l'azione dell'insulina sull'utilizzo e sulla produzione endogena di glucosio, e il controllo di quest'ultimo tramite la secrezione di insulina. I flussi metabolici tenuti in considerazione per lo sviluppo del simulatore sono l'EGP (Endogenous Glucose Production), il Ra_{meal} (Glucose Rate of Appearance), U (Glucose Utilization), ISR (β-cell Insulin Secretion Rate), ed HE (Hepatic Insulin Extraction).

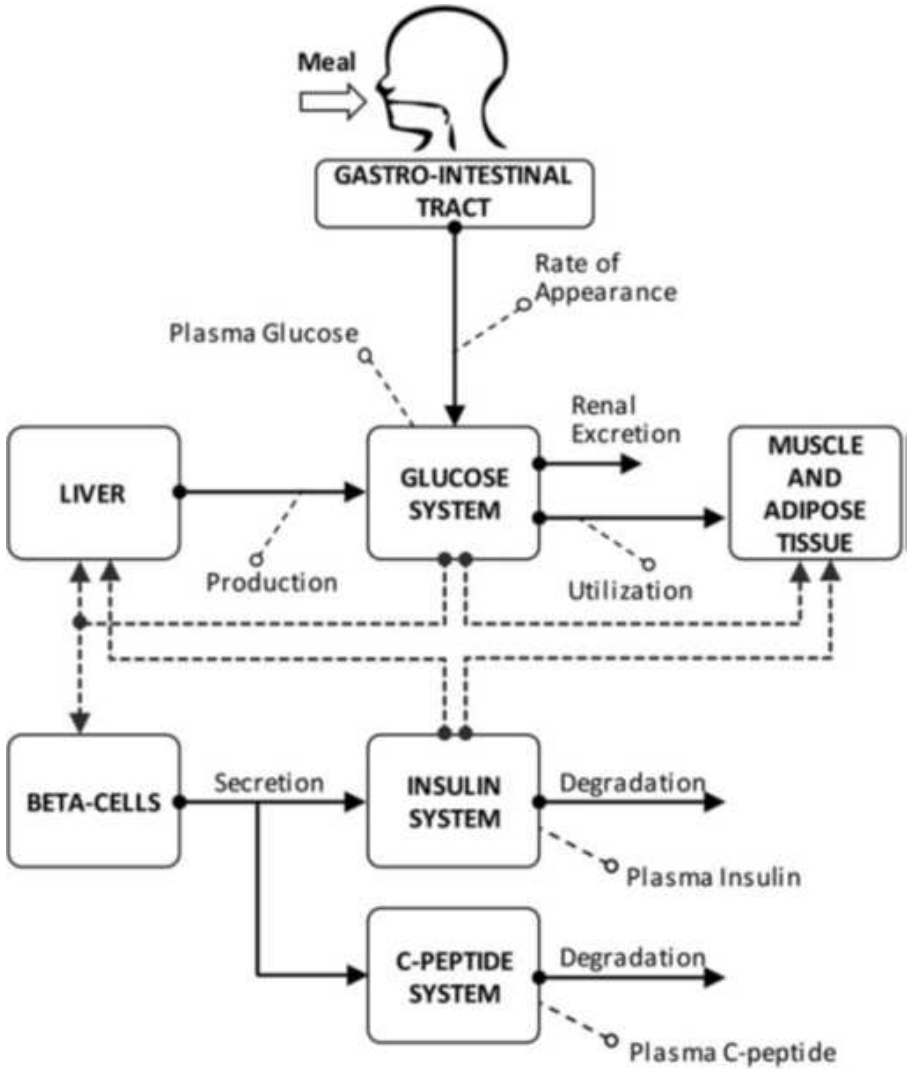

Figura 1.1: Rappresentazione schematizzata della struttura del T2D Simulator. I flussi metabolici sono indicati dalle linee continue, mentre le azioni di controllo sono rappresentate da linee tratteggiate

Il dominio di validità del simulatore è attualmente limitato ad uno scenario comprendente un singolo pasto, ed è programmato per lavorare su una popolazione di 100 soggetti

in silico, che riflette le principali caratteristiche di una popolazione di pazienti affetti da diabete di tipo 2. Esso prevede, inoltre, un comparto utilizzato per la descrizione di farmacocinetica e farmacodinamica per il test dei medicinali di interesse. Fondamentale è l'interfaccia grafica [26], appositamente progettata per facilitare il design degli esperimenti in silico da condurre, in modo da estendere l'usabilità del simulatore anche ad utenti che non hanno conoscenze in ambito di programmazione informatica. Essa permette di selezionare:

- Le dimensioni della popolazione da testare
- Tipologia di esperimento da condurre
- Durata e sampling grid
- Dosi di glucosio ed insulina somministrate
- Dose di farmaco da testare

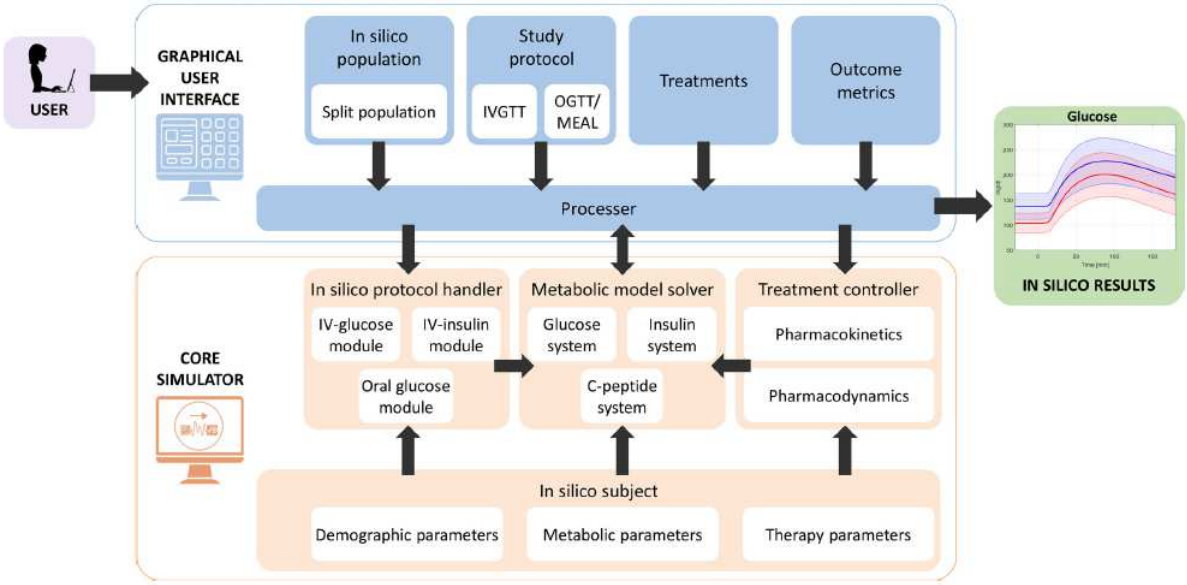

Figura 1.2: Schema completo dell'architettura del T2D Simulator, comprendente: l'interfaccia grafica utilizzata dall'utente per selezionare i parametri dell'esperimento e successivamente leggere i risultati in output; il core del simulatore che si occupa dell'elaborazione delle informazioni e la conduzione dell'esperimento sui soggetti in silico

1.3.1 T2D Simulator - Output

L'interfaccia grafica del simulatore permette, oltre la scelta dei parametri in input utilizzati per la conduzione di un esperimento, anche la visualizzazione in maniera semplificata dei risultati ottenuti dalla simulazione. Di seguito è riportato un esempio della rappresentazione degli output di una simulazione con i seguenti parametri in input:

- 100 soggetti affetti da diabete di tipo 2
- Test OGTT utilizzando 75 g di glucosio, ed un sampling time di 300 min
- Trattamento con 500 mg di Metformina e controllo in placebo

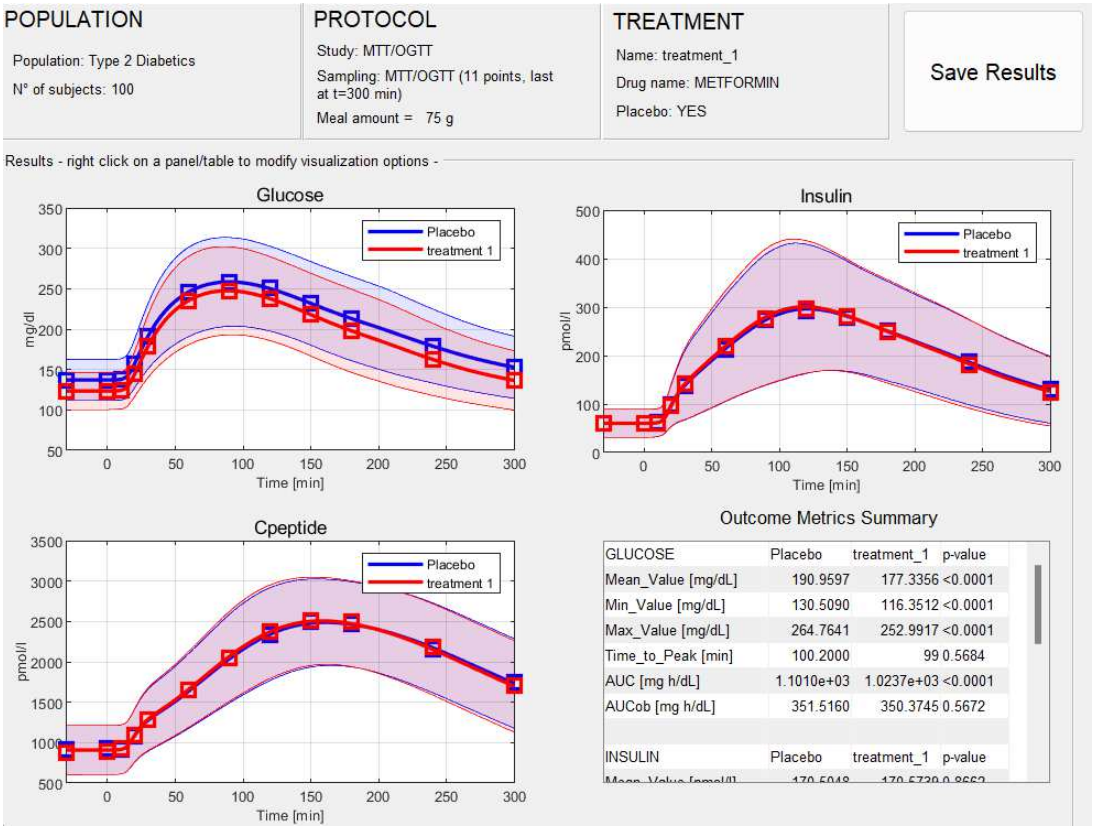

Figura 1.3: Risultati di una simulazione condotta su pazienti affetti da diabete di tipo 2, sottoposti a test OGTT, e trattati con Metformina

Le curve in rosso rappresentano le evoluzioni di Glucosio, Insulina e C-Peptide nel tempo, per quanto riguarda i soggetti trattati con Metformina. Le curve in blu rappresentano il controllo in placebo. I grafici visualizzano i valori medi e le deviazioni standard delle metriche per tutti e 100 i pazienti considerati. Il plot dei grafici è accompagnato dalla creazione di un file denominato "SimResults.mat" che contiene le informazioni specifiche per ogni paziente. Quest'ultimo è stato utilizzato per la creazione di dataset utilizzati nell'applicazione di modelli di Machine Learning.

1.4 Scopi ed obiettivi di tesi

Il lavoro di tesi si compone sostanzialmente di due step: uno preliminare che vede impiegato il T2D Simulator per la generazione di curve da test OGTT (Glucosio, C-Peptide, Insulina) per coorti di pazienti virtuali con diagnosi di diabete di tipo 2 in fase iniziale ed avanzata, trattati con metformina ed utilizzando un controllo in placebo; un secondo step, obiettivo finale dell'elaborato di tesi, che consiste nell'applicazione di diversi algoritmi di Machine Learning (in particolare della branca del Supervised Learning) per lo sviluppo di modelli di classificazione (capaci di classificare pazienti in base allo stato di avanzamento della patologia e/o all'assunzione o meno di Metformina) e di regressione (principalmente per l'estrapolazione di informazioni riguardo le caratteristiche della patologia partendo dalle features collegate ad ogni paziente considerato nello studio).

Capitolo 2

Dati e preprocessing

2.1 Dataset

L'utilizzo del T2D Simulator è stato di fondamentale importanza per lo sviluppo di questo elaborato di tesi. Grazie ad esso è stato possibile condurre 4 simulazioni:

- Una prima simulazione utilizzando 1 coorte di soggetti affetti da diabete di tipo 2 in fase iniziale
- Successive tre simulazioni utilizzando 3 diverse coorti di soggetti affetti da diabete di tipo 2 in fase avanzata

In entrambi i casi, per ognuna delle simulazioni svolte, i parametri scelti per la conduzione degli esperimenti sono stati:

- 100 soggetti affetti da diabete di tipo 2
- Test OGTT utilizzando 75 g di glucosio, ed un sampling time di 300 min
- Trattamento con 500mg di Metformina
- Controllo in placebo

Grazie alle simulazioni svolte è stato possibile ottenere 4 diversi file di tipo "*struct*", ognuno contenente 9 campi differenti, e riferiti sia ai pazienti trattati con metformina, che ai pazienti in placebo. Tramite lo sviluppo di un algoritmo ad hoc in linguaggio MATLAB, è stato possibile estrarre dai file in output dal simulatore le informazioni ritenute necessarie per la costruzione dei dataset successivamente utilizzati nell'applicazione dei vari modelli di Machine Learning presi in considerazione.

Si sono così ottenuti **8 dataset completi** (2 relativi a soggetti con T2D in fase iniziale e 6 relativi a soggetti con T2D in fase avanzata) contenenti ognuno 100 pazienti e 33 features, ed **8 dataset ridotti** (stesso numero di pazienti e stesso rapporto fra fase iniziale e fase avanzata, ma con features in parte differenti e di inferiore numero, nello specifico 17). Per quanto riguarda i datasets completi, le features tenute in considerazione sono così riassunte:

- **AUC_G**: Area sotto la curva dell'evoluzione del glucosio nel tempo $[mg * h/dL]$
- **AUCob_G**: Area sotto la curva soprabasale dell'evoluzione del glucosio nel tempo $[mg * h/dL]$
- **Max_Value_G**: Massimo valore di glucosio $[mg/dL]$

- **Mean_Value_G**: Valore medio di glucosio $[mg/dL]$
- **Min_Value_G**: Minimo valore di glucosio $[mg/dL]$
- **Time_to_Peak_G**: Tempo impiegato per raggiungere il picco di glucosio $[min]$
- **AUC_I**: Area sotto la curva dell'evoluzione dell'insulina nel tempo $[pmol * h/L]$
- **AUCob_I**: Area sotto la curva soprabasale dell'evoluzione dell'insulina nel tempo $[pmol * h/L]$
- **Max_Value_I**: Massimo valore di insulina $[pmol/L]$
- **Mean_Value_I**: Valore medio di insulina $[pmol/L]$
- **Min_Value_I**: Minimo valore di insulina $[pmol/L]$
- **Time_to_Peak_I**: Tempo impiegato per raggiungere il picco di insulina $[min]$
- **AUC_CP**: Area sotto la curva dell'evoluzione del C-Peptide nel tempo $[pmol * h/L]$
- **AUCob_CP**: Area sotto la curva soprabasale dell'evoluzione del C-Peptide nel tempo $[pmol * h/L]$
- **Max_Value_CP**: Massimo valore di C-Peptide $[pmol/L]$
- **Mean_Value_CP**: Valore medio di C-Peptide $[pmol/L]$
- **Min_Value_CP**: Minimo valore di C-Peptide $[pmol/L]$
- **Time_to_Peak_CP**: Tempo impiegato per raggiungere il picco di C-Peptide $[min]$
- **AGE**: Età del paziente $[anni]$
- **BMI**: Indice di massa corporea del paziente $[Kg/m^2]$
- **BW**: Peso del paziente [Kg]
- **BSA**: Area di superificie corporea $[m^2]$
- **EGPb**: Produzione Endogena di glucosio basale $[mg/(Kg * min)]$
- **Gb**: Glucosio basale $[mg/dL]$
- **Gender**: Sesso $[M/F]$
- **k1**: Parametro utilizzato nelle equazioni per la descrizione della cinetica del glucosio $[min^{-1}]$
- **k2**: Parametro utilizzato nelle equazioni per la descrizione della cinetica del glucosio $[min^{-1}]$
- **Height**: Altezza $[cm]$
- **Ib**: Insulina basale $[pmol/L]$
- **PHId (ϕ_d)**: Reattività delle cellule β al variare della velocità di variazione del glucosio $[10^{-9}]$

