Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in
Bioingegneria

APPROCCI ALLA STRATIFICAZIONE NON SUPERVISIONATA DI PAZIENTI CON DIABETE DI TIPO I USANDO SOLO DATI DI MONITORAGGIO IN CONTINUA DELLA GLICEMIA

Relatore:
Prof. Giovanni Sparacino

Laureanda:
Cristina Andolfatto
Matricola: 1211247

Correlatori:
Dott. Giacomo Cappon
Ing. Giulia Noaro

Anno Accademico 2021/2022
Data di laurea 21/02/2022

Ringraziamenti

Ringrazio il Professor Sparacino, il Dottor Giacomo Cappon, la Dottoressa Giulia Noaro, e tutto il team di ricerca per la sempre gentile disponibilità e presenza, il continuo supporto e la fiducia riposta nei miei confronti.

Ringrazio la mia famiglia, per avermi sempre supportato (ma soprattutto sopportato), non sarei quella che sono oggi senza di voi.

Ringrazio te, Seba, per avermi salvato, e amato sempre, incondizionatamente; per essere la mia spalla ed il mio primo supporter in ogni situazione, per spronarmi ad essere ogni giorno migliore.
Sei la cosa più preziosa che ho.

E sì, finalmente ringrazio anche te, Cri, per non aver mai mollato.

Abstract

Il Diabete Mellito è senza dubbio una tra le patologie che più caratterizzano il nostro secolo, e l'andamento della sua incidenza presenta un trend allarmante: secondo la World Health Organization (WHO), i casi di diabete diagnosticati a livello globale nel 1980 erano 180 milioni [1], contro i 422 milioni registrati nel 2014 ed i 600 milioni previsti entro il 2035. La situazione pandemica che ha colpito la nostra società, inoltre, ha ulteriormente aggravato la situazione, già precaria, di soggetti con malattie croniche come il diabete: come riportato in [4], il diabete risulta essere tra i principali fattori di rischio sia per una possibile infezione da Sars-Cov-2, sia per una più grave progressione della malattia. La gravità e l'impatto rilevante che ha però questa patologia sui pazienti sono, in parte, già note da diversi anni. Il diabete fa parte delle cosiddette "malattie metaboliche" e si configura principalmente in un mancato controllo dei livelli di glicemia nel sangue a causa di una secrezione deficitaria (presente nel diabete di tipo II) o nulla (caratteristica invece del tipo I) di insulina da parte del pancreas [5]. Questa disfunzione fa sì che la glicemia di soggetti diabetici possa ritrovarsi a livelli che non risultano essere fisiologici, in particolare più o meno elevati (rispettivamente range iperglicemico e ipoglicemico) rispetto al "range euglicemico"; il prolungarsi di questa condizione può comportare gravi complicanze sistemiche. Risulta di fondamentale importanza quindi il monitoraggio dei livelli glicemici del paziente, che ad oggi può avvenire attraverso l'utilizzo di sensori "CGM" ("Continuous Glucose Monitoring" system) [6], che permettono al paziente di avere informazioni (quasi) continue sui livelli di glicemia, ma anche sulla sua variabilità.
La disponibilità di una sempre più elevata quantità di dati di glicemia, derivanti dall'utilizzo di tali sensori, ha fatto emergere l'idea di sviluppare anche per il diabete una "medicina di precisione", che si propone di sviluppare terapie ottimizzate sul paziente e sul suo stato di salute [7]. In particolare, attraverso metodologie di "clustering", si possono pensare di individuare sottogruppi di pazienti con caratteristiche simili, rendendo queste quindi uno strumento promettente per la medicina nell'era dei "big data" al fine di individuare o sviluppare terapie sempre più personalizzate. Questo è proprio ciò che si propone di fare il seguente lavoro di tesi. L'obbiettivo è quello di applicare le tecniche di clustering a lunghe serie di dati CGM, al fine, in primis, di individuare sottogruppi di soggetti o pattern glicemici simili, e successivamente di indagare su possibili applicazioni future: per fare ciò, sono stati utilizzati due algoritmi appartenenti allo stato dell'arte (k-means e clustering gerarchico) su un database di dati CGM

raccolti con il sistema di pancreas artificiale open-source OpenAPS. Sono stati ottenuti un totale di 6 cluster per i profili dei pazienti e 8 cluster per i profili glicemici settimanali, analizzati in ultima battuta per indagare sul loro utilizzo nello sviluppo di possibili strumenti di ausilio alla terapia dei pazienti.
La tesi è strutturata in 7 capitoli. Nel capitolo 1 vengono introdotte la patologia del diabete, le principali tecniche di monitoraggio glicemico ed infine le nuove tendenze in ambito di studio e controllo della malattia. Nel capitolo 2 vengono presentati dataset e fase di "preprocessing". Nel capitolo 3 si riporta la descrizione della fase di estrazione delle features. Nel capitolo 4 invece vengono descritte le tecniche di clustering utilizzate e le loro caratteristiche. Nel capitolo 5 vengono presentati i risultati sulle due analisi condotte in parallelo su pazienti e pattern glicemici settimanali. Nel capitolo 6 vengono esplorate possibili applicazioni dei risultati ottenuti, mentre, infine, nel capitolo 7 sono riportati brevemente possibili sviluppi futuri del progetto presentato, assieme alle conclusioni sul lavoro svolto.

Abstract

English version: Diabetes is, without any doubt, one of the most characteristic pathology of our time and its trend it's far from being negative: according to the World Health Organization (WHO), global cases of diabetes in 1980 were 180 millions [1], against the 422 millions registered in 2014 and the 600 millions estimated for the 2035. Also, the pandemic situation that affects our society globally has additionally exacerbated the already complicated situation of people with diabetes: in fact, as reported in [4], diabetes it's one of the risk factors of infection and severe disease progression of Covid-19. The severe and huge impact that diabetes has on patients it's however known from several years. Diabetes is a "metabolic disease" and can be characterized as a lack of control of glucose levels in blood caused by a deficient or absent secretion of insulin by pancreas [5]. As a result of this disfunction, glucose levels of diabetic patients can be in ranges that are higher (hyperglicemic) or lower (hypoglycemic) than the basal and physiological one (euglycemic range); if this situation lasts, the risk of severe comorbidities dramatically grows. So, it's fundamental to monitor patients glucose levels with sensors such as the CGM one ("Continuous Glucose Monitoring sensor"), that lead the patient to monitor its blood glucose levels almost continuosly and also glycemic variability. The increased availability of lot of data given by the use of this sensors, caused the idea to develop also for diabetes treatment "precision medicine" approaches, that aims to develop patient-optimized and personalized

therapies [7]. In particular, through the use of "clustering" algorithms, it is possibile to find group of patients with similar characteristics: this lead clustering techniques to represent a promising medical instrument in the "big data" era, in order to find more individualized therapy strategies. This is also what these study aims to do: apply clustering techniques to long CGM data series in order to find possible subgroups of similar patients or weekly glycemic patterns, and then investigate to possible clinical uses and applications: in order to do so, two "state of the art" algorithm (k-means and hierarchical clustering) were used on a database made of CGM data coming from an open-source artificial pancreas system named OpenAPS. We obtained a total of 6 patients clusters and 8 weekly glycemic pattern clusters, subsequently analyzed in order to find new possible instruments to enhance patients therapy.

This thesis is structured in 7 chapters. Chapter 1 introduces the pathology and its management, with classic and new techniques. In the chapter 2 the dataset is presented, along with first preprocessing operations. The description of feature extraction phase is reported in chapter 3. Chapter 4 reports state of the art clustering methodologies. In chapter 5 one can find the results obtained with the application of clustering techniques on two types of datasets, the patients one and the weekly glycemic patterns set. In chapter 6 investigation on possible applications of clustering results are described, whereas, finally, chapter 7 outline possible future improvements and conclusions on the present work.