- **PHIs (ϕ_s)**: Reattività delle cellule β al glucosio $[10^{-9}\ min^{-1}]$
- **Vmx**: Sensibilità dell'insulina all'utilizzo del glucosio $[mg/kg/min\ per\ pmol/L]$
- **Treat.Plac**: Variabile discreta utilizzata per tenere traccia del trattamento scelto $[Meformina/Placebo]$

Mentre per quanto riguarda i datasets ridotti, le features utilizzate sono:

- **G_30min**: Valore di glucosio nel sangue a 30 minuti $[mg/dL]$
- **I_30min**: Valore di insulina nel sangue a 30 minuti $[mg/dL]$
- **G_60min**: Valore di glucosio nel sangue a 60 minuti $[mg/dL]$
- **I_60min**: Valore di insulina nel sangue a 60 minuti $[mg/dL]$
- **G_90min**: Valore di glucosio nel sangue a 90 minuti $[mg/dL]$
- **I_90min**: Valore di insulina nel sangue a 90 minuti $[mg/dL]$
- **G_120min**: Valore di glucosio nel sangue a 120 minuti $[mg/dL]$
- **I_120min**: Valore di insulina nel sangue a 120 minuti $[mg/dL]$
- **AGE**: Età del paziente $[anni]$
- **BMI**: Indice di massa corporea del paziente $[Kg/m^2]$
- **BW**: Peso del paziente [Kg]
- **BSA**: Area di superificie corporea $[m^2]$
- **Gb**: Glucosio basale $[mg/dL]$
- **Gender**: Sesso $[M/F]$
- **Height**: Altezza $[cm]$
- **Ib**: Insulina basale $[pmol/L]$
- **Treat.Plac**: Variabile discreta utilizzata per tenere traccia del trattamento scelto $[Meformina/Placebo]$

Sia nel caso completo che nel caso ridotto, i datasets ottenuti sono stati successivamente uniti seguendo 2 step:

1. Dagli 8 datasets inizialmente ottenuti dal simulatore, che ricordiamo essere 2 relativi a soggetti con T2D in fase iniziale (1 derivante dal trattamento con Metformina ed 1 derivante dal controllo in placebo) e 6 relativi a soggetti con T2D in fase avanzata (3 derivanti dal trattamento con Metformina e 3 derivanti dal controllo in placebo), si sono ottenuti in totale 4 nuovi datasets:
   (a) T2D Early Phase. Treatment + Placebo. Dimensioni 200x33
   (b) T2D Advanced Phase A. Treatment + Placebo. Dimensioni 200x33

(c) T2D Advanced Phase B. Treatment + Placebo. Dimensioni 200x33

(d) T2D Advanced Phase C. Treatment + Placebo. Dimensioni 200x33

2. I 4 datasets appena presentati sono stati, per metà del lavoro di tesi utilizzati in maniera separata, e per un'altra metà uniti a formare due datasets distinti (ognuno dei quali utilizzato in specifiche applicazioni che saranno discusse nelle sezioni successive del seguente elaborato):

   (a) T2D Complete. Early + Advanced (Treatment + Placebo). Dimensioni 800x34. Da notare che successivamente all'unione dei 4 dataset, è stata aggiunta una feature che identifica i pazienti con diabete di tipo 2 in fase iniziale e quelli con diabete di tipo 2 in fase avanzata: **T2DType** [Early/Advanced]

   (b) T2D Reduct. Early + Advanced (Treatment + Placebo). Dimensioni 800x18. Anche in questo caso è stata aggiunta una feature che identifica lo stadio della patologia: **T2DType** [Early/Advanced]

A scopo di semplificazione, di seguito (Figura 2.1) è rappresentato l'iter di ottenimento dei datasets precedentemente descritto. Le scelte del numero di pazienti e la tipologia di features utilizzate per la creazione dei datasets saranno discusse nei successivi capitoli del seguente elaborato di tesi.

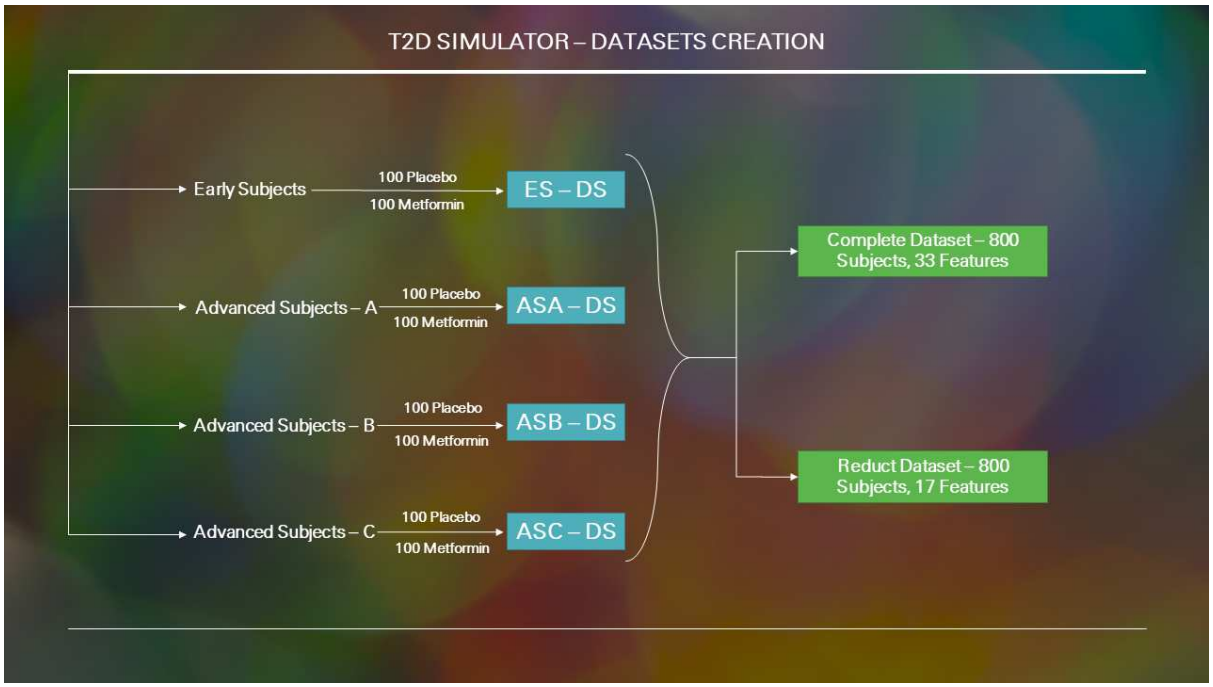

Figura 2.1: Schema riassuntivo del processo di ottenimento dei datasets a partire dai dati raccolti dalle simulazioni svolte su 4 diverse coorti di pazienti

2.2 Preprocessing

Su ognuno dei dataset ricavati dal simulatore sono state eseguite operazioni di analisi e preprocessing. Avendo utilizzato diversi modelli per diverse applicazioni, ogni dataset è stato adattato in maniera ottimale al caso in esame. Si sono però eseguite delle operazioni preliminari comuni ad ogni caso. Esse sono:

1. Ricerca e successiva rimozione degli outliers presenti nelle features
2. Feature Selection e analisi di correlazione:
   (a) Pre-selezione fatta tramite matrice di correlazione per tutti i modelli
   (b) Metodo di selezione delle feature "stepwise backward" applicato nel caso della regressione
3. Split dei dataset in una porzione per il training (75%) ed il test (25%) dei modelli
4. Taratura degli iperparametri dei modelli tramite "Cross Validation", applicata ai training set

Da notare che lo step di eventuale imputazione dei "missing value", comunemente eseguito in ogni studio nell'ambito del Machine Learning, in questo caso specifico è stato omesso, dal momento che sappiamo con sicurezza non esserci valori mancanti nei dataset generati.

2.2.1 Rimozione degli outliers

Soggetti in fase iniziale

Il primo step di preprocessing dei dati è stato quello relativo alla ricerca ed eventuale rimozione degli outliers, passo fondamentale per evitare che valori troppo distanti dalla distribuzione media di ogni feature possano influenzare negativamente i risultati. Ai fini di questo scopo sono stati analizzati i boxplot di ogni feature presente all'interno dei dataset, e la presenza di outlier è stata riscontrata in 14 features, nello specifico:

- **AUCob_G**: 1 outlier
- **Min_Value_G**: 2 outliers
- **AUC_I**: 10 outliers
- **Max_Value_I**: 7 outliers
- **Mean_Value_I**: 6 outliers
- **Min_Value_I**: 5 outliers
- **AUCob_CP**: 2 outliers
- **Min_Value_CP**: 3 outliers
- **AGE**: Presi in considerazione solo soggetti sotto i 100 anni
- **EGPb**: 1 outlier
- **k1**: 2 outliers

- **k2**: 5 outliers
- **PHId** (ϕ_d): 14 outliers
- **Vmx**: 5 outliers

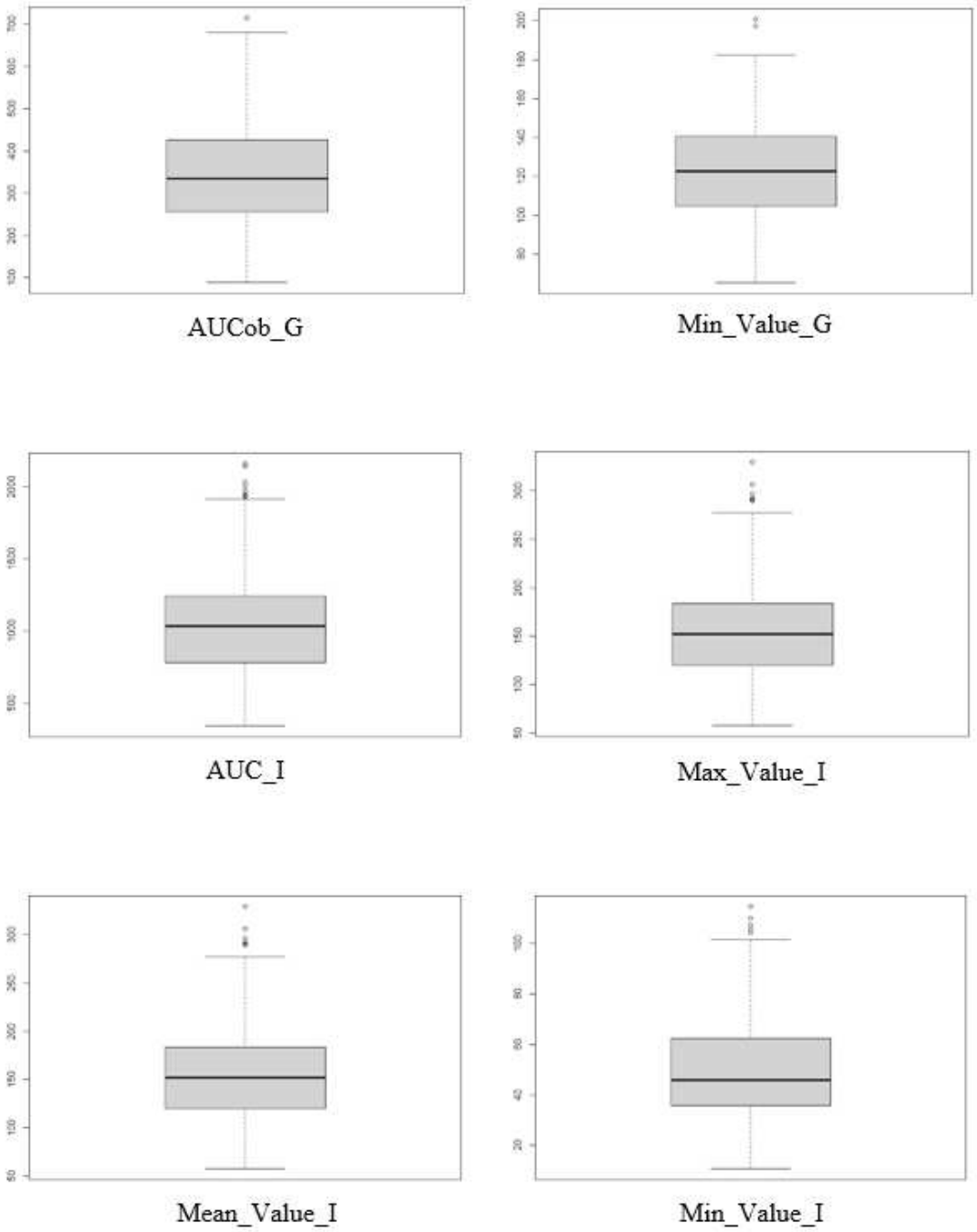

Figura 2.2: Boxplot delle features che presentano outliers visibili: AUCob_G, Min_Value_G, AUC_I, Max_Value_I, Mean_Value_I, Min_Value_I

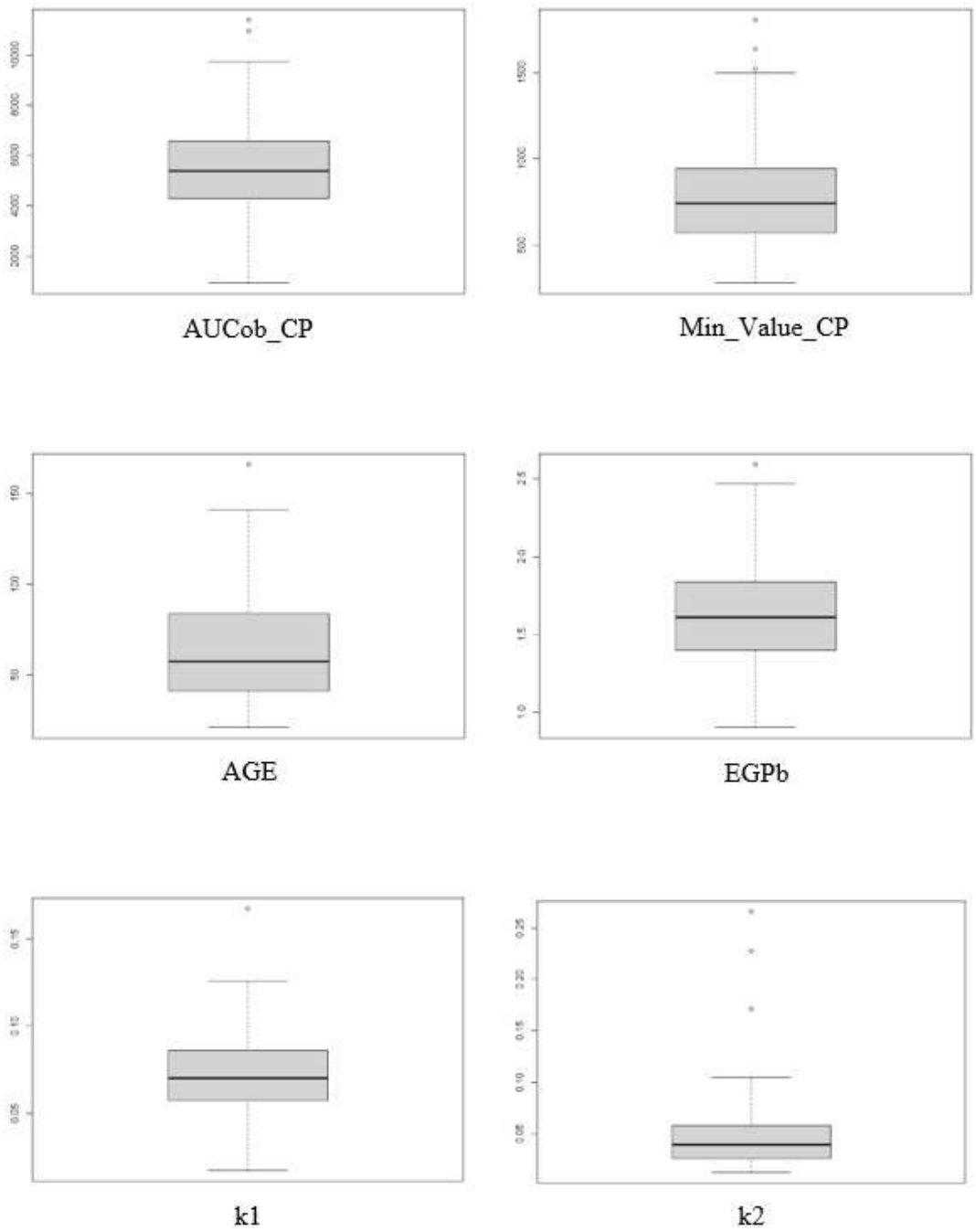

Figura 2.3: Boxplot delle features che presentano outliers visibili: AUCob_CP, Min_Value_CP, AGE, EGPb, k1, k2

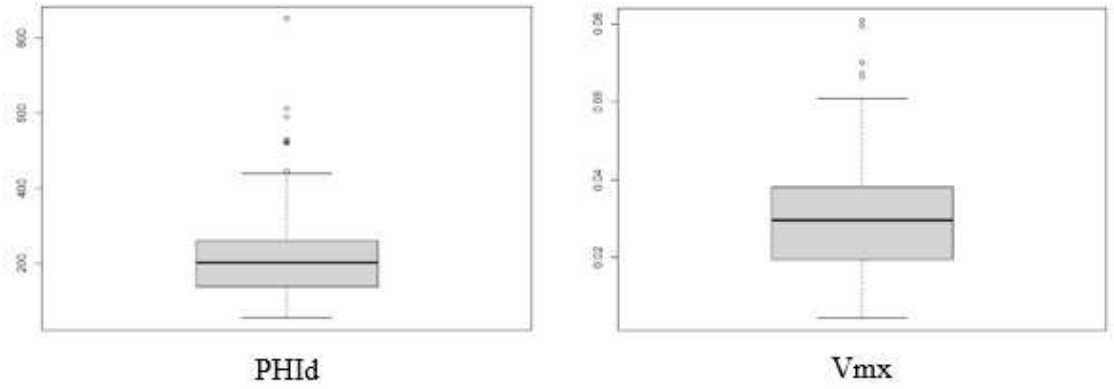

Figura 2.4: Boxplot delle features che presentano outliers visibili: PHId, Vmx

Dai grafici mostrati è immediato osservare che, pur essendoci degli outliers, essi non sono troppo distanti dai valori medi delle distribuzioni di ogni feature, di conseguenza è stata fatta una considerazione sul numero di sample a disposizione, ed essendo esso relativamente piccolo (solo 200 soggetti con T2D in fase iniziale: 100 trattati con metformina + 100 in placebo), è stato deciso di tenere in considerazione tutti i sample, eliminando solo i soggetti con età superiore a 100 anni. Da notare che la decisione presa è confermata dal fatto che, seppure alcuni valori si allontanino dalle medie di appartenenza, risultano comunque essere coerenti con la fisiologia di un essere umano.

Soggetti in fase avanzata

Per i soggetti in fase avanzata valgono le stesse considerazioni precedentemente esposte. Gli outliers trovati non si distaccano in maniera eccessiva dalle distribuzioni di appartenenza, di conseguenza nessuno di essi è stato eliminato dai dataset. Queste considerazioni si sono ritenute necessarie per lo svolgimento del lavoro di tesi a causa del numero ridotto di sample a disposizione.

2.2.2 Feature Selection

Correlation Matrix

Il secondo step è stato quello della "feature selection", fondamentale per assicurare una corretta stima dei coefficienti dei modelli utilizzati nello studio. Inizialmente si è eseguita una preselezione delle features tramite l'utilizzo della cosiddetta "Correlation Matrix". Avendo lavorato con 8 dataset diversi, sono state valutate altrettante matrici di correlazione, per brevità nella descrizione del metodo, in questa sezione dell'elaborato di tesi si fa riferimento solo al caso di soggetti affetti da diabete di tipo 2 in fase iniziale e si ragioni per analogia sugli altri casi. Di seguito sono rappresentate due matrici di correlazione, una riferita al dataset prima della rimozione delle features altamente correlate (Figura 2.5), l'altra riferita al dataset successivamente alla rimozione delle features correlate (Figura 2.6).

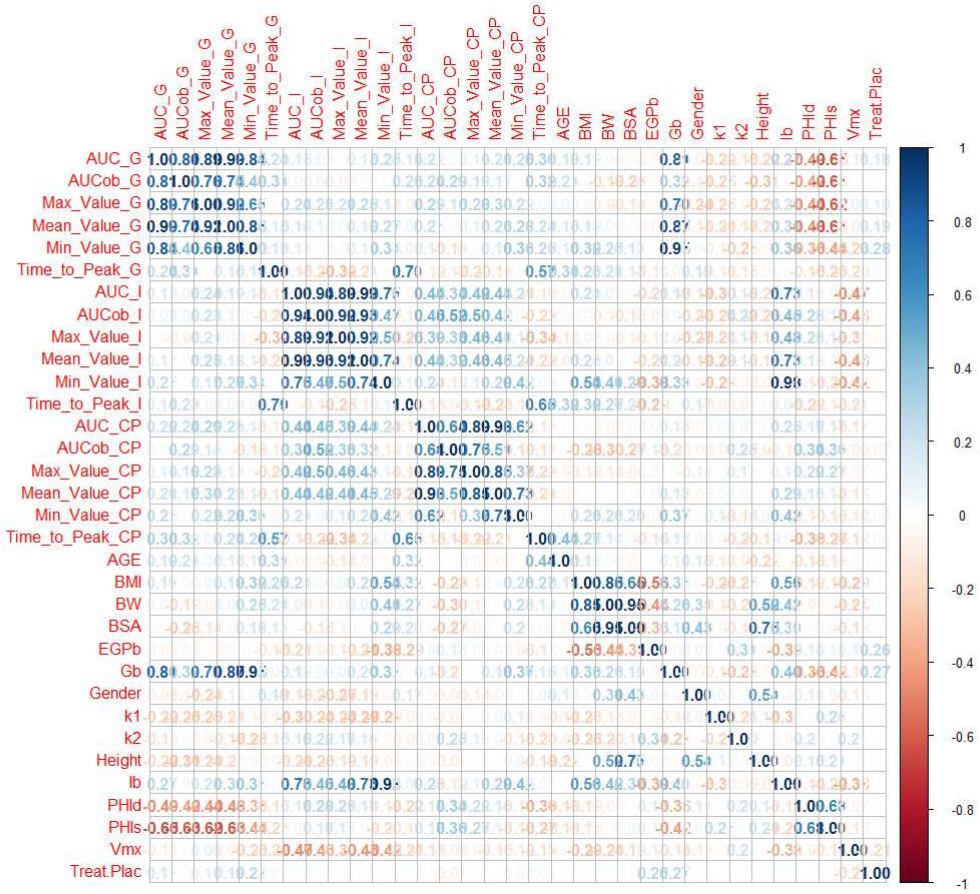

Figura 2.5: Matrice di correlazione di dati relativi a soggetti affetti da diabete di tipo 2 in fase iniziale che tiene in considerazione tutte le features presenti nel dataset

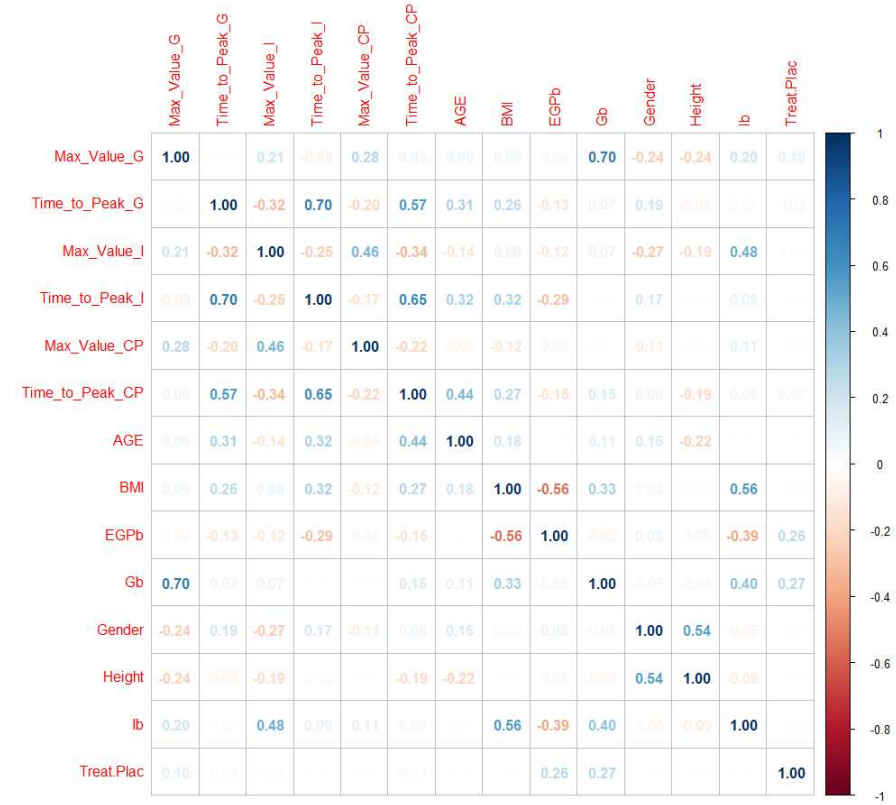

Figura 2.6: Matrice ottenuta successivamente alla rimozione delle features con alto coefficiente di correlazione

Stepwise Method

Per alcuni modelli di regressione si è proceduto all'applicazione di un metodo più strutturato per la selezione delle features ottimali: lo "Stepwise Backward Selection Method" [7]. Il suddetto metodo agisce andando a fittare tutti i possibili subsets di features, e selezionando il modello con le migliori performance basandosi su un criterio specifico. Il criterio utilizzato in questo caso è il cosiddetto "Akaike's Information Criterion (AIC)" [2]:

$$AIC = -2log(\hat{L}) + 2q$$

Dove $\hat{L}$ rappresenta la Likelihood del modello, e q rappresenta il numero di parametri. L'algoritmo di Stepwise Backward Selection agisce andando a minimizzare il valore di AIC, trovando quindi il giusto compromesso fra complessità e fit dei parametri. Esso viene definito "Stepwise Backward" perché parte dal modello completo, ovvero che contiene tutte le features a disposizione, ed una per volta va a rimuovere le features che risultano essere meno significative per la predizione dell'outcome di interesse. Grazie all'utilizzo di questo metodo si sono ottenuti dei modelli con un numero di features inferiore rispetto a quello dei dataset ottenuti dalla simulazione, quindi con un risparmio dal punto di vista computazionale e della complessità, senza però avere problemi in termini di performance dei modelli stessi.

2.2.3 Data Split

Uno step necessario per il corretto svolgimento di ogni studio nell'ambito del Machine Learning è la suddivisione di un dataset iniziale in una porzione utilizzata per il training del modello (solitamente una percentuale molto alta del totale, circa 70 - 80%) ed in una porzione utilizzata per il test del modello sviluppato (solitamente una percentuale minore, circa il 20 - 30%). Suddetto step è stato applicato ai datasets precedentemente presentati, in particolare utilizzando una percentuale del 75% per la creazione dei training set, e lasciando la restante parte (25%) a scopo di test set. La creazione dei subsets è stata eseguita andando ad estrarre in maniera randomica i sample dai datasets originari, così da garantire la massima casualità e l'assenza di bias che avrebbero potuto condizionare le future misure. Da sottolineare che è stato eseguito uno step preliminare, necessario per correggere lo sbilanciamento fra le classi "0 = early phase" ed " 1 = advanced phase" come mostrato in Figura 2.7.

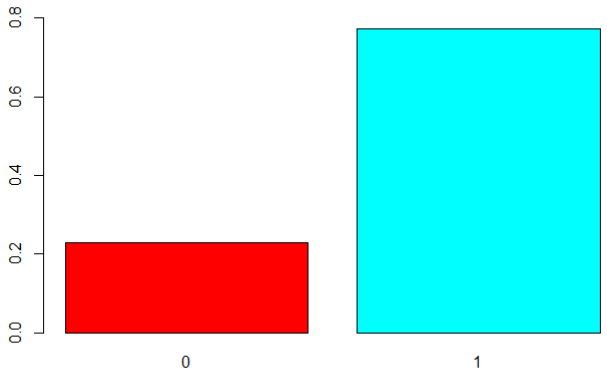

Figura 2.7: Distribuzione percentuale dei soggetti all'interno delle due classi utilizzate per la predizione (0 per "early phase" ed 1 per "advanced phase") precedentemente all'applicazione dell'algoritmo di resampling

Come si osserva in Figura 2.7, le classi risultano essere fortemente sbilanciate, con un valore di 200 soggetti affetti da diabete di tipo 2 in fase iniziale, e ben 600 soggetti in fase avanzata. Visto il numero limitato di sample a disposizione si è scelto, piuttosto che eliminare soggetti dalla classe a numero maggiore per equilibrare le due classi, di utilizzare un algoritmo di resampling che opera contemporaneamente over-sampling sulla classe a numero inferiore, ed under-sampling sulla classe a numero maggiore. In sostanza l'algoritmo elimina randomicamente alcuni sample dalla classe più popolata e duplica randomicamente un certo numero di sample nella classe meno popolata. Il risultato è mostrato in Figura 2.8, nella quale possiamo notare un bilanciamento delle due classi attorno ad un valore del 50%.

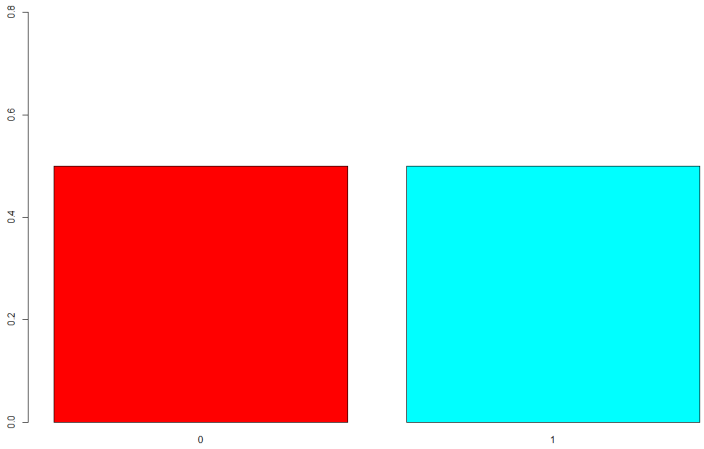

Figura 2.8: Distribuzione percentuale dei soggetti all'interno delle due classi utilizzate per la predizione (0 per "early phase" ed 1 per "advanced phase") successivamente all'applicazione dell'algoritmo di resampling

La sintesi dei dati risulta essere un processo quasi sempre sconsigliato, ma in questo caso specifico necessario ai fini della conduzione dello studio, avendo a disposizione un numero di soggetti molto ridotto. A scopo di smentita sono state condotte delle analisi parallele su un dataset composto da soli 400 pazienti: 200 in early phase e 200 in advanced phase, in modo tale da avere un equilibrio perfetto fra le classi, ed i valori ottenuti sono risultati essere sostanzialmente uguali a quelli ottenuti utilizzando il dataset per intero dopo l'applicazione del metodo di resampling. Di conseguenza si è ritenuto corretto lavorare con il dataset post-resampling, dividendolo in training set (75%) e test set (25%).

2.2.4 Cross Validation

L'ultimo step di preprocessing risulta essere la Cross Validation [22], fondamentale per una quantificazione delle potenziali performance del modello e per la successiva taratura di eventuali iperparametri. Il metodo è stato applicato ai training set ottenuti dai datasets iniziali, e nello specifico è definito "k-fold Cross Validation". Esso consiste nella suddivisione dell'insieme di dati in k parti di uguale numero e, ad ogni passo, la k-esima parte dell'insieme di dati viene utilizzata come "validation set", mentre la restante parte costituisce il training set. In questo modo è possibile addestrare il modello scelto su ognuna delle k parti ed ottenere una misura delle performance del modello andando a mediare le singole performance ottenute su ogni parte. Di seguito (Figura 2.9) è riportato uno schema riassuntivo del procedimento generico di Cross Validation, dove k = 5.