Indice

Capitolo 1

Nuove prospettive nella terapia del diabete aperte dai sensori di monitoraggio in continua della glicemia

In questo primo capitolo viene brevemente descritta la patologia del Diabete Mellito, con una rapida analisi circa la sua incidenza e di quelli che sono i sintomi della malattia; nello specifico, viene descritto come questa impatti sulla vita dei pazienti che ne sono affetti e ne vengono brevemente riportate le due tipologie più diffuse (tipo I e tipo II). Vengono poi descritte le principali tecniche di monitoraggio della glicemia appartenenti allo stato dell'arte, con un successivo sguardo a quelle che sono invece le nuove prospettive emergenti. In coda al capitolo, viene infine esposto l'obbiettivo del presente lavoro di tesi, e come questo cerchi di inserirsi proprio all'interno di queste nuove prospettive.

1.1 Il Diabete Mellito

1.1.1 Il problema globale del diabete

Il Diabete Mellito colpisce ogni anno milioni di persone: solo negli Stati Uniti infatti, e solamente nell'anno 2018, sono stati registrati 1.5 milioni di nuovi casi [2] . L'Italia non è da meno: il diabete infatti è stato inserito nell'annua-

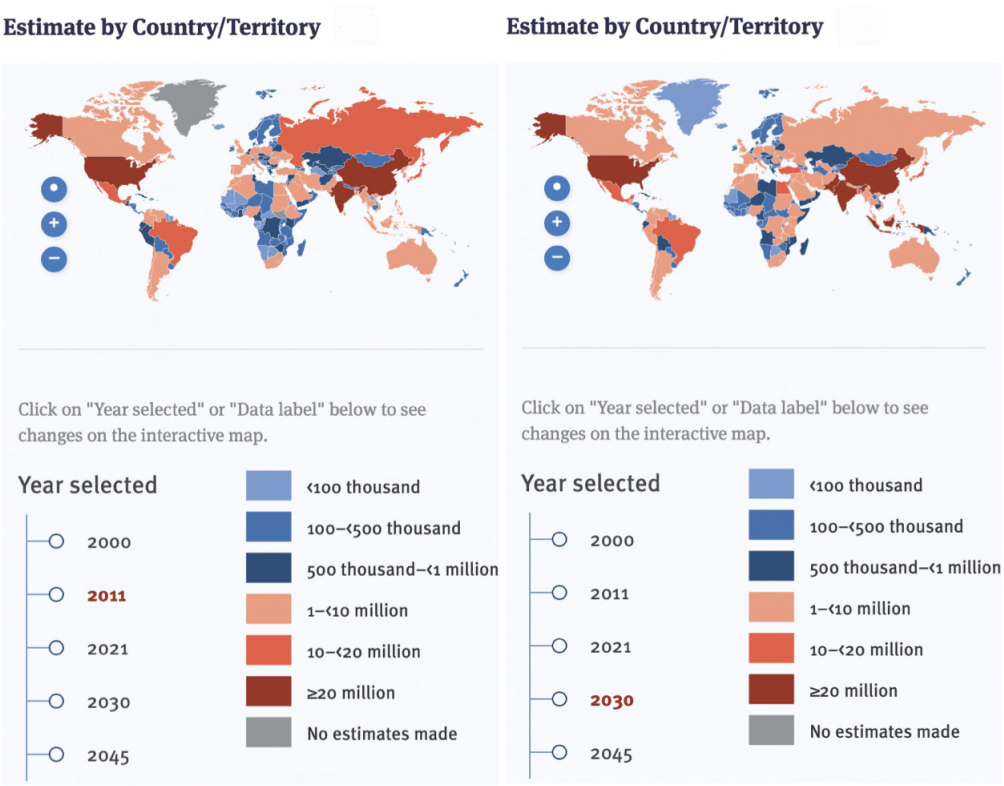


Figura 1.1: Stime casi di diabete a livello mondiale: anno 2011 (a sinistra) e crescita stimata per il 2030 (a destra) (fonte:[1.1] p.51)

rio statistico prodotto dall'Istituto Nazionale di Statistica (ISTAT) all'interno lista delle principali malattie croniche che affliggono il nostro paese, con una percentuale di incidenza, riferita all'anno 2016, del 5,3%, pari a oltre 3 milioni di persone. Quello del diabete è un problema però globalmente diffuso, tanto che la World Health Organization (WHO), assieme all'Organizzazione delle Nazioni Unite (ONU), hanno inserito questa patologia nelle lista delle emergenze sanitarie, assieme a malaria e tubercolosi [8]: le stime globali, infatti, riportano che gli individui affetti da questa patologia sono vicini alla soglia dei 400 milioni, e si prevede che possano raggiungere i 600 milioni entro il 2035; l'entità e la gravità di questo trend positivo si possono facilmente desumere dalla figura 1.1.

I numeri precedentemente riportati però, seppur non trascurabili, possono riferirsi solamente ai casi diagnosticati e non tengono conto di tutti quegli individui nei quali la patologia rimane latente negli anni (e quindi non identificata): i dati quindi potrebbero essere ben peggiori di quelli riportati. La sua incidenza va via via incrementando e caratterizza non solo i paesi del mondo maggiormente sviluppati (Europa, Asia, Nord America), ma anche quelli emergenti o in via di sviluppo;

anzi, proprio in questi ultimi la sua crescita risulta essere maggiore. Quello del diabete non è però soltanto un problema sanitario ma anche socio-economico: rimanendo in Italia, la quota di spesa che il Fondo Sanitario Nazionale destina alla cura del diabete è pari al 10% del totale (circa 15 miliardi di euro l'anno); a questa ingente quantità di denaro pubblico, si deve poi aggiungere gli stimati 3 miliardi di spese dirette delle persone affette da diabete e le famiglie loro vicine. Mai come nell'ultimo periodo infine ci si è ricordati dell'esigenza di proteggere i soggetti più fragili e i pazienti affetti da malattie croniche: la pandemia che ha colpito la nostra società infatti, come ben noto, ha ulteriormente gravato la condizione, già precaria, di queste categorie; tanto che, come riportato in [4], dalle prime analisi sembra emergere che il diabete sia tra i principali fattori di rischio sia per l'infezione del virus Sars-Cov-2, sia per una più grave progressione della malattia. Il diabete, quindi, senza ombra di dubbio rappresenta uno dei più grandi problemi del nostro secolo, e al contempo una delle più grandi sfide che la medicina e l'ingegneria si ritrovano, oggi come ieri, a dover affrontare. Per riuscire in questo, fondamentale è la conoscenza della patologia, delle sue caratteristiche, della sua eziologia e dei sintomi. Questi aspetti nello specifico verrano brevemente descritti nei paragrafi a seguire.