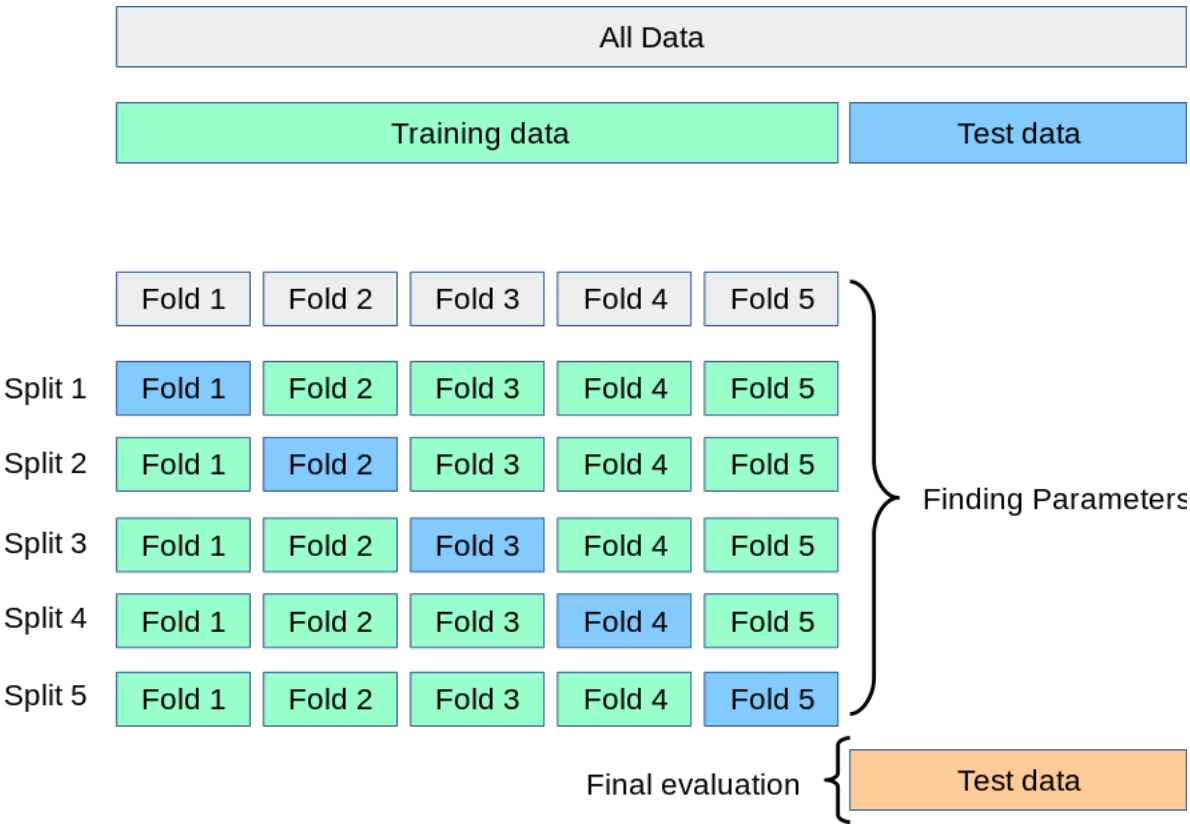

Figura 2.9: Schema esplicativo del processo di k-fold Cross Validation, con k = 5

Capitolo 3

Metodi e Modelli

Premessa

In questo capitolo sono esposte le metodologie ed i modelli utilizzati nelle fasi di ricerca. Sebbene sia il T2D Simulator, che lo script programmato ad hoc per l'estrazione delle features e la creazione dei datasets siano scritti in linguaggio MATLAB, per l'applicazione dei vari modelli di Machine Learning si è deciso di utilizzare il linguaggio "R". La scelta è motivata dalla maggiore versatilità e semplicità di utilizzo, oltre che alla presenza di numerosi strumenti essenziali ai fini delle analisi condotte sui dati raccolti. Come già anticipato, sono state intraprese due strade differenti: la **Classificazione** e la **Regressione**, descritte di seguito. Si riportano inoltre il metodo della **Principal Component Analysis**, utilizzato per l'analisi esplorativa dei dati a disposizione, ed una breve digressione sulle metriche utilizzate per la misura delle performance dei modelli sviluppati.

3.1 Classificazione

La prima parte dello studio ha riguardato, come già detto, la Classificazione. Trattasi di un'attività di apprendimento automatico che ha l'obiettivo di utilizzare le caratteristiche di un oggetto (features) per identificare a quale classe (o gruppo) appartiene. Le considerazioni preliminari che hanno portato alla scelta di questa tipologia di analisi sono basate su 2 idee:

- Studiare la possibilità di distinguere fra soggetti diabetici trattati con Metformina e soggetti diabetici non trattati
- Valutare e classificare in maniera accurata i soggetti diabetici basandosi sullo stadio della propria patologia (fase iniziale o fase avanzata)

A tal proposito sono state condotte 3 diverse tipologie di test:

1. Classificazione fra soggetti trattati con Metformina e soggetti usati come controllo in Placebo, utilizzando il dataset completo
2. Classificazione fra soggetti con diabete di tipo 2 in fase iniziale e soggetti in fase avanzata, utilizzando il dataset completo
3. Classificazione fra soggetti con diabete di tipo 2 in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto

Per tutte le tipologie di analisi condotte appena elencate, si sono testate le performance dei principali modelli utilizzati per la classificazione nell'ambito del "Supervised Learning", nello specifico:

- Regressione Logistica
- k-Nearest Neighbor
- Naive-Bayes Classifier
- Classification Trees
- Random Forests
- Support Vector Machine

3.1.1 Regressione Logistica

In statistica e in Machine Learning, la regressione logistica, è un modello di regressione non lineare utilizzato in caso di variabile dipendentente di tipo dicotomico [21]. L'obiettivo del modello è di stabilire la probabilità con cui una osservazione può generare uno o l'altro valore della variabile dipendente, e di classificare le osservazioni partendo dalle caratteristiche di queste ultime, in due categorie, grazie alla selezione di una soglia di probabilità. Il modello di regressione logistica è definito dalla cosiddetta "funzione logit":

$$\phi(z) = 1/[1 + exp(-z)] \tag{3.1}$$

Suddetta funzione descrive un tipico andamento ad "S", ed assume valori compresi fra [0,1]. Il modello di regressione logistica fa parte dei modelli lineari generalizzati, così come il modello "probit" ed il modello "loglineare", dai quali differisce sostanzialmente per la scelta della funzione fondamentale.

3.1.2 k-Nearest Neighbor

L'algoritmo k-NN (k-Nearest Neighbors) [18], è un classificatore non parametrico che si basa sulle caratteristiche degli oggetti vicini a quello considerato per il riconoscimento di pattern per la classificazione. Un oggetto è classificato da un voto di pluralità dei suoi vicini, con l'oggetto assegnato alla classe più comune tra i suoi k vicini più vicini (con k numero intero positivo, tipicamente piccolo). La scelta di k risulta essere di fondamentale importanza e dipende dalle caratteristiche dei dati. Generalmente all'aumentare di k si riduce il rumore che compromette la classificazione, ma il criterio di scelta per la classe diventa più labile. La scelta del parametro k può essere fatta grazie all'ausilio di metodi di "cross validation".

3.1.3 Naive-Bayes Classifier

Il classificatore Bayesiano è un classificatore basato sull'applicazione del teorema di Bayes [9], definito come:

$$p(A|B) = [p(A) * p(B|A)]/p(B) \tag{3.2}$$

Il classificatore richiede dunque la conoscenza delle probabilità a priori ($p(A)$ e $p(B)$) e condizionali ($p(B|A)$) relative al problema, quantità che in generale non sono note ma sono tipicamente stimabili. Se è possibile ottenere delle stime affidabili delle probabilità coinvolte nel teorema, il classificatore bayesiano risulta generalmente affidabile e potenzialmente compatto.

3.1.4 Classification Trees

Un classification tree è un algoritmo di apprendimento supervisionato non parametrico utilizzato per la classificazione [19]. Si compone di una struttura ad albero gerarchica, che consiste di un nodo radice, di rami, di nodi interni e di nodi finali. Questo tipo di struttura crea una rappresentazione facile da interpretare per quanto riguarda un eventuale processo decisionale. I metodi basati su alberi possono essere usati sia per problemi di classificazione che problemi di regressione. Questi problemi richiedono la stratificazione o segmentazione dello spazio dei predittori in un numero di regioni semplici. Una previsione per una data osservazione è quindi ottenuta tipicamente usando la media o la moda delle osservazioni di training nella regione in cui ricade l'osservazione. Per la crescita di un classification tree, usualmente si utilizza una suddivisione binaria, ricercando ad ogni nodo la minima variabilità nel rispetto di specifiche metriche, nello specifico:

- L'indice Gini [16], che fornisce una misura dell'incertezza totale tra le k classi. Spesso è indicato anche con il nome di misura di "purezza" del nodo
- La cross-entropia [3], che, come l'indice Gini, assume un valore piccolo se il nodo è puro

3.1.5 Random Forests

Una random forest è un classificatore ottenuto dall'aggregazione tramite bagging di un numero elevato di decision tree [8]. L'algoritmo presenta tre iperparametri principali, che devono essere impostati prima dell'addestramento. Questi parametri includono la dimensione dei nodi, il numero di decision tree e il numero di caratteristiche campionate.

3.1.6 Support Vector Machine

Le Support Vector Machine (SVM [4]) sono dei modelli di apprendimento supervisionato utilizzabili sia per la classificazione che per la regressione. Dato un insieme di esempi per l'addestramento, ognuno dei quali etichettato con la classe di appartenenza fra due possibili classi, un algoritmo di addestramento per le SVM costruisce un modello che assegna i nuovi esempi a una delle due classi, ottenendo quindi un classificatore lineare binario. Un modello SVM è una rappresentazione degli esempi come punti nello spazio, mappati in modo tale che gli esempi appartenenti alle due diverse categorie siano chiaramente separati da uno spazio il più possibile ampio. I nuovi esempi sono quindi mappati nello stesso spazio e la predizione della categoria alla quale appartengono viene fatta sulla base del lato nel quale ricade. Formalmente, una SVM costruisce un iperpiano (o un insieme di iperpiani) in uno spazio a più dimensioni o ad infinite dimensioni, il quale può essere usato per la classificazione.

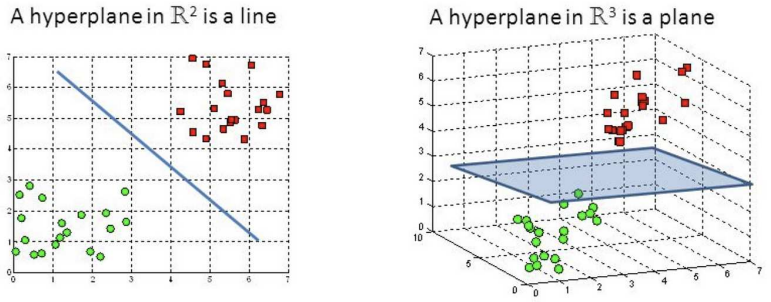

Figura 3.1: Esempio di costruzione di iperpiani nelle SVM in 2-D (a sinistra) e 3-D (a destra)

Oltre alla classificazione lineare, è possibile fare uso delle SVM per svolgere efficacemente la classificazione non lineare utilizzando il "metodo Kernel" [11], molto utile nel caso in cui gli insiemi da distinguere non siano linearmente separabili. Per questo motivo è stato proposto che lo spazio originale di dimensioni finite venisse mappato in uno spazio con un numero di dimensioni maggiore, rendendo presumibilmente più facile trovare una separazione. Per mantenere il carico computazionale accettabile, le mappature utilizzate dalle SVM sono fatte in modo tale che i prodotti scalari dei vettori delle coppie di punti in ingresso siano calcolati facilmente in termini delle variabili dello spazio originale, attraverso la loro definizione in termini di una funzione Kernel scelta in base al problema da risolvere. Esistono varie tipologie di funzioni Kernel ma, nel caso specifico di questo elaborato di tesi, ne sono state tenute in considerazione solamente 3: Kernel polinomiale, Kernel radiale, Kernel lineare. Fra le 3 funzioni testate, quella di tipo radiale ha consentito l'ottenimento di performance ottimali.

3.2 Regressione

Per la seconda parte del seguente elaborato di tesi sono state condotte analisi di regressione sui dataset a disposizione. In generale si tratta di tecniche utilizzate per analizzare le relazioni fra una variabile dipendente ed una o più variabili indipendenti, e stimare un'eventuale relazione funzionale. Nello specifico le variabili dipendenti tenute in considerazione e rispetto le quali sono state condotte le analisi di regressione sono:

- ϕ_d: Reattività delle cellule β al variare della velocità di variazione del glucosio $[10^{-9}]$
- ϕ_s: Reattività delle cellule β al glucosio $[10^{-9}\ min^{-1}]$
- V_{mx}: Sensibilità dell'insulina all'utilizzo del glucosio $[mg/kg/min\ per\ pmol/L]$

Per ognuno dei parametri appena elencati sono stati applicati diversi modelli, preliminarmente utilizzando il dataset completo, e solo successivamente utilizzando il dataset ridotto. I modelli usati in entrambi i casi nello specifico sono:

- Regressione Lineare
- Regression Tree
- Random Forest
- Boosting

Di seguito sono riportate delle brevi digressioni sulla Regressione Lineare e il Boosting. Per Regression Tree e Random Forest si ragioni per analogia con i discorsi già fatti nel caso della classificazione.

3.2.1 Regressione Lineare

Nella regressione lineare, il modello assume che la variabile dipendente sia una combinazione lineare dei parametri (ma non è necessario che sia lineare nella variabile indipendente) [20]. Ad esempio nella regressione lineare semplice con N osservazioni ci sono una variabile indipendente x_i e due parametri β_0 e β_1:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

$$i = 1, ..., N$$

Dove y_i è la variabile dipendente, x_i è la variabile indipendente, β_0 è l'intercetta della retta di regressione della popolazione, β_1 è il coefficiente angolare della retta di regressione, ed ϵ è l'errore.

3.2.2 Boosting

Il boosting è una tecnica di machine learning che rientra nella categoria dell'apprendimento d'ensemble [6]. Nel boosting più modelli vengono generati consecutivamente dando sempre più peso agli errori effettuati nei modelli precedenti, ottenendo infine un modello aggregato avente migliore accuratezza di ciascun modello che lo costituisce. Ogni qual volta un modello viene addestrato, ci sarà una fase di ripesaggio delle istanze. L'algoritmo di boosting tenderà a dare un peso maggiore alle istanze misclassificate, nella speranza che il successivo modello sia più esperto su quest'ultime.

3.3 PCA - Principal Component Analysis

La Principal Component Analysis, abbreviata PCA [23], è una tecnica per la semplificazione dei dati, che ha lo scopo di ridurre il numero più o meno elevato di variabili che descrivono un insieme di dati ad un numero minore di variabili latenti, limitando il più possibile la perdita di informazioni. Ciò avviene tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano in cui la nuova variabile con la maggiore varianza viene proiettata sul primo asse, la variabile nuova, seconda per dimensione della varianza, sul secondo asse e così via. Diversamente da altre trasformazioni lineari di variabili praticate nell'ambito della statistica, in questa tecnica sono gli stessi dati che determinano i vettori di trasformazione. La PCA è stata utilizzata, per lo sviluppo di questo elaborato di tesi, per l'analisi esplorativa dei datasets descritti precedentemente, di seguito se ne riporta un breve riassunto.