1.1.2 Descrizione della malattia

Il Diabete Mellito è una malattia cronica facente parte delle patologie endocrine e metaboliche. Si configura principalmente in una secrezione mancata o deficitaria di insulina da parte del pancreas, che porta ad un errato controllo dei livelli di glucosio nel sangue. Per comprendere meglio però l'intimo rapporto che lega queste due molecole, schematizzato in figura 1.2, è necessario prima un breve excursus sul metabolismo di entrambe.
Il glucosio è una molecola fondamentale per il funzionamento dell'organismo umano (ma in generale anche per quello di molti altri esseri viventi) e alcuni dei nostri organi vengono solitamente definiti infatti "glucosio-dipendenti" [9]. Questi, quali ad esempio quelli appartenenti al sistema nervoso, ma in generale anche molte cellule del corpo umano come i globuli rossi, necessitano proprio di substrati di glucosio per adempiere alle loro funzioni e per mantenersi in vita. Oltre a questi, la cui sopravvivenza dipende quindi da una presenza costante di glucosio, vi sono anche altre tipologie di organi o tessuti, come quello muscolare e adiposo, per i quali l'utilizzo di glucosio è fasico e dipendente dal livello di insulina circolante (e sono detti per questo "insulino-dipendenti"). L'insulina è un ormone secreto dalle

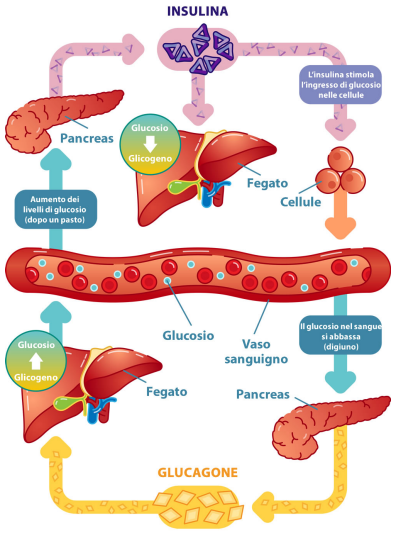


Figura 1.2: Schema meccanismo di controllo del glucosio grazie al pancreas) (fonte:[1.2] p.51)

β-cellule del pancreas endocrino, e, come tutti gli ormoni, viene rilasciato nel sangue allo scopo di regolare le funzioni biochimiche di cosiddetti "organi bersaglio"; nello specifico, l'insulina ha un effetto anabolico, in quanto predispone e favorisce l'assorbimento di nutrienti da parte dell'organismo e la sintesi di nuova materia prima; possiede di conseguenza anche un effetto ipoglicemizzante: stimola infatti l'utilizzazione di glucosio da parte degli organi insulino-dipendenti, abbassandone i livelli nel sangue a seguito di un suo repentino innalzamento (dovuto ad esempio ad un pasto). Se vi sono malfunzionamenti o addirittura assenza di produzione di insulina endogena, come nel caso di pazienti diabetici , questo equilibrio viene a mancare, causando il mantenimento dei livelli di glicemia del sangue, per periodi di tempo più o meno lunghi, all'interno di range non fisiologici: quello iperglicemico e ipoglicemico. Come stabilito in [10], il range iperglicemico in particolare è definito per valori di glucosio maggiori di 180 mg/dl: il prolungarsi del mantenimento dei livelli di glucosio all'interno di questo range non causa gravi problematiche nel breve ma piuttosto nel lungo periodo; infatti, le principali conseguenze di iperglicemia cronica sono [1]:

- Nefropatia: la "Diabetic Kidney Disease" è la complicanza più diffusa nei pazienti diabetici [12] e si manifesta con una costante secrezione elevata di albumina ("albuminuria") nelle urine e un basso tasso di filtrazione renale; queste condizioni compromettono progressivamente la salute renale, tanto che il diabete risulta come causa principale del 50% dei casi di malattia re-

nale allo stadio terminale ("End-Stage Renal Disease"), comportando trattamenti invasivi come dialisi o trapianto renale. La nefropatia contribuisce inoltre ad aumentare i rischi di complicanze cardiovascolari e ovviamente i costi dell'assistenza sanitaria;

- Demenza precoce e Neuropatia [13] : la Neuropatia in particolare è particolarmente critica in quanto ad oggi risulta difficile valutare le condizioni di salute dei nervi e può inoltre configurarsi in diverse sottocategorie, tutte con sintomi diversi tra loro, a volte anche generici e che spesso è difficile ricondurre clinicamente ad esse; esempi di queste sono la Neuropatia Periferale, Cardiaca o Gastrointestinale, che possono manifestare sintomi quali sensazione di dolore cronico, tachicardia a riposo, diarrea, disfunzioni dell'apparato sessuale;

- Retinopatia: presente sia nei pazienti con diabete di tipo I che di tipo II, la retinopatia diabetica risulta essere la principale causa di cecità nei pazienti adulti (20-74 anni) nei paesi sviluppati; questa patologia risulta essere inoltre particolarmente critica in quanto può rimanere asintomatica per diverso tempo e per questo richiede screening e controlli periodici frequenti;

- Malattia cardiovascolare arteriosclerotica ("AtheroSclerotic CardioVascular Disease"): è la causa di mortalità più diffusa tra i pazienti diabetici e comprende malattie coronariche ("Coronary Heart Disease") o arteriose; queste in particolare sono state inserite dall'OMS nella lista delle "Non Communicable Diseases", malattie che ogni anno causano il decesso di oltre 41 milioni di persone, l'equivalente del 71% delle morti globali; tra i sintomi principali compaiono l'ipertensione e, nei casi più gravi, l'arresto cardiaco;

- Complicanze microvascolari, ulcere e nei casi più gravi amputazione negli arti inferiori (piedi sopratutto);

A partire invece dall'estremo inferiore del range fisiologico dei livelli di glicemia (range "euglicemico"), troviamo invece il range denominato come "range ipoglicemico". Viene configurato quindi come quella fascia di valori inferiori ai 70 mg/dl e risulta essere particolarmente critico: se il glucosio infatti si attesta attorno a questi valori, può provocare gravi conseguenze anche nel breve periodo, come coma o addirittura il decesso del paziente.
Al fine di evitare le precedenti descritte complicanze, la terapia del diabete prevede il monitoraggio dei livelli di glucosio del sangue e conseguenti infusioni esogene

di insulina, al fine di ottenere un buon controllo della glicemia ed evitare eventi iper- o ipo-glicemici; l'inizio di uno di questi, nello specifico, è definito all'interno dell' "International Consensus on the Use of Continuous Glucose Monitoring" [10] come il mantenimento per almeno 15 minuti dei livelli di glucosio nella fascia ipoglicemica (evento di ipoglicemia) e iperglicemica (evento di iperglicemia).
Una gestione ottimale della glicemia richiede di ottenere preferibilmente quindi parametri come ridotta variabilità glicemica, alte percentuali di tempo speso nella fascia euglicemica (e di conseguenza basse percentuali nei range ipo-/iperglicemici), basso numero di eventi iper- e ipo-glicemici; come già ribadito però, al fine di ottenere ciò, è necessaria la conoscenza dei valori della glicemia in circolo e quindi l'implementazione di tecniche o dispositivi per il monitoraggio del glucosio nei pazienti diabetici; successivamente ad una breve descrizione dei due principali tipi di diabete (paragrafi 1.1.3-1.1.4) , verranno quindi riportare le principali tecniche utilizzate per il monitoraggio della glicemia (paragrafi 1.2.1-1.2.2).
Spesso non risulta semplice distinguere in maniera netta i diversi tipi di diabete e soprattutto le forme che colpiscono i pazienti, ma grazie agli studi condotti negli anni si è riusciti ad ottenere una classificazione di massima delle diverse sottocategorie, le cui principali, come anticipato, verranno riportate nei paragrafi a seguire.

1.1.3 Diabete di tipo I

Questa forma interessa solamente il 5-10% dei casi totali diagnosticati di diabete [5], e si manifesta nella maggior parte dei casi con una assenza di produzione di insulina da parte delle cellule del pancreas; questa assenza è dovuta ad una reazione autoimmune dell'organismo, che porta le cellule ad "autodistruggere" le β-cells del pancreas; il tasso di velocità con cui avviene questa distruzione è variabile e può essere rapido in alcuni individui (concentrandosi quindi nelle fasce d'età infantile) o più lento in altri (principalmente adulti). Alcuni dei primi sintomi della malattia includono chetoacidosi (insieme di iperglicemia, iperchetonemia e acidosi metabolica) e iperglicemia a digiuno. Le cause del diabete di tipo I non sono ancora del tutto comprovate e note, ma sembrano essere ricollegate a diverse predisposizioni genetiche e fattori ambientali che possono scatenare l'insorgenza della malattia, che spesso avviene in età adolescenziale e giovane (ma non ne è esclusa la comparsa anche in età più avanzata). Gli individui affetti da questa forma risultano anche essere maggiormente predisposti ad altre malattie autoimmuni, come la sindrome di Hashimoto, epatite autoimmune, anemia perniciosa.