3.3.1 PCA - Dataset Completo

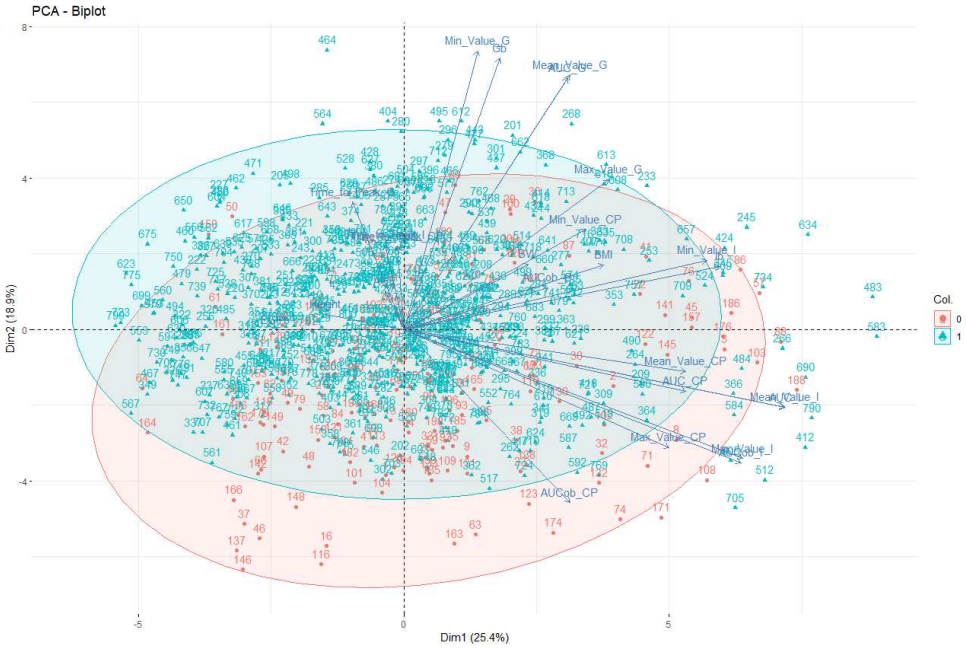

Figura 3.2: Biplot PCA su dataset completo, relativo alle prime due componenti principali

Figura 3.3: Biplot PCA su dataset completo, relativo alla seconda ed alla terza componente principale

I grafici mostrati in Figura 3.2 e Figura 3.3 mostrano rispettivamente i biplot relativi alla prima componente principale in relazione alla seconda, ed alla seconda componente principale in relazione alla terza. In entrambi i casi sembra esserci una leggera distinzione fra i soggetti diabetici in fase iniziale ("0") e soggetti diabetici in fase avanzata ("1").

3.3.2 PCA - Dataset Ridotto

Figura 3.4: Biplot PCA su dataset ridotto, relativo alle prime due componenti principali

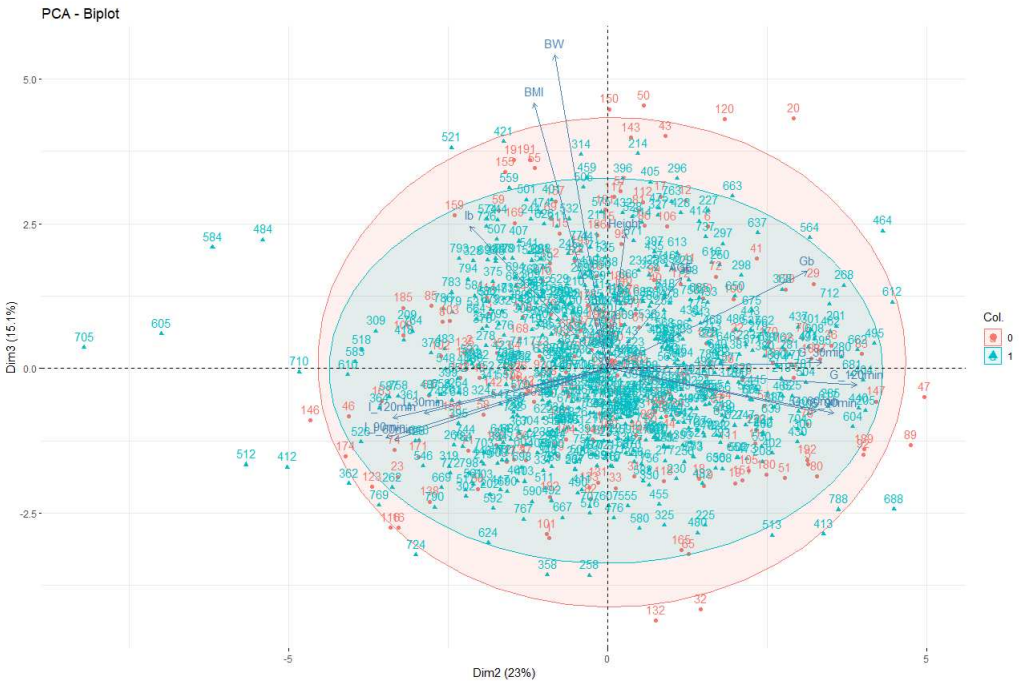

Figura 3.5: Biplot PCA su dataset ridotto, relativo alla seconda ed alla terza componente principale

Per quanto riguarda il dataset ridotto, il discorso è diverso, in questo caso sembra non esserci una distinzione fra i due gruppi da classificare, come si può osservare dalle Figure 3.4 e 3.5.

3.4 Metriche di valutazione

Per la valutazione delle performance in termini di predizione dei modelli precedentemente elencati, si sono utilizzate delle metriche differenti per classificazione e regressione. Gli indici scelti sono stati valutati, successivamente all'applicazione degli algoritmi adeguatamente addestrati, sui dataset di test, per verificare le performance dei modelli nel predire su nuovi dati in input.

3.4.1 Classificazione

Per quanto riguarda la classificazione, gli indici utilizzati sono stati ricavati a partire dalla cosiddetta confusion matrix, nello specifico si sono calcolate:

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$

Dove TP sta per $TruePositive$, TN sta per $TrueNegative$, FP sta per $FalsePositive$, e FN sta per $FalseNegative$. Tutti e tre gli indici hanno valori compresi fra [0,1]. Più i valori delle metriche si avvicinano ad 1 e migliori risultano essere le performance dei modelli.

3.4.2 Regressione

Per la regressione, vista la natura differente dei modelli e non avendo a disposizione delle misure di predizione discrete, bensì valori continui, si sono utilizzate due metriche differenti per la misura delle performance:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$$

$$NRMSE = \frac{RMSE}{mean(y)}$$

Dove RMSE sta per "Root Mean Square Error", e NRMSE sta per "Normalized Root Mean Square Error". Il primo indice rappresenta la media del quadrato delle differenze fra i valori predetti dal modello e i valori realmente osservati. Il secondo indice è semplicemente RMSE diviso per la media delle osservazioni. Quest'ultimo valore è stato calcolato per facilitare la comparazione fra modelli e dati con scale differenti. In generale, più gli indici si avvicinano a 0 e migliori risultano essere le performance dei modelli.

Capitolo 4

Risultati e discussione

4.1 Classificazione

4.1.1 Classificazione fra soggetti trattati con Metformina e soggetti in placebo - Dataset Completo [800x34]

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.750	0.568	0.636	0.546	0.659	0.727
Sensitivity	0.741	0.296	0.630	0.647	0.706	0.609
Specificity	0.765	1.000	0.640	0.482	0.630	0.857

Tabella 4.1: Risultati dei modelli testati per la classificazione fra soggetti trattati con Metformina e soggetti in placebo

I modelli testati per la classificazione fra soggetti trattati con Metformina e soggetti in placebo utilizzando il dataset completo, hanno portato all'ottenimento dei risultati riassunti in Tabella 4.1. Per semplicità di lettura ricordiamo che gli algoritmi utilizzati sono: Logistic Regression (LR), k-Nearest Neighbor (k-NN), Naive-Bayes (NB), Classification Trees (CT), Random Forests (RF), Support Vector Machines (SVM). Dai dati mostrati è immediato osservare che le performance medie non sono particolarmente buone, sebbene la Logistic Regression e le Support Vector Machine abbiano valori migliori rispetto agli altri modelli testati, è comunque risultato complesso classificare in maniera adeguata i soggetti trattati con Metformina e i soggetti in placebo.

4.1.2 Classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata - Dataset Completo [800x34]

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.935	0.852	0.834	0.785	0.935	0.975
Sensitivity	0.930	0.717	0.792	0.870	0.920	0.979
Specificity	0.940	0.930	0.867	0.700	0.950	0.970

Tabella 4.2: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset completo

La seconda tranche di analisi ha riguardato la classificazione fra soggetti con diabete di tipo 2 in fase iniziale e soggetti in fase avanzata. Come si può osservare dai dati riassunti in Tabella 4.2, le performance medie risultano essere molto buone, con una particolare attenzione da dedicare alla Support Vector Machine, avente valori prossimi ad 1. Gli

algoritmi riescono a distinguere in maniera accurata i pazienti in fase iniziale dai pazienti in fase avanzata. L'ottenimento dei risultati appena presentati è diretta conseguenza della scelta di utilizzare un dataset ridotto, contenente solo valori di glucosio ed insulina a vari punti nel tempo, e specifici parametri fisiologici misurabili in maniera semplice e rapida.

4.1.3 Classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata - Dataset Ridotto [800x18]

La terza fase dello studio di classificazione vede impiegato l'utilizzo del dataset ridotto già ampiamente descritto. Per osservare in maniera accurata il contributo dell'insulina e del glucosio al variare del tempo, si è deciso di proseguire lo studio svolgendo 6 test diversi:

1. Test dei modelli usando una porzione del dataset ridotto contenente solo i seguenti parametri:

   - Age
   - BMI
   - BW
   - Gb
   - Gender
   - Height
   - Ib

2. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 30 minuti durante il test OGTT simulato

3. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 60 minuti

4. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 90 minuti

5. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 120 minuti

6. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 60 e 120 minuti

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.764	0.796	0.736	0.880	0.975	0.955
Sensitivity	0.769	0.684	0.718	0.795	0.980	0.951
Specificity	0.759	0.895	0.752	0.955	0.970	0.960

Tabella 4.3: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente solo i parametri fisiologici

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.856	0.796	0.728	0.884	0.968	0.984
Sensitivity	0.889	0.726	0.718	0.863	0.940	0.991
Specificity	0.827	0.857	0.737	0.902	0.993	0.978

Tabella 4.4: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 30 minuti

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.916	0.840	0.844	0.872	0.964	0.988
Sensitivity	0.891	0.717	0.696	0.837	0.924	0.983
Specificity	0.930	0.911	0.930	0.892	0.987	0.992

Tabella 4.5: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 60 minuti

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.880	0.772	0.800	0.912	0.968	0.980
Sensitivity	0.846	0.846	0.752	0.889	0.940	0.975
Specificity	0.910	0.707	0.842	0.932	0.993	0.985

Tabella 4.6: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 90 minuti

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.840	0.772	0.788	0.888	0.960	0.984
Sensitivity	0.812	0.745	0.761	0.846	0.949	0.983
Specificity	0.865	0.797	0.812	0.925	0.970	0.985

Tabella 4.7: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 120 minuti

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.928	0.804	0.804	0.900	0.972	0.996
Sensitivity	0.923	0.795	0.803	0.880	0.949	1.000
Specificity	0.932	0.812	0.805	0.917	0.993	0.993

Tabella 4.8: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 60 e 120 minuti

Dai risultati appena mostrati è immediato osservare che non ci sono sostanziali differenze in termini di performance medie fra l'utilizzo del dataset completo e quello ridotto. Da notare che, nel caso in cui si sono utilizzati solo parametri fisiologici (Tabella 4.3), le perfomance ottenute con Random Forest e SVM sono sostanzialmente uguali ai casi con dataset ridotto che considerano i valori di glucosio ed insulina a vari punti nel tempo (Tabelle 4.4, 4.5, 4.6, 4.7, 4.8), ed addirittura leggermente migliori rispetto all'utilizzo del dataset completo (Tabella 4.2). Questo risultato risulta essere di notevole importanza e suggerisce la possibilità di poter sviluppare una soluzione alternativa ai comuni test OGTT, dispendiosi in termini economici e di tempistiche.

4.2 Regressione

4.2.1 Regressione su ϕ_d

Risultati con dataset completo [800x34]

	GLM	RT	RF	Boost
RMSE	128.759	173.717	125.815	140.384
NRMSE	0.546	0.737	0.534	0.586

Tabella 4.9: Risultati dei modelli testati per la regressione su ϕ_d utilizzando il dataset completo

Per semplicità di lettura si ricorda che i modelli utilizzati per tutte le sezioni di regressione sono: Regressione Lineare (GLM), Regression Trees (RT), Random Forest (RF) e Boosting (Boost). Dalla Tabella 4.9 è immediato osservare che i modelli ricavati non descrivono sufficientemente bene le relazioni che intercorrono fra la variabile dipendente e i regressori utilizzati, di fatto le performance in termini di predizione sono pessime.

Risultati con dataset ridotto [800x18]

	GLM	RT	RF	Boost
RMSE	143.122	141.494	107.650	145.486
NRMSE	0.492	0.487	0.370	0.500

Tabella 4.10: Risultati dei modelli testati per la regressione su ϕ_d utilizzando una porzione di dataset ridotto contenente solo i parametri fisiologici

	GLM	RT	RF	Boost
RMSE	113.848	130.289	130.253	146.491
NRMSE	0.392	0.448	0.355	0.504

Tabella 4.11: Risultati dei modelli testati per la regressione su ϕ_d utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 30 minuti

	GLM	RT	RF	Boost
RMSE	124.650	134.605	104.814	127.241
NRMSE	0.429	0.463	0.361	0.438

Tabella 4.12: Risultati dei modelli testati per la regressione su ϕ_d utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 60 minuti

	GLM	RT	RF	Boost
RMSE	125.932	135.642	102.813	128.352
NRMSE	0.433	0.467	0.354	0.442

Tabella 4.13: Risultati dei modelli testati per la regressione su ϕ_d utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 90 minuti

	GLM	RT	RF	Boost
RMSE	128.220	134.466	106.531	130.615
NRMSE	0.441	0.463	0.366	0.449

Tabella 4.14: Risultati dei modelli testati per la regressione su ϕ_d utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 120 minuti

	GLM	RT	RF	Boost
RMSE	124.723	134.605	109.367	129.245
NRMSE	0.429	0.463	0.376	0.445

Tabella 4.15: Risultati dei modelli testati per la regressione su ϕ_d utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 60 e 120 minuti

Anche nel caso di dataset ridotto è immediato osservare che le performance medie dei modelli ottenuti non sono particolarmente positive. Si osserva un leggero miglioramento rispetto all'utilizzo del dataset completo, ma i risultati ricadono comunque in un range di valori non accettabili.

4.2.2 Regressione su ϕ_s

Risultati con dataset completo [800x34]

	GLM	RT	RF	Boost
RMSE	4.362	6.299	5.585	5.826
NRMSE	0.243	0.350	0.311	0.309

Tabella 4.16: Risultati dei modelli testati per la regressione su ϕ_s utilizzando il dataset completo

Anche nel caso di regressione rispetto la variabile dipendente ϕ_s utilizzando il dataset completo, i risultati ottenuti non sono particolarmente soddisfacenti come si può osservare in Tabella 4.16. Da osservare che il modello di regressione lineare sembra comportarsi leggermente meglio rispetto agli altri algoritmi testati.

Risultati con dataset ridotto [800x18]

	GLM	RT	RF	Boost
RMSE	5.751	5.943	4.214	5.708
NRMSE	0.361	0.373	0.264	0.358

Tabella 4.17: Risultati dei modelli testati per la regressione su ϕ_s utilizzando una porzione di dataset ridotto contenente solo i parametri fisiologici

	GLM	RT	RF	Boost
RMSE	5.452	5.912	4.562	5.581
NRMSE	0.342	0.371	0.286	0.350

Tabella 4.18: Risultati dei modelli testati per la regressione su ϕ_s utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 30 minuti

	GLM	RT	RF	Boost
RMSE	5.189	5.368	4.292	5.410
NRMSE	0.325	0.337	0.269	0.339

Tabella 4.19: Risultati dei modelli testati per la regressione su ϕ_s utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 60 minuti

	GLM	RT	RF	Boost
RMSE	4.861	5.381	4.110	5.889
NRMSE	0.305	0.337	0.258	0.369

Tabella 4.20: Risultati dei modelli testati per la regressione su ϕ_s utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 90 minuti

	GLM	RT	RF	Boost
RMSE	4.633	5.212	4.011	5.026
NRMSE	0.291	0.327	0.252	0.315

Tabella 4.21: Risultati dei modelli testati per la regressione su ϕ_s utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 120 minuti

	GLM	RT	RF	Boost
RMSE	4.622	5.225	4.137	5.079
NRMSE	0.289	0.328	0.259	0.319

Tabella 4.22: Risultati dei modelli testati per la regressione su ϕ_s utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 60 e 120 minuti

Come è osservabile dai dati raccolti, nel caso di dataset ridotto non si presenta nessuna sostanziale differenza in termini di performance dei modelli rispetto al caso in cui è stato utilizzato il dataset completo.

4.2.3 Regressione su V_{mx}

Risultati con dataset completo [800x34]

	GLM	RT	RF	Boost
RMSE	0.01742	0.01970	0.01750	0.01750
NRMSE	0.457	0.517	0.459	0.459

Tabella 4.23: Risultati dei modelli testati per la regressione su V_{mx} utilizzando il dataset completo

Come nei due precedenti casi, anche per quanto riguarda la regressione su V_{mx}, i risultati non sono particolarmente positivi. I modelli sembrano non riuscire a predire in maniera corretta i valori di V_{mx} a partire dalle features presenti all'interno del dataset completo.

Risultati con dataset ridotto [800x18]

	GLM	RT	RF	Boost
RMSE	0.01092	0.01110	0.00802	0.01025
NRMSE	0.392	0.399	0.288	0.368

Tabella 4.24: Risultati dei modelli testati per la regressione su V_{mx} utilizzando una porzione di dataset ridotto contenente solo i parametri fisiologici

	GLM	RT	RF	Boost
RMSE	0.01073	0.01100	0.00834	0.01027
NRMSE	0.385	0.395	0.299	0.369

Tabella 4.25: Risultati dei modelli testati per la regressione su V_{mx} utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 30 minuti

	GLM	RT	RF	Boost
RMSE	0.01074	0.01064	0.00819	0.01020
NRMSE	0.386	0.382	0.294	0.366

Tabella 4.26: Risultati dei modelli testati per la regressione su V_{mx} utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 60 minuti

	GLM	RT	RF	Boost
RMSE	0.01060	0.01070	0.00819	0.01020
NRMSE	0.383	0.384	0.294	0.365

Tabella 4.27: Risultati dei modelli testati per la regressione su V_{mx} utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 90 minuti

	GLM	RT	RF	Boost
RMSE	0.01070	0.01077	0.00839	0.01021
NRMSE	0.384	0.387	0.301	0.367

Tabella 4.28: Risultati dei modelli testati per la regressione su V_{mx} utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 120 minuti

	GLM	RT	RF	Boost
RMSE	0.01059	0.01050	0.00851	0.01022
NRMSE	0.380	0.377	0.306	0.367

Tabella 4.29: Risultati dei modelli testati per la regressione su V_{mx} utilizzando una porzione di dataset ridotto contenente i parametri fisiologici e i valori di insulina e glucosio a 60 e 120 minuti

Sebbene i valori medi di RMSE ed NRMSE siano leggermente migliori in questo caso specifico, le stesse considerazioni fatte sul caso con dataset completo valgono anche nel caso di dataset ridotto.

4.3 Dati raccolti utilizzando solo 400 soggetti

Si riportano in questa sezione i dati relativi ai modelli ed ai metodi precedentemente mostrati, raccolti utilizzando solo una porzione dei datasets, contente un totale di 400 soggetti (200 con diabete di tipo due in fase iniziale e 200 in fase avanzata), in modo da avere un bilanciamento perfetto fra le due classi da predire, e di conseguenza di non necessitare di un algoritmo di sintesi dei sample, come nel caso precedentemente descritto.

4.3.1 Classificazione fra soggetti trattati con Metformina e soggetti in placebo - Dataset Completo [400x34]

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.691	0.583	0.593	0.577	0.644	0.650
Sensitivity	0.705	0.457	0.543	0.551	0.596	0.615
Specificity	0.674	0.73	0.652	0.600	0.686	0.680

Tabella 4.30: Risultati dei modelli testati per la classificazione fra soggetti trattati con Metformina e soggetti in placebo (400 sample)

Come già osservato nel caso precedente, i modelli riescono a distinguere a fatica i soggetti trattati con Metformina dai soggetti del gruppo di controllo in placebo. Confrontando i risultati raccolti Tabella 4.30 con i risultati raccolti in Tabella 4.1, si può osservare che non c'è nessuna sostanziale differenza in termini di performance medie.

4.3.2 Classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata - Dataset Completo [400x34]

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.904	0.800	0.713	0.745	0.809	0.936
Sensitivity	0.962	0.774	0.755	0.736	0.849	0.873
Specificity	0.829	0.829	0.659	0.756	0.756	1.000

Tabella 4.31: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset completo (400 sample)

Nel caso di classificazione fra soggetti con diabete di tipo 2 in fase iniziale e soggetti in fase avanzata, utilizzando un dataset contenente solo 400 soggetti e che tiene in considerazione tutte le features, otteniamo dei valori (Tabella 4.31) leggermente inferiori rispetto a quelli riassunti in Tabella 4.2, ma che in ogni caso non presentano differenze sostanziali dai dati precedentemente osservati.

4.3.3 Classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata - Dataset Ridotto [400x18]

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.798	0.872	0.819	0.766	0.904	0.930
Sensitivity	0.793	0.868	0.868	0.868	0.962	0.923
Specificity	0.805	0.878	0.756	0.634	0.829	0.939

Tabella 4.32: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente solo i parametri fisiologici (400 sample)

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.894	0.830	0.787	0.840	0.904	0.915
Sensitivity	0.906	0.792	0.868	0.906	0.981	0.902
Specificity	0.878	0.878	0.683	0.756	0.805	0.925

Tabella 4.33: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 30 minuti (400 sample)

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.862	0.755	0.798	0.745	0.904	0.957
Sensitivity	0.887	0.755	0.868	0.736	0.981	0.980
Specificity	0.829	0.756	0.707	0.756	0.805	0.930

Tabella 4.34: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 60 minuti (400 sample)

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.787	0.745	0.702	0.798	0.851	0.904
Sensitivity	0.793	0.925	0.755	0.849	0.906	0.923
Specificity	0.781	0.512	0.634	0.732	0.781	0.881

Tabella 4.35: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 90 minuti (400 sample)

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.787	0.766	0.702	0.777	0.809	0.915
Sensitivity	0.793	0.792	0.774	0.849	0.868	0.959
Specificity	0.781	0.731	0.610	0.683	0.732	0.867

Tabella 4.36: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 120 minuti (400 sample)

	LR	k-NN	NB	CT	RF	SVM
Accuracy	0.915	0.766	0.723	0.755	0.862	0.957
Sensitivity	0.943	0.792	0.811	0.793	0.925	1.000
Specificity	0.878	0.732	0.610	0.707	0.781	0.911

Tabella 4.37: Risultati dei modelli testati per la classificazione fra soggetti con T2D in fase iniziale e soggetti in fase avanzata, utilizzando il dataset ridotto contenente parametri fisiologici e valori di insulina e glucosio a 60 e 120 minuti (400 sample)

Per quanto riguarda i modelli addestrati utilizzando il dataset ridotto, anche in questo caso le performance risultano essere leggermente peggiori rispetto a quelle ottenute utilizzando il dataset contenente 800 soggetti, ma comunque non eccessivamente diverse da quelle già descritte nei precedenti paragrafi. Se ne conclude che la sintesi dei sample non influenza negativamente la veridicità delle performance dei modelli.

4.4 Approfondimento Classificazione

Dai dati raccolti e mostrati nel paragrafo 4.2 è immediato osservare che, per quanto riguarda la regressione rispetto ai 3 parametri considerati, è risultato complesso riuscire a sviluppare un modello che, partendo dai predittori a nostra disposizione, riuscisse ad avere performance adeguate. Discorso diverso vale per la classificazione fra soggetti con diabete di tipo due in fase iniziale e soggetti in fase avanzata (sezioni 4.1.2 e 4.1.3 del paragrafo 4.1). In questo caso specifico si sono ottenute delle performance medie molto buone dai modelli, con particolare attenzione da dedicare a Regressione Logistica, Classification Trees, Random Forests e Support Vector Machine, che sono discussi in maniera dettagliata di seguito.

4.4.1 Approfondimento Classificazione - Dataset Completo

Regressione Logistica

Il primo modello testato, come già detto, è stato quello della Regressione Logistica. Si è innanzitutto condotto un test preliminare addestrando un modello contenente tutti i predittori a disposizione nel dataset, ottenendo i risultati forniti in Figura 4.1.

```
Call:
glm(formula = T2DType ~ ., family = "binomial", data = ds_class_trn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4360  -0.0171   0.0000   0.0051   2.7972

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.4881205 45.9905772  -0.054 0.956855
AUC_G            0.3637962  0.0780239   4.663 3.12e-06 ***
AUCob_G         -0.0460529  0.0415581  -1.108 0.267794
Max_Value_G      0.0272201  0.0429145   0.634 0.525895
Mean_Value_G    -2.3501596  0.5000134  -4.700 2.60e-06 ***
Min_Value_G      0.1284055  0.1092060   1.176 0.239671
Time_to_Peak_G   0.0890773  0.0256222   3.477 0.000508 ***
AUC_I           -0.1134309  0.0312608  -3.629 0.000285 ***
AUCob_I          0.0277919  0.0186024   1.494 0.135177
Max_Value_I     -0.0372556  0.0115716  -3.220 0.001284 **
Mean_Value_I     0.7293974  0.1576503   4.627 3.72e-06 ***
Min_Value_I     -0.1712176  0.0815988  -2.098 0.035880 *
Time_to_Peak_I  -0.0810787  0.0217543  -3.727 0.000194 ***
AUC_CP          -0.0019039  0.0024808  -0.767 0.442811
AUCob_CP        -0.0023231  0.0011211  -2.072 0.038260 *
Max_Value_CP    -0.0016301  0.0021097  -0.773 0.439726
Mean_Value_CP    0.0218463  0.0140213   1.558 0.119215
Min_Value_CP    -0.0102521  0.0054069  -1.896 0.057945 .
Time_to_Peak_CP -0.0009495  0.0119697  -0.079 0.936776
AGE              0.0436802  0.0231762   1.885 0.059471 .
BMI              0.1500971  0.6117582   0.245 0.806183
BW              -0.4195487  0.3848202  -1.090 0.275605
BSA             14.5041190 39.6658133   0.366 0.714620
EGPb            -3.6276154  1.5514004  -2.338 0.019373 *
Gb               0.3586519  0.2016641   1.778 0.075328 .
Gender          -1.9053556  0.9642590  -1.976 0.048157 *
Height           0.0111007  0.4753111   0.023 0.981367
Ib               0.0546347  0.0898494   0.608 0.543141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.777  on 599  degrees of freedom
Residual deviance:  99.831  on 572  degrees of freedom
AIC: 155.83
```

Figura 4.1: Regressione logistica su dataset completo utilizzando tutti i predittori a disposizione

Si è poi proceduto con l'applicazione di un algoritmo di "Stepwise Backward Selection", che si occupa di selezionare il numero ottimale di features basandosi sulla minimiz-

zazione del valore dell'Akaike Information Criterion (AIC). Il modello ottenuto è stato successivamente addestrato ed i risultati sono mostrati in Figura 4.2.

```
Call:
glm(formula = T2DType ~ AUC_G + Max_Value_G + Mean_Value_G +
    Min_Value_G + Time_to_Peak_G + AUC_I + Max_Value_I + Mean_Value_I +
    Min_Value_I + Time_to_Peak_I + AUC_CP + AUCob_CP + Mean_Value_CP +
    Time_to_Peak_CP + AGE + BMI + EGPb + Gb + Gender + Ib, family = "binomial",
    data = ds_class_trn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39631  -0.04029   0.00000   0.02940   2.78181

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     12.9139080  4.3519307   2.967 0.003003 **
AUC_G            0.2672348  0.0521599   5.123 3.00e-07 ***
Max_Value_G      0.0468355  0.0355346   1.318 0.187495
Mean_Value_G    -2.0092690  0.4112531  -4.886 1.03e-06 ***
Min_Value_G      0.1716355  0.0617960   2.777 0.005479 **
Time_to_Peak_G   0.0507686  0.0149168   3.403 0.000665 ***
AUC_I           -0.0603451  0.0138043  -4.371 1.23e-05 ***
Max_Value_I     -0.0298979  0.0104289  -2.867 0.004146 **
Mean_Value_I     0.5245276  0.1093232   4.798 1.60e-06 ***
Min_Value_I     -0.2149555  0.0630977  -3.407 0.000658 ***
Time_to_Peak_I  -0.0496567  0.0118051  -4.206 2.60e-05 ***
AUC_CP          -0.0024105  0.0020383  -1.183 0.236964
AUCob_CP        -0.0007293  0.0005277  -1.382 0.166984
Mean_Value_CP    0.0133338  0.0113783   1.172 0.241251
Time_to_Peak_CP -0.0106611  0.0083755  -1.273 0.203054
AGE              0.0533522  0.0176729   3.019 0.002537 **
BMI             -0.4583129  0.0965602  -4.746 2.07e-06 ***
EGPb            -2.6577896  1.1591644  -2.293 0.021857 *
Gb               0.4418030  0.1024354   4.313 1.61e-05 ***
Gender          -3.2007237  0.7169636  -4.464 8.03e-06 ***
Ib               0.0114151  0.0617729   0.185 0.853392
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 122.12  on 579  degrees of freedom
AIC: 164.12
```

Figura 4.2: Regressione logistica su dataset completo successivamente all'applicazione dell'algoritmo Stepwise Backward Selection

Il modello addestrato mostrato in Figura 4.2, è stato poi utilizzato per svolgere delle predizioni sulla porzione di "test set" inizialmente ottenuta dallo split del dataset. I risultati ottenuti sono riassunti in Figura 4.3.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 94  7
         1  6 93

               Accuracy : 0.935
                 95% CI : (0.8914, 0.9649)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.87