Esiste infine una piccola percentuale di pazienti nei quali la secrezione di insulina non è del tutto assente, ma piuttosto deficitaria e soprattutto variabile nel tempo: questi individui vengono comunemente raggruppati come affetti da una sottocategoria di diabete di tipi I chiamata "diabete idiopatico"; nello specifico, soffrono di episodi di chetoacidosi ma non presentano evidenze di reazioni autoimmuni e risultano più frequenti nella popolazione di origine africana e asiatica.
Ad oggi non esiste una cura definitiva per il diabete di tipo I ma viene gestito tramite l'infusione esogena di insulina in dosi soggetto-dipendenti (stabilite attraverso visite specialistiche) chiamate "boli", che possono essere di due tipi: "meal bolus" (ovvero dosi di insulina conseguenti ad un pasto) e "basal insulin"; quest'ultima in particolare permette al paziente di mantenere, auspicabilmente per più tempo possibile, i livelli di glicemia all'interno del range normoglicemico lontano dai pasti o durante la notte (dove non sono poco frequenti episodi di ipoglicemie notturne particolarmente critici).

1.1.4 Diabete di tipo II

Questa tipologia di diabete è quella più largamente diffusa: il 95% dei pazienti diabetici presenta infatti questa forma, detta comunemente anche "adult-onset diabetes" o "non insulin-dependent diabetes" [5]; infatti, per lo meno inizialmente, i pazienti affetti da questa tipologia non necessitano obbligatoriamente di infusione di boli di insulina per la sopravvivenza: sono per lo più soggetti insulino-resistenti o deficitari, per i quali quindi la secrezione di insulina endogena è scarsa o il suo effetto è blando e non sufficiente per la normale regolazione glicemica. L'eziologia specifica del tipo II non è nota, ma in questo caso non è presente la reazione autoimmune di distruzione delle β-cells del pancreas come nel tipo I; anche il grado di chetoacidosi è inferiore e raramente si manifesta in maniera spontanea, ma piuttosto emerge in seguito alla comparsa di altre cause esterne come infezioni o altre patologie.
Nonostante non ci siano state individuate ad oggi delle cause specifiche, il diabete di tipo II sembra essere una malattia "multifattoriale" , con forte predisposizione genetica ma che trova correlazioni non trascurabili anche con altri aspetti, legati soprattutto al protrarsi nel tempo di stili di vita e abitudini poco salutari; l'obesità in primis (assieme anche a scarsa attività fisica e abitudini alimentari errate) sembra avere un forte legame con l'insorgere in età adulta non solo di questo tipo di diabete, ma anche in generale di una ridotta "sensibilità insulinica" dell'individuo (che se protratta nel tempo rischia appunto di diventare patologica): come

riportato in [14], un valore di BMI compreso tra 27.2 e 54.2 kg/m^2, è risultata essere associata ad un rischio di sviluppare diabete di tipo II 8 volte maggiori rispetto ad individui con BMI inferiore a 22.8 kg/m^2 ; inoltre, i pazienti affetti da questa tipologia, nonostante all'apparenza sembrino avere un controllo glicemico migliore rispetto ai pazienti del tipo I, sono quelli più esposti a complicanze micro e macro vascolari, oltre che allo sviluppo di patologie cardiovascolari. Il trattamento del diabete di tipo II prevede quindi, oltre al supporto farmacologico, anche una particolare attenzione all'attuazione di cambiamenti nello stile di vita, dieta ed esercizio fisico.

1.2 Tecniche di monitoraggio della glicemia

A seguito di questo breve excursus circa caratteristiche e classificazione generali del diabete mellito, viene riportata una rapida descrizione delle principali tecniche che consentono al paziente di monitorare i livelli di glicemia presenti in circolo e quindi la stima corretta di boli di insulinca ed il trattamento della patologia. Ciò infatti, come precedentemente descritto, consente auspicabilmente al paziente di mantenere la glicemia all'interno dell'intervallo euglicemico 70-180mg/dl, limitare escursioni,variabilità glicemica ed eventi di ipo- o iper-glicemia.

1.2.1 Self Monitoring Blood Glucose

La tecnica dell' "automonitoraggio" (appunto "Self Monitoring Blood Glucose") è quella largamente più diffusa e tradizionale: i primi dispositivi risalgono infatti ai primi anni 70 [15] e già allora davano prova di migliorare il controllo glicemico dei pazienti. Questa tecnica prevede l'utilizzo del cosiddetto "glucometro", un dispositivo medico portatile che, assieme ad una "penna pungidito" (utile per il prelievo di una piccola quantità di sangue) e delle strisce reattive, consente al paziente di poter monitorare i propri livelli di glucosio in maniera completamente autonoma. La figura 1.3 ne riporta un esempio. Il dispositivo quindi analizza la piccola goccia di sangue che risale per capillarità lungo la striscia reattiva; a questo punto, nella maggior parte dei dispositivi avviene una reazione chimica di ossidazione del glucosio, che comporta la variazione cromatica o la produzione di corrente elettrica in una misura proporzionale alla glicemia presente nel campione. L'accuratezza di questi dispositivi è buona, l'errore di misura limitato (5-10% CV) ed il loro utilizzo risulta semplice ed intuitivo; questa tecnica di monitoraggio però soffre di problemi di contaminazione delle misure [16] (non solo temperatura

Figura 1.3: Dispositivo utilizzato per misure SMBG (fonte:[1.3] p.51)

ed umidità dell'ambiente ma anche valore di ematocrito, quantità di colesterolo e trigliceridi o presenza di ipotensione); infine, un problema non trascurabile di questa tecnica è l'invasività nei confronti della quotidianità dei pazienti: le misure eseguite giornalmente infatti posso essere addirittura 7-8, ed in ogni caso la disponibilità così bassa e diluita nel tempo di informazioni circa i livelli di glicemia non sempre assicura un trattamento ottimale della patologia. Per risolvere questa problematica, entrano in gioco verso la fine degli anni 90 i primi sensori di monitoraggio in continua della glicemia, i cosiddetti sensori "CGM", brevemente descritti nel paragrafo successivo.

1.2.2 Il sensore CGM

Come anticipato nel paragrafo precedente, il sensore CGM ("Continuous Glucose Monitoring") appare nel mercato verso la fine degli anni 90 e si ripropone di migliorare il monitoraggio della glicemia con una rilevazione pressochè continua (le frequenze di campionamento vanno da 1 a 10 minuti [6]) e minimamente invasiva, visto che non richiedono alcun prelievo da parte del paziente; i sensori CGM in particolare sfruttano anch'essi una reazione di ossidazione del glucosio e ne misurano la concentrazione sfruttando la corrente che si genera dalla reazione. Se da un lato quindi consentono di avere un numero di informazioni molto maggiori rispetto ai glucometri, dall'altro soffrono di un grosso problema di accuratezza, dovuto al fatto che la corrente sfruttata per misurare la glicemia non è proporzionale alla quantità di glucosio nel sangue (come appunto nel caso delle misure SMBG) bensì ai livelli di glucosio nel tessuto interstiziale; ciò comporta un ritar-

Figura 1.4: Dispositivo Continuous Glucose Monitoring (fonte:[1.4] p.51)

do ed una accuratezza inferiore nella misurazione, e ne ha tardato l'approvazione come dispositivo medicale dalla parte della "Food and Drug Administration", avvenuta soltanto nel 2016 [17] I dispositivi CGM però presentano anche notevoli vantaggi: l'elevato numero di misurazioni giornaliere consentono infatti di avere informazioni fondamentali circa l'andamento del glucosio, la variabilità glicemica, direzione, velocità, magnitudo e frequenza delle oscillazioni: permettono quindi in generale di avere una panoramica molto più approfondita sull'andamento della glicemia del paziente; inoltre, il loro grado di accuratezza è andato via via migliorando nel tempo grazie alla possibilità di implementare all'interno del sensore algoritmi "smart". Algoritmi che hanno anche aggiunto la possibilità di generare allarmi ogni qual volta i livelli di glucosio uscissero dai range fisiologici o di prevedere (anche grazie al crescente sviluppo e interesse nei confronti di algoritmi di machine learning) con un certo anticipo eventi di ipo- o iper-glicemia.
Il sensore CGM (visibile in figura 1.4) è composto principalmente da 3 dispositivi: un sensore, un trasmettitore ed un ricevitore. Il sensore è posto a contatto con la pelle del paziente, solitamente nella zona addominale o negli arti superiori, e tramette la misura rilevata ad un trasmettitore; questo, grazie ad una connessione wireless, invia l'informazione al ricevitore, che può memorizzare, elaborare o anche semplicemente visualizzare l'informazione relativa alla glicemia. La maggior versatilità del sensore CGM emerge anche in questo aspetto: il ricevitore può anche essere ad esempio uno smartphone o un tablet, rendendo il monitoraggio decisamente più "user-friendly", oltre che ad aprire le porte anche ad altre possibili funzioni; possono essere infatti memorizzate ed integrate informazioni relative a pasti e attività fisica, rendendo l'inquadramento generale dello stato di salute

del paziente molto più completo. Ciò ha decisamente contribuito all'avvento e sviluppo di nuove prospettive di ricerca, descritte successivamente al paragrafo 1.3. Ultimo aspetto, ma non meno importante, il sensore CGM ha permesso anche l'implementazione di nuove tecniche di infusione di insulina, decisamente meno invasive rispetto alle tradizionali punture addominali, contribuendo ulteriormente a migliorare la qualità della vita dei pazienti affetti da questa patologia. Queste tecniche emergenti e decisamente promettenti sono ad esempio le pompe insuliniche [18] o gli ancor più recenti sistemi di pancreas artificiale "Do-It-Yourself", ai quali in particolare appartengono i dati di glicemia utilizzati nel presente lavoro di tesi e che verranno descritti successivamente in maggior dettaglio in 2.1.

1.3 Verso la precision medicine: nuove prospettive basate sulla stratificazione dei pazienti

L'avvento di nuove tecnologie come quelle del sensore CGM hanno contribuito parallelamente allo sviluppo anche di nuove esigenze: la moltitudine di dati infatti raccolti da questi nuovi dispositivi ne comporta anche adeguati metodi di gestione e utilizzo; per questo motivo, nei tempi più recenti la ricerca accademica si è concentrata sulla possibilità di utilizzare algoritmi appartenenti all'area del "machine learning" per sfruttare al meglio questa grande quantità d'informazione. Come noto, le tecniche di machine learning si possono suddividere in due macro categorie: "supervised" e "unsupervised" . Nelle tecniche "supervised", all'algoritmo solitamente è richiesto di classificare in maniera autonoma degli oggetti di cui però sia nota la tipologia o "classe"; esempi di ciò possiamo riconoscerli facilmente anche nella pratica quotidiana. Gli algoritmi di guida autonoma di veicoli si basano infatti proprio su questo: il sistema viene "addestrato" (attraverso la cosiddetta fase di "training") a distinguere un tipo di oggetto da un altro, per poi essere in grado di riconoscerne la tipologia nel momento in cui arriva un nuovo input. La categoria invece di algoritmi "unsupervised" si pone come obbiettivo quello di creare algoritmi che dato un gruppo di oggetti, ognuno con le proprie caratteristiche ("features"), senza alcuna conoscenza a priori circa le loro categorie di appartenenza, siano in grado di creare raggruppamenti di oggetti simili tra loro ("clusters"); algoritmi di questo tipo sono nello specifico denominati "algoritmi di clustering" o di "stratificazione non supervisionata". Questi algoritmi quindi risultano essere particolarmente utili quando non vi siano conoscenze a priori su un problema o un gruppo di "oggetti" di cui se ne voglia studiare le

diverse caratteristiche.
Come anticipato e brevemente discusso nei capitoli precedenti, il diabete risulta essere ad oggi una patologia in parte sconosciuta, sia nelle cause che nelle possibili diverse sfumature. Gli algoritmi di clustering quindi incontrano l'esigenza di comprendere maggiormente questa patologia: la tendenza in letteratura è quella di utilizzarli in particolare per individuare possibili "sottocategorie" di pazienti. In [21] ad esempio, Rui Tao et al. hanno utilizzato un algoritmo di clustering (k-means nello specifico) per individuare quattro sottocategorie di pazienti diabetici di tipo II: LLLFD (Low Level and Low Fluctuations Diabetes), HLHFD (High Level and High Fluctuations Diabetes), MLMFD (Moderate Level and Moderate Fluctuations Diabetes) ed infine MLHFD (Moderate Levels and High Fluctuations Diabetes); utilizzando solamente dati provenienti da sensore CGM, i ricercatori sono quindi riusciti ad isolare in maniera automatica delle classi di pazienti diabetici con fenotipo clinico diverso che potrebbero potenzialmente beneficiare di terapie maggiormente personalizzate; Lyvia Biagi et al. in [22] invece hanno sfruttato l'algoritmo di clustering k-means per ottenere quattro tipologie di "pattern giornalieri": sono stati ottenuti ad esempio cluster di profili giornalieri con una percentuale di tempo spesa nel range ipoglicemico elevata e cluster invece con basse percentuali di tempo nei range ipo- e iper-glicemici; e ancora, Corinna Schröder et al. [23] si sono riproposti di utilizzare gli algoritmi di clustering per identificare i "pattern postprandiali" di pazienti creati utilizzando il simulatore UVA/Padova; nello specifico, gli autori hanno scelto di utilizzare una versione modificata di k-means, nella quale i cluster non hanno forma circolare ma ellissoidale: queste ellissoidi sono state successivamente fittate in ognuna delle fasi postprandiali attraversate dalla curva glicemica; le dimensioni degli assi di queste ellissoidi rappresentavano il pattern glicemico del paziente considerato, permettendo quindi di ottenere delle sorte di "impronte glicemiche" per ogni paziente; infine Ivan Contreras et al. in [24] sfruttando un algoritmo di clustering gerarchico sono riusciti a categorizzare, sia in silico che in vivo, i profili giornalieri dei pazienti: in questo modo sono riusciti a raggruppare giornate con caratteristiche simili tra loro (in termini di profili glicemici), tutte appartenenti allo stesso paziente.
Parallelamente, vi è anche un'altra branca della medicina che si pone come obbiettivo quello di offrire terapie sempre più individualizzate e "paziente-centriche" [7]: la cosiddetta "precision medicine". Terapie che tengano conto quindi delle caratteristiche, delle risposte fisiologiche e dello stato di salute del singolo individuo,

permettendo auspicabilmente un'ottimizzazione elevata del trattamento terapeutico attraverso un supporto di tipo decisionale. Questa disciplina si pone quindi l'obbiettivo di risolvere un grosso limite della medicina, ovvero il dover generalizzare indicazioni terapeutiche sul maggior numero di individui possibile per poter fornire una direzione unanime nella pratica clinica.
Ciò che si propone di fare questo lavoro di tesi è porsi in mezzo a queste due discipline, clustering e precision medicine, e di esplorare possibili nuove strade che sposino entrambe, almeno negli intenti.