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9300
            Specificity : 0.9400
```

Figura 4.3: Confusion matrix e indici delle performance del modello di regressione logistica in fase di predizione sulla porzione di dataset per il test

Come si può osservare dalla Figura 4.3, il modello sviluppato riesce a classificare i pazienti con un'accuracy del 93.5%, una sensitivity del 93% ed una specificity del 94%. Dalla Figura 4.2 possiamo inoltre estrarre informazioni riguardo la coerenza fisica del modello andando ad analizzare il comportamento delle singole features che compongono il modello. Prendendo ad esempio in considerazione la feature "Gb", ovvero il Glucosio Basale, osserviamo che essa presenta segno positivo, di conseguenza un aumento in valore di questa variabile comporterà un aumento della probabilità di un soggetto di essere classificato come patologico in fase avanzata. Dicesi lo stesso per valori di glucosio massimo e minimo (rispettivamente "Max_Value_G", "Min_Value_G") e per le altre variabili significative, segno del fatto che il modello è stato addestrato in maniera corretta e che potrebbe riuscire a classificare nuovi soggetti in maniera ottimale.

Classification Tree

Il secondo modello oggetto del seguente approfondimento è quello relativo all'albero di classificazione. Sebbene le performance di questo modello risultino essere leggermente inferiori rispetto a quelle degli altri 3 modelli presentati in questo paragrafo, questa tipologia di algoritmo andrebbe comunque tenuta in considerazione vista la semplicità di lettura e di interpretazione dei risultati, che non di rado vengono utilizzati in ambito clinico-decisionale. In Figura 4.4 è mostrato il primo modello ottenuto utilizzando il dataset completo.

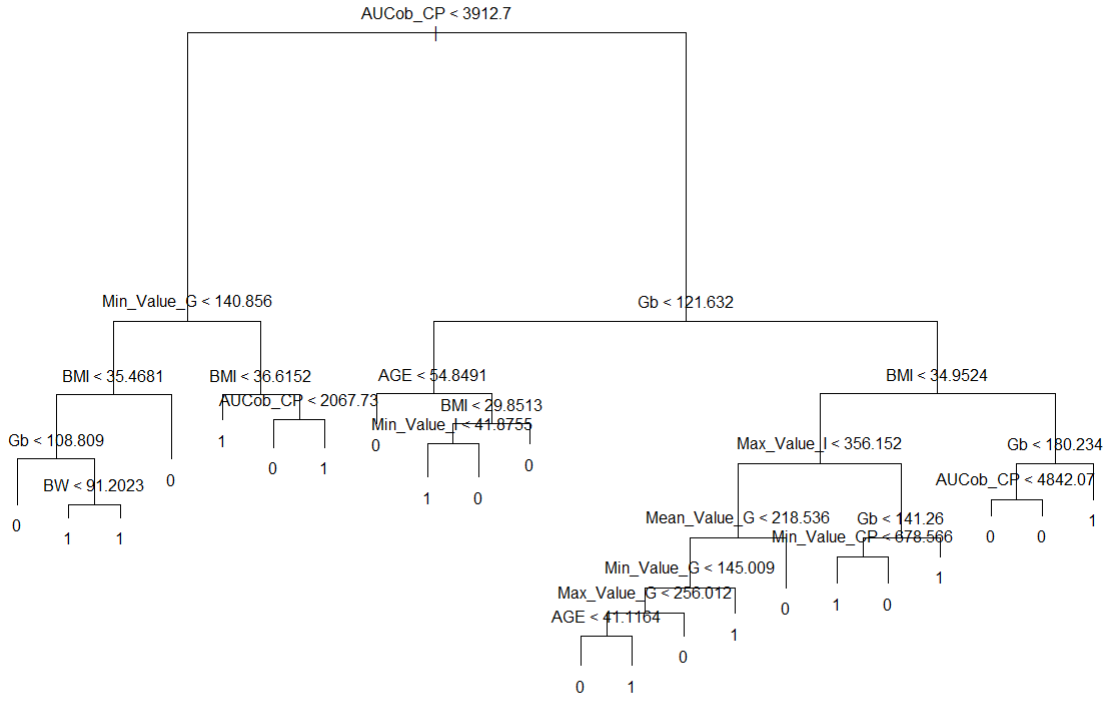

Figura 4.4: Classification Tree utilizzando il dataset completo

Il modello appena presentato, per costruzione, risulta essere difficilmente leggibile, si è dunque proceduto ad una operazione di "pruning" dell'albero, orientata alla riduzione delle dimensioni dell'albero, rimuovendone alcune sezioni, ed al miglioramento delle performance in termini di predizione evitando problemi di overfitting dei dati. La scelta del

parametro di pruning ottimale è stata eseguita tramite una k-fold Cross Validation, ed il risultato finale è mostrato in Figura 4.5.

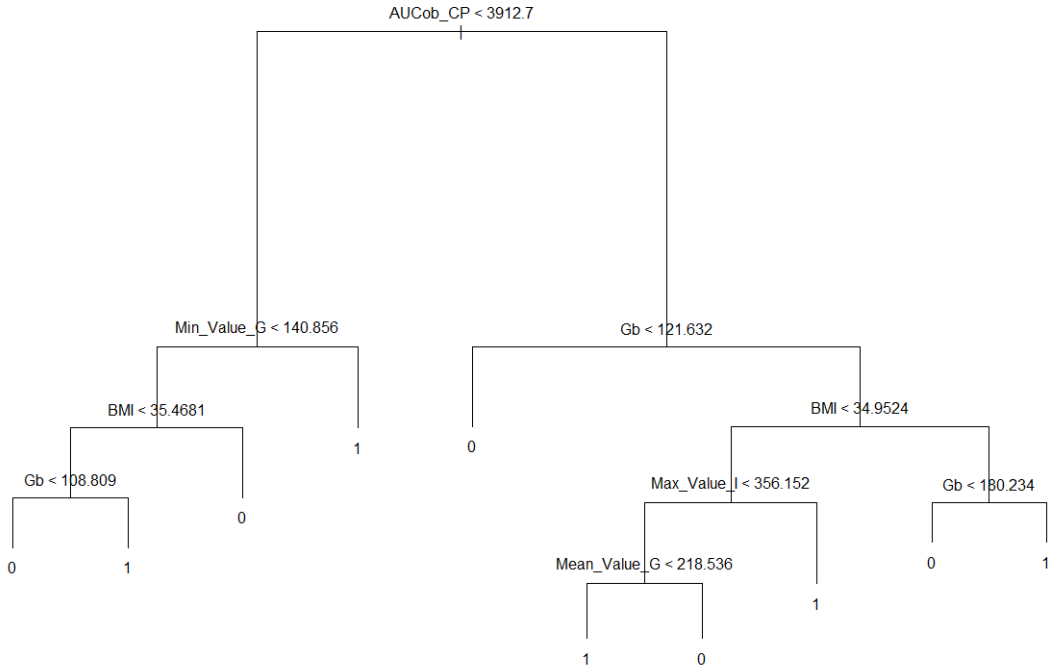

Figura 4.5: Classification Tree utilizzando il dataset completo dopo il pruning

Si riportano inoltre, in Figura 4.6, le performance del modello appena presentato, che ricordiamo avere un'accuracy del 78.5%, una sensitivity dell'87% ed una specificity del 70%, ne traiamo che il modello tende a classificare in maniera discreta i pazienti diabetici in fase avanzata, un po' meno accurata quelli in fase iniziale, ma in ogni caso non in maniera efficiente come gli altri modelli presentati.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 70 13
         1 30 87

               Accuracy : 0.785
                 95% CI : (0.7215, 0.8398)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.57

 Mcnemar's Test P-Value : 0.01469

            Sensitivity : 0.8700
            Specificity : 0.7000
```

Figura 4.6: Confusion matrix e indici delle performance dell'albero in fase di predizione sulla porzione di dataset per il test

Random Forest

Il terzo modello da tenere in considerazione è la Random Forest, notoriamente fra i modelli migliori in termine di performance fra quelli disponibili nell'ambito del machine learning. Una peculiarità di questa tipologia di modelli è la possibilità di ottenere in output una tipologia di grafico, chiamata "Importance Plot", che quantifica il peso (l'importanza) di ogni variabile nello sviluppo del modello stesso, fattore molto utile per una più dettagliata esplorazione e comprensione dei dati a nostra disposizione. In Figura 4.7 sono mostrati i due Importance Plot relativi all'algoritmo addestrato sulla porzione di training del dataset completo.

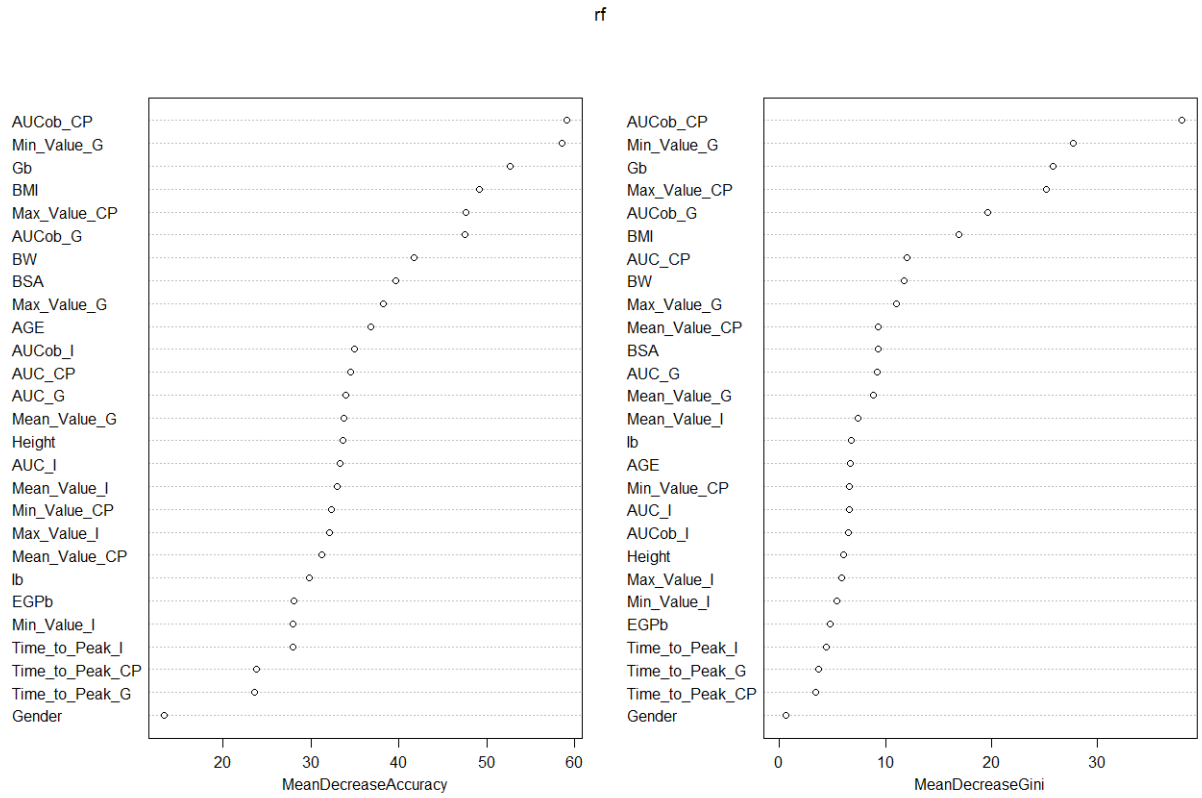

Figura 4.7: Importance Plot della random forest addestrata sul dataset completo

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 95  8
         1  5 92

               Accuracy : 0.935
                 95% CI : (0.8914, 0.9649)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.87

 Mcnemar's Test P-Value : 0.5791

            Sensitivity : 0.9200
            Specificity : 0.9500
```

Figura 4.8: Confusion matrix e indici delle performance della random forest in fase di predizione sulla porzione di dataset per il test

I grafici in Figura 4.7 mostrano in ordine decrescente di importanza, rispetto l'accuracy (grafico di sinistra) e rispetto l'indice Gini (grafico a destra), quali sono state le features che l'algoritmo ha tenuto in considerazione per la generazione dei 2000 alberi utilizzati per la costruzione della random forest. L'algoritmo così ottenuto è stato poi testato utilizzando il test set, ed i risultati sono riassunti in Figura 4.8, dalla quale si può notare che il modello sviluppato ha classificato i soggetti di test con una accuracy del 93.5%, una sensitivity del 92% ed una specificity del 95%.

Support Vector Machine

L'ultimo modello da tenere in considerazione è quello relativo alle Support Vector Machines. Fra i 4 presentati in questo paragrafo, risulta essere l'algoritmo con le migliori performance in termini di predizione. Il modello sviluppato, è risultato di un discreto numero di prove per l'identificazione del Kernel corretto da utilizzare. L'attenzione è ricaduta sul Kernel di tipo radiale, e gli iperparametri sono stati tarati tramite k-fold Cross Validation. Il modello ottenuto è stato successivamente testato sulla porzione di dataset completo adibita a test set, ed i risultati ottenuti sono riassunti in Figura 4.9.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 98  2
         1  3 97

               Accuracy : 0.975
                 95% CI : (0.9426, 0.9918)
    No Information Rate : 0.505
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.95

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9798
            Specificity : 0.9703
```

Figura 4.9: Confusion matrix e indici delle performance della support vector machine in fase di predizione sulla porzione di dataset per il test

Dalla Figura 4.9 si può osservare che il modello ha un'accuracy del 97.5%, una sensitivity del 97.98%, ed un valore di specificity pari al 97%. Trattasi di risultati estremamente positivi.

4.4.2 Approfondimento Classificazione - Dataset Ridotto

Preliminarmente alla descrizione dei modelli sviluppati e dei risultati ottenuti in questo caso, è bene precisare che sul dataset ridotto sono state eseguite 6 diverse tranche di analisi, che ricordiamo essere:

1. Test dei modelli usando una porzione del dataset ridotto contenente solo i seguenti parametri:
   - Age
   - BMI
   - BW
   - Gb
   - Gender
   - Height
   - Ib
2. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 30 minuti durante il test OGTT simulato
3. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 60 minuti
4. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 90 minuti
5. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 120 minuti
6. Test dei modelli usando la porzione di dataset ridotto contenente solo i parametri fisiologici, ed in aggiunta i valori di insulina e glucosio rilevati a 60 e 120 minuti

Avendo osservato nel paragrafo 4.1.3 nessuna sostanziale differenza in termini di performance dei modelli fra i 6 casi precedentemente elencati, per brevità nella descrizione dei risultati si riportano di seguito solo quelli relativi al test n°1 (usando una porzione del dataset ridotto contenente solo i parametri fisiologici scelti). La scelta di suddetto test rispetto agli altri è inoltre motivata dal fatto che i modelli sviluppati porterebbero consentire la diagnosi della patologia, in fase iniziale o avanzata, mediante l'utilizzo dei soli parametri fisiologici e di un singolo prelievo di sangue per la misura dei valori di Glucosio ed Insulina basale, senza di fatto dover ricorrere all'intero test OGTT, comunemente eseguito in questi casi, riducendo di fatto l'invasività della metodologia ed i relativi costi. Si riportano per brevità solo i modelli con le performance migliori, che risultano essere in questo caso, Random Forest e SVM.

Random Forest - Solo parametri Fisiologici

Come già discusso, la random forest addestrata utilizzando una porzione del dataset ridotto che contenesse solo i parametri fisiologici, ha garantito l'ottenimento di risultati ottimali in termini di performance, che sono riassunti in Figura 4.12.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 97  2
         1  3 98

               Accuracy : 0.975
                 95% CI : (0.9426, 0.9918)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.95