1.4 Obiettivo del lavoro di tesi

Come anticipato nel paragrafo precedente, il presente lavoro di tesi si propone di "fondere" le esigenze delle diverse discipline accademiche citate (machine learning, clustering e precision medicine) al fine di indagare su possibili sotto categorie di pazienti con diabete di tipo I; categorizzazioni di questo tipo permetterebbero infatti da un lato di approfondire maggiormente una patologia ad oggi ancora non del tutto conosciuta, e dall'altro di spianare la strada verso la definizione di terapie sempre più individualizzate ed ottimizzate sulle caratteristiche del paziente stesso. La scelta del tipo di dato utilizzato per perseguire ciò però non è casuale: l'obbiettivo è quello di cercare di "stratificare" pazienti con diabete di tipo I, in maniera "unsupervised", ma allo stesso tempo parsimoniosa dal punto di vista del numero delle tipologie di dati in input all'algoritmo: sono stati utilizzati in particolare solamente dati provenienti da sensore CGM, senza nessuna informazione a priori su pasti, insulina o caratteristiche cliniche dei pazienti stessi. Lo schema generale delle varie fasi del presente lavoro di tesi è riporta in figura 1.5: anche qui è possibile ritrovare il percorso già anticipato nell'abstract, composto di estrazione dei dati dal database (fase 1 in figura), preprocessing (fase 2) e partizionamento (fase 3) dei dati, estrazione delle features (fase 4); una volta terminate queste prime fasi di elaborazione dei dati ed estrazione delle informazioni, sono state condotte parallelamente due tipi di analisi, utili al perseguimento dei seguenti obbiettivi: da un lato, la ricerca di possibili gruppi di pazienti con caratteristiche simili tra loro (fase 6a in figura); dall'altro di possibili sottoinsiemi di pattern glicemici settimanali (fase 6b). Successivamente, nell'ultima fase si cercherà di dare una descrizione ed una possibile interpretazione dei cluster ottenuti in entrambi i tipi di analisi, concludendo infine con una breve digressione su possibili applicazioni dei risultati conseguiti.

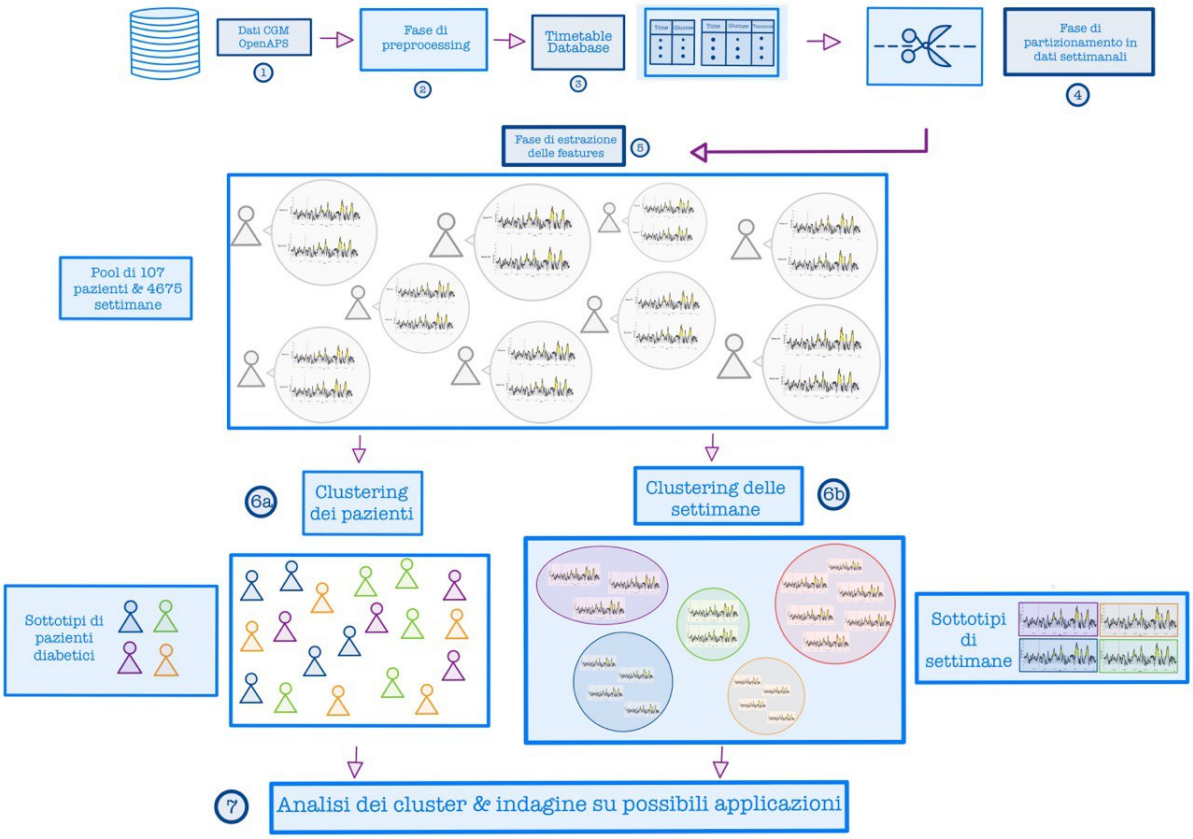


Figura 1.5: Schema generale delle fasi del presente lavoro di tesi: estrazione e preprocessing dei dati CGM del database OpenAPS (fasi 1-2); costruzione delle "timetables" con valori glicemia e corrispondenti istanti temporali (fase 3); partizionamento dei profili dei pazienti in dati settimanali (fase 4); estrazione delle features (fase 5) e successivo clustering dei pazienti (6a) e dei profili glicemici settimanali (6b); analisi dei cluster ottenuti e indagine su possibili applicazioni dei risultati di clustering (fase 7);

Nel capitolo successivo, viene riportata la descrizione del database utilizzato e delle prime fasi di elaborazione ("fase di preprocessing") di quest'ultimo.

Capitolo 2

Database e fase di pre-processing

Come anticipato nel corso del capitolo precedente, il database utilizzato nel presente lavoro di tesi è composto da lunghe serie di dati glicemici raccolti da un sensore CGM inserito all'interno di un sistema di pancreas artificiale DIY ("Do-It-Yourself") open source denominato OpenAPS. Ne verranno di seguito in particolare descritte composizione, struttura e prime fasi di elaborazione. Nello specifico, alcuni passaggi della fase di preprocessing e della successiva fase di estrazione delle features (3) sono state eseguite tramite l'utilizzo del tool "Agata", che comprende una serie di funzioni implementate in ambiente Matlab per l'elaborazione di segnali provenienti da sensore CGM.

2.1 Il database OpenAPS

2.1.1 OpenAPS: il sistema open-source di pancreas artificiale

OpenAPS è un sistema di pancreas artificiale DIY ("Do-It-Yourself") open source, caratterizzato dalla possibilità di essere configurato autonomamente dal paziente stesso e i cui algoritmi di funzionamento sono di libera consultazione e aggiornamento da parte degli utenti. I cosiddetti DIY Artificial Pancreas infatti, sono dei sistemi di pancreas artificiale "closed-loop" e open source che, grazie alla loro semplicità e versatilità, si ripropongono di rendere accessibile sia la tecnologia del pancreas artificiale a quanti più pazienti possibili, sia la moltitudine di dati che ne deriva, alla comunità scientifica e non solo: così facendo, permettono da un lato il miglioramento continuo del sistema nell'adattarsi a soddisfare le esi-

genze terapeutiche dei pazienti stessi, dall'altro contribuiscono ad incrementare la quantità di conoscenze disponibili riguardo questa patologia. Il termine "open source" indica che algoritmo e istruzioni di funzionamento del sistema possono essere scaricati liberamente sul sito internet della community e, ad oggi, vi sono nello specifico tre sistemi DIY disponibili: OpenAPS, AndroidAPS e Loop. OpenAPS è proprio il database scelto ed utilizzato all'interno del lavoro di tesi e sarà quindi quello approfondito maggiormente nel dettaglio nel prossimo capitolo.
OpenAPS fa la sua comparsa con la prima release nel 2015 su iniziativa di un gruppo di ricercatori (anch'essi pazienti diabetici) e non è un circuito di pancreas artificiale regolarmente approvato dalla Food and Drug Administration; nonostante ciò, decine di migliaia di utenti ne fanno utilizzo ed inoltre diversi studi hanno evidenziato come i pazienti diabetici che hanno utilizzato questo sistema ne hanno tratto beneficio in termini di ridotta variabilità glicemica, un incremento del tempo in range della glicemia ed anche di miglioramento di qualità della vita [19]- [20]- [25]. L'obbiettivo di OpenAPS, e anche di altri sistemi di pancreas artificiale open source come AndroidAPS o Loop, è proprio quello di migliorare le condizioni terapeutiche e di vita dei pazienti affetti da diabete di tipo I (soprattutto per quanto riguarda il controllo glicemico notturno) e di rendere questa tecnologia di facile e rapido accesso a chiunque, evitando lunghe attese di approvazione di nuovi dispositivi terapeutici. Tutto ciò che deve fare l'utente è consultare la documentazione di OpenAPS, costruire autonomamente il circuito con le componenti compatibili indicate al suo interno ed installarvici il software open source in grado di attivare il circuito. Se quindi forti miglioramenti si erano registrati già grazie all'avvento del sensore CGM (rispetto a test SMBG seguiti da multiple infusioni "manuali"), i sistemi DIY risultano essere ancora più promettenti nel migliorare ulteriormente la terapia diabetica e la qualità di vita dei pazienti; la loro esistenza nello specifico, dimostra come la libera condivisione di algoritmi e dati provenienti direttamente dall'esperienza quotidiana degli utenti hanno contribuito a rendere questa tecnologia facilmente accessibile, disponibile all'implementazione di nuove features (vista la possibilità di collegare il sistema al proprio smartphone) e potrebbero potenzialmente portare ad una conoscenza ancor più approfondita della patologia.
Dal punto di vista hardware, il sistema OpenAPS è costituito principalmente dai seguenti elementi: una pompa insulinica, un sensore CGM, un microcontrollore ("rig") ed un dispositivo di terze parti (come uno smartphone) appartenente allo stesso utente. Il suo funzionamento è semplice ed intuitivo; il microcon-

trollore comunica attraverso una piccola antenna radio con gli altri componenti del sistema: la pompa insulinica, che gli permette di conoscere la quantità di insulina che viene infusa; il sensore CGM, che rilascia dati di monitoraggio in continua della glicemia del paziente; ed infine, il dispositivo di terze parti (tablet o appunto smartphone), il quale consente all'utente di caricare informazioni (ad esempio su pasti o attività fisica) o visualizzare quelle presenti nel rig attraverso un software appositamente implementato chiamato "Nightscout". In base alle informazioni raccolte, il microcontrollore controlla l'infusione di insulina, la cui quantità viene determinata attraverso un algoritmo euristico implementato dal team di openAPS. Il sistema prevede due tipi di funzionamento, corrispondenti a due versioni del software: 0ref0, nel quale viene calcolato il tasso di infusione di insulina basale che permette di mantenere la glicemia all'interno del range euglicemico tra un pasto e l'altro e durante la notte; 0ref01, che aggiunge alla versione precedente alcune features, tra cui la possibilità di inserire pasti e quindi di calcolare boli di insulina aggiuntivi ("meal boluses"): quest'ultima modalità in particolare è ciò che fa sì che spesso il sistema di pancreas artificiale OpenAPS venga definito "hybrid close loop".
L'algoritmo installato all'interno del microcontrollore, come precedentemente ribadito, è molto semplice e determina la quantità di insulina che deve essere infusa combinando una serie di "scenari" predetti dall'algoritmo (come ad esempio se l'assorbimento di carboidrati è nullo o scarso, se cessa all'improvviso o quanto tempo sarebbe destinata a salire la glicemia in base al pasto inserito); la combinazione di questi scenari, permette al "rig" di stimare il più basso livello di glucosio predetto che è probabile osservare nel periodo valido per effettuare un'infusione di insulina; successivamente, viene calcolata la dose di insulina che sarebbe necessaria per far sì che il minimo valore "futuro" di glicemia precedentemente predetto rientri all'interno del range target normoglicemico; questa dose di insulina calcolata viene poi utilizzata per la stima del livello di insulina basale che è necessario assumere.
Dopo questa breve descrizione del sistema di pancreas artificiale OpenAPS, vengono di seguito presentate struttura e file contenuti all'interno del database.

2.1.2 Struttura e descrizione del database

Il database con i dati raccolti con il sistema OpenAPS contiene, per ogni paziente, quattro tipi di file:

- Entries: contiene i dati provenienti dal sensore CGM, ovvero una serie di coppie di valori "data-valore glicemia", dove il valore data riporta la data del giorno, l'orario in cui è stato registrato il campione (formato da ora, minuti, secondi e fuso orario) ed infine il valore di glucosio in mg/dl;

- Treatment: questo tipo di file contiene tutte le informazioni utili per il corretto funzionamento del sistema ed è costituito da una tabella, nella quale ogni riga identifica un evento, che può essere accompagnato da un numero variabile di informazioni memorizzate nelle corrispondenti colonne; gli eventi registrati in questa tabella possono essere inseriti in maniera automatica dal sistema o manualmente dall'utente e riguardano pasti, attività fisica, boli di insulina, target temporaneo di glucosio presente nel circuito o ancora dati sulla corrente misurazione glicemica;

- Device Status: qui sono invece registrati i parametri utilizzati dal circuito di pancreas artificiale quindi ad esempio quantità di insulina basale e "on board" del soggetto, il valore della glicemia o una breve descrizione delle ragioni per cui l'algoritmo ha stimato una certa quantità di insulina basale;

- Profile: contiene informazioni utili alla visualizzazione di diversi parametri utili da parte del paziente;

In base allo scopo del presente lavoro di tesi, che si propone di stratificare i pazienti presenti nel database utilizzando solo i dati provenienti dal sensore di monitoraggio in continua della glicemia, l'unico tipo di dati che è stato estratto dal database e successivamente utilizzato è il tipo "entries", come facilmente deducibile dalla precedente lista di file. Viene di seguito quindi riportata una breve analisi preliminare del contenuto di questo tipo di file.

2.1.3 Estrazione e descrizione del dataset iniziale

L'obbiettivo che ci si è posti con la presente tesi, come già esposto nei paragrafi precedenti, comprende il clustering di soggetti diabetici tramite il solo utilizzo di dati provenienti da un sensore CGM. Dal database di dati raccolti con OpenAPS quindi sono stati estratti solo i file di tipo "entries", di numero variabile per ogni paziente. I file di questo tipo sono costituiti da tabelle in formato ".csv" contenenti due colonne: nella prima sono registrati data (giorno, mese, anno), orario (ore, minuti, secondi) e fuso orario ("timezone") del valore di glicemia del paziente, registrato invece in mg/dl nella seconda colonna. Si sono quindi

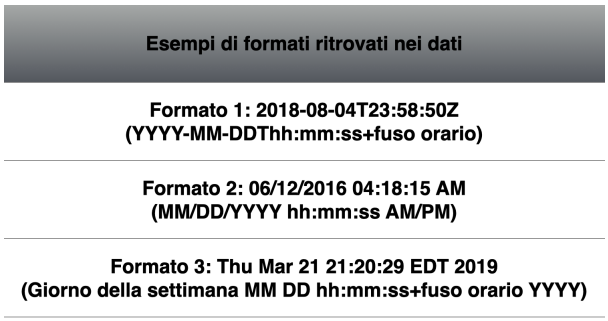


Figura 2.1: Esempi di formati diversi di data presenti nei dati (Y=cifra dell'anno, M=cifra del mese, D=cifra del giorno, h=cifra ora, m=cifra minuti, s=cifra secondi)

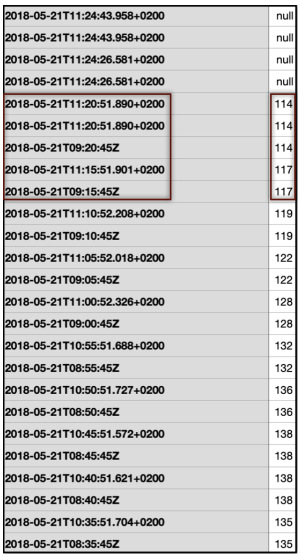
importati questi file in ambiente Matlab per la successiva fase di pre-processing, riuscendo a raccogliere in particolare i profili glicemici di 120 pazienti totali, con una frequenza di campionamento approssimativamente attorno ai 5 minuti.