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9800
            Specificity : 0.9700
```

Figura 4.10: Confusion matrix e indici delle performance della random forest in fase di predizione sulla porzione di dataset ridotto, per il test

Dalla Figura 4.12 si può osservare che la random forest ha un'accuracy del 97.5%, una sensitivity del 98% ed una specificity del 97%, il modello di fatto ha classificato in maniera scorretta solo 5 soggetti su 200 testati. Si riportano inoltre gli Importance Plot relativi al modello sviluppato, in Figura 4.11, dalla quale possiamo osservare che il valore di glucosio basale (Gb) sembra avere un peso maggiore rispetto alle altre features, che presentano una distribuzione più omogenea.

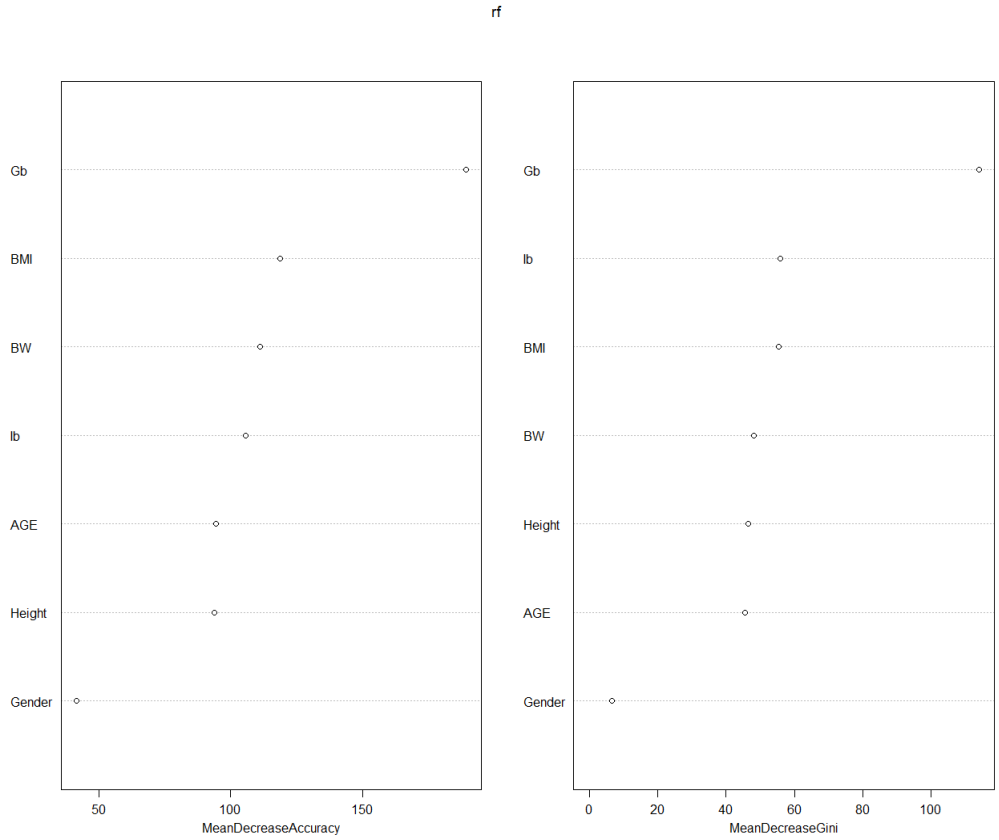

Figura 4.11: Importance Plot della random forest addestrata sul dataset ridotto

Support Vector Machine - Solo parametri Fisiologici

L'ultimo modello presentato e la SVM che, come nel caso mostrato per il dataset completo (Paragrafo 4.4.1), è stata addestrata utilizzando un Kernel di tipo radiale, che ci ha permesso di ottenere gli ottimi risultati mostrati in Figura 4.12.

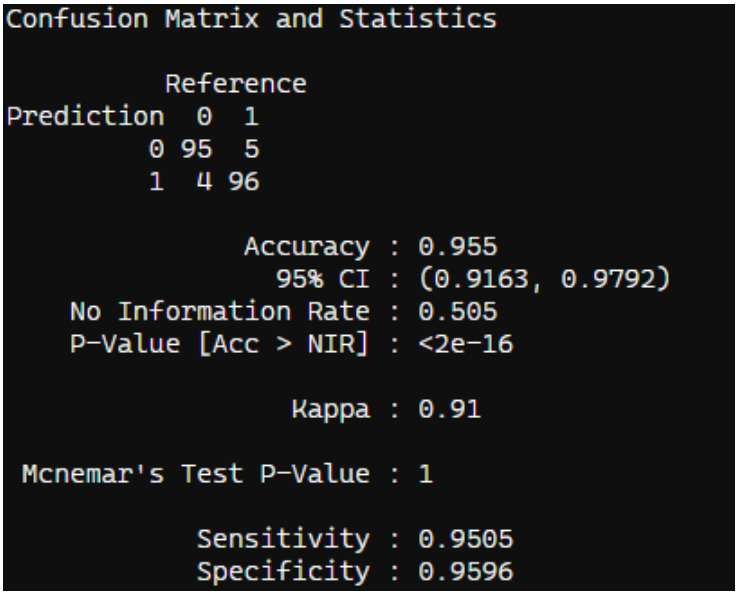

Figura 4.12: Confusion matrix e indici delle performance della support vector machine in fase di predizione sulla porzione di dataset per il test

Il modello è riuscito a classificare correttamente 191 dei 200 soggetti di test, ottenendo quindi una accuracy del 95.5%, una sensitivity del 95.1% ed una specificity del 95.96%.

4.4.3 Approfondimento Classificazione - Classificatore Multiplo

Si è deciso di concludere lo studio riguardo la classificazione andando a sviluppare un classificatore multiplo, aggiungendo ai dataset già ampiamente discussi una nuova coorte di soggetti virtuali sani. Si è dunque creato un nuovo dataset contenente le caratteristiche fisiologiche di 100 soggetti affetti da diabete di tipo 2 in fase iniziale, 100 soggetti affetti da diabete di tipo 2 in fase avanzata e 100 soggetti non affetti da diabete di tipo 2. Dal momento che i modelli si sono comportati in maniera ottimale anche nel caso del dataset ridotto (che ricordiamo avere pochi parametri fisiologici facilmente misurabili), anche in questo caso si sono tenute in considerazione le stesse features per la creazione del dataset multiclasse, nello specifico:

- **G_30min**: Valore di glucosio nel sangue a 30 minuti $[mg/dL]$
- **I_30min**: Valore di insulina nel sangue a 30 minuti $[mg/dL]$
- **G_60min**: Valore di glucosio nel sangue a 60 minuti $[mg/dL]$
- **I_60min**: Valore di insulina nel sangue a 60 minuti $[mg/dL]$
- **G_90min**: Valore di glucosio nel sangue a 90 minuti $[mg/dL]$
- **I_90min**: Valore di insulina nel sangue a 90 minuti $[mg/dL]$
- **G_120min**: Valore di glucosio nel sangue a 120 minuti $[mg/dL]$
- **I_120min**: Valore di insulina nel sangue a 120 minuti $[mg/dL]$
- **AGE**: Età del paziente $[anni]$
- **BMI**: Indice di massa corporea del paziente $[Kg/m^2]$
- **BW**: Peso del paziente [Kg]
- **BSA**: Area di superificie corporea $[m^2]$
- **Gb**: Glucosio basale $[mg/dL]$
- **Gender**: Sesso $[M/F]$
- **Height**: Altezza $[cm]$
- **Ib**: Insulina basale $[pmol/L]$
- **Treat.Plac**: Variabile discreta utilizzata per tenere traccia del trattamento scelto $[Meformina/Placebo]$

Il suddetto dataset è stato poi scremato fino all'ottenimento di un dataset contenente solo i parametri fisiologici fondamentali, e le misure di Insulina e Glucosio basale (lo stesso utilizzato per la raccolta dei dati riassunti in Tabella 4.3 del Paragrafo 4.1.3). Lo scopo di questa appendice è verificare se i modelli testati riescono a distinguere correttamente i soggetti affetti da diabete dai soggetti sani, partendo da parametri facilmente misurabili e rendendo di conseguenza una eventuale diagnosi quanto meno invasiva possibile. A tal proposito, basandoci sulle performance dei modelli dei casi esposti precedentemente, si è scelto di lavorare con i 3 algoritmi che mediamente hanno restituito i risultati migliori, nello specifico: Classification Trees, Random Forests e SVM. Si riassumono di seguito le performance ottenute.

Classification Tree - Classificatore Multiclasse

	Early Phase	Advanced Phase	Healthy Subjects
Sensitivity	0.704	0.905	1.000
Specificity	0.958	0.870	0.979
Balanced Accuracy	0.831	0.888	0.990
Overall Accuracy		0.867	

Tabella 4.38: Risultati del Classification Tree utilizzato come classificatore multiclasse, per la classificazione fra soggetti affetti da diabete di tipo 2 in fase iniziale, soggetti in fase avanzata e soggetti sani

Random Forest - Classificatore Multiclasse

	Early Phase	Advanced Phase	Healthy Subjects
Sensitivity	0.852	0.905	1.000
Specificity	0.958	0.926	1.000
Balanced Accuracy	0.905	0.915	1.000
Overall Accuracy		0.920	

Tabella 4.39: Risultati della Random Forest utilizzata come classificatore multiclasse, per la classificazione fra soggetti affetti da diabete di tipo 2 in fase iniziale, soggetti in fase avanzata e soggetti sani

Support Vector Machine - Classificatore Multiclasse

	Early Phase	Advanced Phase	Healthy Subjects
Sensitivity	0.778	0.773	0.962
Specificity	0.875	0.925	0.959
Balanced Accuracy	0.826	0.849	0.960
Overall Accuracy		0.840	

Tabella 4.40: Risultati della SVM utilizzata come classificatore multiclasse, per la classificazione fra soggetti affetti da diabete di tipo 2 in fase iniziale, soggetti in fase avanzata e soggetti sani

Dai risultati riassunti nelle Tabelle 4.38, 4.39 e 4.40, si evince che tutti i modelli riescono a distinguere accuratamente i soggetti sani dai soggetti diabetici. Performano meno meglio nella distinzione fra soggetti in fase iniziale e soggetti in fase avanzata. In termini di accuracy media, la Random Forest risulta essere il modello migliore fra quelli testati, con un valore del 92%, come si può osservare in Tabella 4.39.

Capitolo 5

Conclusioni

Il diabete mellito di tipo 2 è una malattia metabolica caratterizzata da glicemia alta in un contesto di insulino-resistenza ed insulino-deficienza relativa. Rappresenta circa il 90% dei casi di diabete nel mondo e lo sviluppo della patologia è causato da una combinazione tra lo stile di vita, endocrinopatie, e fattori genetici. L'Organizzazione Mondiale della Sanità riconosce la condizione di diabete dopo una rilevazione di elevati valori di glucosio nel sangue con la presenza di sintomi tipici, ed in questo contesto risulta di fondamentale importanza la somministrazione di un test chiamato "Oral Glucose Tolerance Test", comunemente OGTT, tramite il quale è possibile avere una diagnosi precisa della patologia. Grazie allo sviluppo e all'utilizzo di un software, il "T2D Simulator", di un gruppo di ricerca dell'Università di Padova, che fra le tante cose ci ha permesso di simulare il suddetto test su soggetti virtuali, è stato possibile ricostruire delle vere e proprie condizioni sperimentali, che hanno consentito l'ottenimento di un certo numero di datasets successivamente analizzati ed utilizzati per l'addestramento di specifici algoritmi di Machine Learning. L'applicazione degli algoritmi di apprendimento automatico risulta motivata in primis dalla costante esigenza di approfondire le nostre conoscenze riguardo le patologie in generale, fattibile grazie alla possibilità di esplorare grandi quantità dati tramite l'utilizzo di strumenti ad hoc, ed in secondo luogo per lo sviluppo di strumenti informatici potenzialmente utilizzabili a supporto di attività cliniche, che come nel caso del test OGTT, risultano essere leggermente invasive per i pazienti a causa dell'elevato numero di prelievi ematici, ed economicamente dispendiose per le aziende ospedaliere che le conducono. Sono dunque state condotte preliminarmente delle analisi di regressione sui dataset accuratamente organizzati, con l'obiettivo di sviluppare dei modelli capaci di poter stimare in maniera accurata specifici parametri funzionali per le cinetiche di secrezione ed assorbimento di glucosio ed insulina. I modelli testati sono di comune utilizzo nell'ambito delle analisi regressive se si parla di supervised learning, ed in particolare sono: Linear Regression, Regression Tree, Random Forest e Boosting. Uno step successivo ha riguardato lo sviluppo di modelli per la classificazione di soggetti diabetici, preliminarmente in relazione all'assunzione o meno di Metformina, e solo successivamente in base allo stadio di avanzamento della patologia. Si menzionano a tal proposito la Logistic Regression, il k-NN Classifier, il Naive-Bayes Classifier, i Classification Tree, le Random Forest e le Support Vector Machine. Da precisare che per tutti gli algoritmi elencati si è seguita la tipica pipeline di operazioni di preprocessing dei dataset e di taratura degli iperparametri tramite k-Fold Cross Validation, caratteristica degli studi nell'ambito del Machine Learning. Per quanto riguarda gli algoritmi di regressione del primo step, e la classificazione di pazienti trattati o meno con Metformina, i risultati ottenuti non ci permettono di estrapolare nuove informazioni o di sviluppare modelli utilizzabili per fare predizioni e/o classificazioni in ambito clinico. Viceversa, per quanto riguarda gli algoritmi sviluppati

per la classificazione di soggetti diabetici in base all'avanzamento della patologia, si sono ottenuti risultati eccellenti, che suggeriscono la possibilità di poter sviluppare in futuro dei metodi alternativi al test OGTT per la diagnosi di diabete e la catalogazione dello stadio di avanzamento della patologia. In termini di performance, i modelli che hanno restituito mediamente i risultati migliori sono: Logistic Regression, Random Forest ed SVM con Kernel Radiale. Suddetti modelli hanno restituito in output valori di accuracy, in alcuni casi, ben superiori al 95%, ma in generale sempre al di sopra del 90%, segno che i modelli riescono a classificare in maniera decisamente accurata eventuali nuovi soggetti. Sebbene, come già detto, i risultati siano promettenti, è doveroso ricordare che questo elaborato di tesi prende in esame si, caratteristiche coerenti con la fisiologia di soggetti affetti da diabete di tipo 2, ma pur sempre di soggetti virtuali estratti dal T2D Simulator. Non è dunque da escludere la possibilità che nell'applicazione degli stessi algoritmi su coorti di pazienti reali, si possano ottenere risultati diversi, in particolare inferiori in termini di performance. Si consiglia, dunque, come sviluppo futuro, proprio la ricerca su soggetti reali, possibilmente in numero più elevato rispetto agli 800 trattati nel presente elaborato di tesi, relativamente pochi per un efficace addestramento di algoritmi come quelli presentati.

Ringraziamenti

Il primo ringraziamento va al Professore Morten Gram Pedersen, che grazie alla sua professionalità e alla sua costante disponibilità e presenza, mi ha guidato passo dopo passo e mi ha permesso di condurre questo studio e di arrivare a conclusione di questo bellissimo percorso con un bagaglio di conoscenze che sono sicuro non sarei riuscito ad ottenere altrove.

Il secondo ringraziamento è da dedicare ai Professori R.Visentin e C.Della Man, ideatori del T2D Simulator e delle coorti di soggetti virtuali, senza i quali tutto questo non sarebbe potuto esistere.

Un enorme grazie va alla mia famiglia, tutta, ma in particolare a mio Padre, esempio di resilienza e caparbietà, a mia Madre fonte di bontà e di saggezza, e a mia Sorella vicina sempre, seppur adesso lontana. Ai miei zii, ai miei cugini, ai miei nonni, fonte costante di insegnamenti e crescita.

L'ennesimo grazie, e non sarà mai abbastanza, va al mio carissimo amico Rino, senza il quale io non sarei dove sono adesso, non solo per la sua ospitalità patavina, ma per essere stato una guida costante in questo percorso frastagliato che grazie a lui ho intrapreso anni fa. Auguro a tutti di avere un Rino nella propria vita, anche a Rino stesso.

Un grazie particolare va ad Edoardo, fratello più che amico, che dopo 20 anni mi dimostra ancora che è meglio averne pochi, ma sceglierseli bene.

Ringrazio poi la Residenza Cornaro, che è riuscita a regalarmi i 2 anni più belli che potessi immaginare, e mi ha dato la possibilità di conoscere persone eccezionali che occuperanno sempre un posto speciale nel mio cuore, prime fra tutte Fratm Lorenzo e Nonna Ani, grazie di tutto.

Ultima ma non per importanza, grazie Padova, non è un addio, ma un arrivederci.

Bibliografia

[1] Graziella Bruno, Franco Merletti, Antonio Vuolo, Elisabetta Pisu, Mauro Giorio, and Gianfranco Pagano. Sex differences in incidence of iddm in age-group 15-29 yr: Higher risk in males in province of turin, italy. *Diabetes Care*, 16(1):133–136, 1993.

[2] Kenneth P Burnham. Model selection and multimodel inference. *A practical information-theoretic approach*, 1998.

[3] Xiaowei Chen, Samarjit Kar, and Dan A Ralescu. Cross-entropy measure of uncertain variables. *Information Sciences*, 201:53–60, 2012.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[5] Gardner David, D Gardner, and R Dolores. Greenspan's basic & clinical endocrinology. *New York: McGram Hill Medical*, 2011.

[6] Andrea De Mauro. *Big Data Analytics: Analizzare e interpretare dati con il machine learning*. Apogeo Editore, 2019.

[7] James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[8] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[9] James Joyce. Bayes' theorem. 2003.

[10] Wolfgang Kerner and J Brückel. Definition, classification and diagnosis of diabetes mellitus. *Experimental and clinical endocrinology & diabetes*, 122(07):384–386, 2014.

[11] Stan Lipovetsky. Numerical recipes: The art of scientific computing. *Technometrics*, 51(4):481, 2009.

[12] Kenneth Maiese. Diabetic stress: new triumphs and challenges to maintain vascular longevity. *Expert Review of Cardiovascular Therapy*, 6(3):281–284, 2008.

[13] E Mannucci, R Candido, L Delle Monache, M Gallo, A Giaccari, ML Masini, A Mazzone, G Medea, B Pintaudi, G Targher, et al. La terapia del diabete mellito di tipo 2. linea guida della società italiana di diabetologia (sid) e dell'associazione medici diabetologi (amd). metodologia e sintesi. *J. AMD*, 24(3):232–240, 2021.

[14] Payal H Marathe, Helen X Gao, and Kelly L Close. American d iabetes a ssociation s tandards of m edical c are in d iabetes 2017, 2017.

[15] Hugh O McDevitt and Emil R Unanue. Autoimmune diabetes mellitus—much progress, but many challenges. *Advances in immunology*, 100:1–12, 2008.

[16] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285, 2004.

[17] World Health Organization et al. Definition, diagnosis and classification of diabetes mellitus and its complications: report of a who consultation. part 1, diagnosis and classification of diabetes mellitus. Technical report, World health organization, 1999.

[18] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[19] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[20] James H Stock and Mark W Watson. *Introduzione all'econometria*. Pearson Italia Spa, 2005.

[21] James H Stock and Mark W Watson. Introduction to econometrics (3rd updated edition). *Age (X3)*, 3(0.22), 2015.

[22] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

[23] Stéphane Tufféry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011.

[24] Sandeep Vijan. Type 2 diabetes. *Annals of internal medicine*, 152(5):ITC3–1, 2010.

[25] Roberto Visentin, Claudio Cobelli, and Chiara Dalla Man. The padova type 2 diabetes simulator from triple-tracer single-meal studies: In silico trials also possible in rare but not-so-rare individuals. *Diabetes technology & therapeutics*, 22(12):892–903, 2020.

[26] Roberto Visentin, Claudio Cobelli, and Chiara Dalla Man. A software interface for in silico testing of type 2 diabetes treatments. *Computer Methods and Programs in Biomedicine*, 223:106973, 2022.

[27] Wikipedia. Diabete mellito di tipo 1 — wikipedia, l'enciclopedia libera, 2022. [Online; in data 6-aprile-2023].

[28] Wikipedia. Diabete mellito di tipo 2 — wikipedia, l'enciclopedia libera, 2023. [Online; in data 6-aprile-2023].

[29] Wikipedia. Test orale di tolleranza al glucosio — wikipedia, l'enciclopedia libera, 2023. [Online; in data 6-aprile-2023].