2.2 Fase di pre-processing del dataset iniziale

A seguito dell'importazione dei file di tipo "entries" descritta nel paragrafo precedente, è stata effettuata un'analisi qualitativa e preliminare dei dati ottenuti, al fine di individuare quali fossero i passaggi fondamentali di pre-processing per eliminare dai dati eventuale rumore o errori di acquisizione. Di seguito quindi vengono riportate le anomalie rilevate con la corrispondente procedura che na permesso l'eliminazione.

2.2.1 Risoluzione del problema della timezone e presenza di formati multipli

Il primo problema che si è dovuto risolvere è stata la presenza di formati multipli all'interno dei dati indicanti gli istanti temporali dei campioni di glucosio, come si evince dalla figura 2.1. È stato necessario nello specifico implementare una funzione che riconoscesse in maniera automatica il formato della data, convertisse tale data in formato "datetime" (consente di rappresentare una data in ambiente Matlab) inserendo nel campo "InputFormat" il formato appena riconosciuto, ed infine producesse in uscita dati temporali espressi nell'unico formato "giorno/mese/anno/orario", accompagnati dal corrispondente fuso orario.
Successivamente si è dovuto risolvere il problema della presenza cambi improvvisi di fuso orario, dovuti probabilmente alla disconnessione del sensore o ad un qual-



2018-05-21T11:24:43.958+0200	null
2018-05-21T11:24:43.958+0200	null
2018-05-21T11:24:26.581+0200	null
2018-05-21T11:24:26.581+0200	null
2018-05-21T11:20:51.890+0200	114
2018-05-21T11:20:51.890+0200	114
2018-05-21T09:20:45Z	114
2018-05-21T11:15:51.901+0200	117
2018-05-21T09:15:45Z	117
2018-05-21T11:10:52.208+0200	119
2018-05-21T09:10:45Z	119
2018-05-21T11:05:52.018+0200	122
2018-05-21T09:05:45Z	122
2018-05-21T11:00:52.326+0200	128
2018-05-21T09:00:45Z	128
2018-05-21T10:55:51.688+0200	132
2018-05-21T08:55:45Z	132
2018-05-21T10:50:51.727+0200	136
2018-05-21T08:50:45Z	136
2018-05-21T10:45:51.572+0200	138
2018-05-21T08:45:45Z	138
2018-05-21T10:40:51.621+0200	138
2018-05-21T08:40:45Z	138
2018-05-21T10:35:51.704+0200	135
2018-05-21T08:35:45Z	135

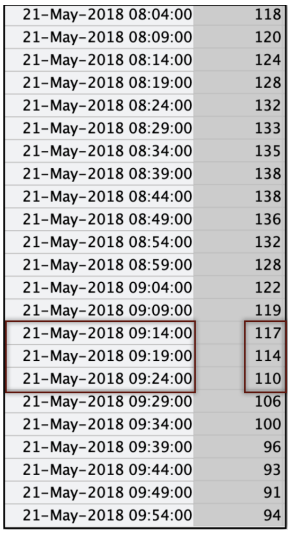
Figura 2.2: Esempio di oscillazione di fuso orario nei dati (nella prima colonna è registrata la data del campione, mentre nella seconda il valore del campione di glucosio in mg/dl))

che malfunzionamento del sistema OpenAPS; un esempio di questa problematica lo si può osservare nella figura 2.2. I dati di alcuni pazienti quindi presentavano oscillazioni improvvise di fuso orario, più volte nella stessa giornata, o valori uguali e consecutivi di glucosio associati a timezone diverse. Questo problema è stato risolto con un algoritmo che può essere riassunto nei seguenti passaggi:

1. Individuazione del "vero" fuso orario della giornata: l'idea consiste nell'individuare il corretto fuso orario della giornata ed eliminare quelli spuri; si sono quindi, per ogni paziente, isolati i dati di ogni giornata e si è calcolato il fuso orario più presente all'interno di essa; si è successivamente confrontato questo valore con quello dei due giorni successivi (estratto seguendo la stessa procedura); se entrambi i giorni successivi possedevano la stessa timezone prevalente, allora la timezone del giorno corrente veniva convertita a quest'ultima per tutti i campioni della giornata; se invece le due timezone erano diverse, la timezone dei dati del giorno esaminato veniva posta come quella più vicina tra quella del giorno precedente e del giorno successivo;

2. Applicazione del fuso orario corretto, determinato nella fase precedente, ai dati della giornata considerata;

3. Redistribuzione dei dati temporali su una griglia di campionamento omogenea nel tempo attraverso il tool "Agata"; successivamente a questa fase quindi i dati temporali e i corrispondenti valori di glucosio si presentano



21-May-2018 08:04:00	118
21-May-2018 08:09:00	120
21-May-2018 08:14:00	124
21-May-2018 08:19:00	128
21-May-2018 08:24:00	132
21-May-2018 08:29:00	133
21-May-2018 08:34:00	135
21-May-2018 08:39:00	138
21-May-2018 08:44:00	138
21-May-2018 08:49:00	136
21-May-2018 08:54:00	132
21-May-2018 08:59:00	128
21-May-2018 09:04:00	122
21-May-2018 09:09:00	119
21-May-2018 09:14:00	117
21-May-2018 09:19:00	114
21-May-2018 09:24:00	110
21-May-2018 09:29:00	106
21-May-2018 09:34:00	100
21-May-2018 09:39:00	96
21-May-2018 09:44:00	93
21-May-2018 09:49:00	91
21-May-2018 09:54:00	94

Figura 2.3: Esempio del risultato finale dopo l'applicazione dell'algoritmo per risolvere il problema dei fusi orari ai dati della figura 2.2 (anche qui, la prima colonna rappresenta data e orario del campione, nella seconda colonna è riportato invece il valore di glucosio in mg/dl)

distribuiti su una griglia con frequenza di campionamento pari a 5 minuti esatti.

Grazie all'utilizzo dell'algoritmo descritto, si è quindi potuto rimuovere il problema ed avere dei dati con data e fuso orario attendibili (come si può evincere dalla figura 2.3, risultato dell'applicazione dell'algoritmo ai dati della figura 2.2).

2.2.2 Eliminazione di falsi eventi di ipoglicemia prolungata e interpolazione gap temporali

Una ulteriore analisi del risultato ottenuto ha fatto emergere la presenza di falsi eventi di ipoglicemia prolungata;come si può apprezzare dalla figura 2.4, nei profili glicemici sono stati rilevati improvvisi cali dei livelli di glucosio a valori inferiori ai 40 mg/dl, prolungati nel tempo, dovuti probabilmente ad un qualche malfunzionamento del sensore; l'ipotesi è che quest'ultimo, in presenza di disconnessioni o anomalie, registri come valore glicemico il più basso valore rilevabile (rappresentante come noto la sensibilità del sensore stesso). Per ovviare il problema, si è deciso di porre a "NaN" tutti i valori di glucosio inferiori alla soglia di 40 mg/dl. Il risultato ottenuto a seguito di questa operazione è apprezzabile in figura 2.5. A questo punto si è resa necessaria una fase di interpolazione al fine di colmare i "gap temporali" presenti nelle lunghe serie di dati rappresentanti i

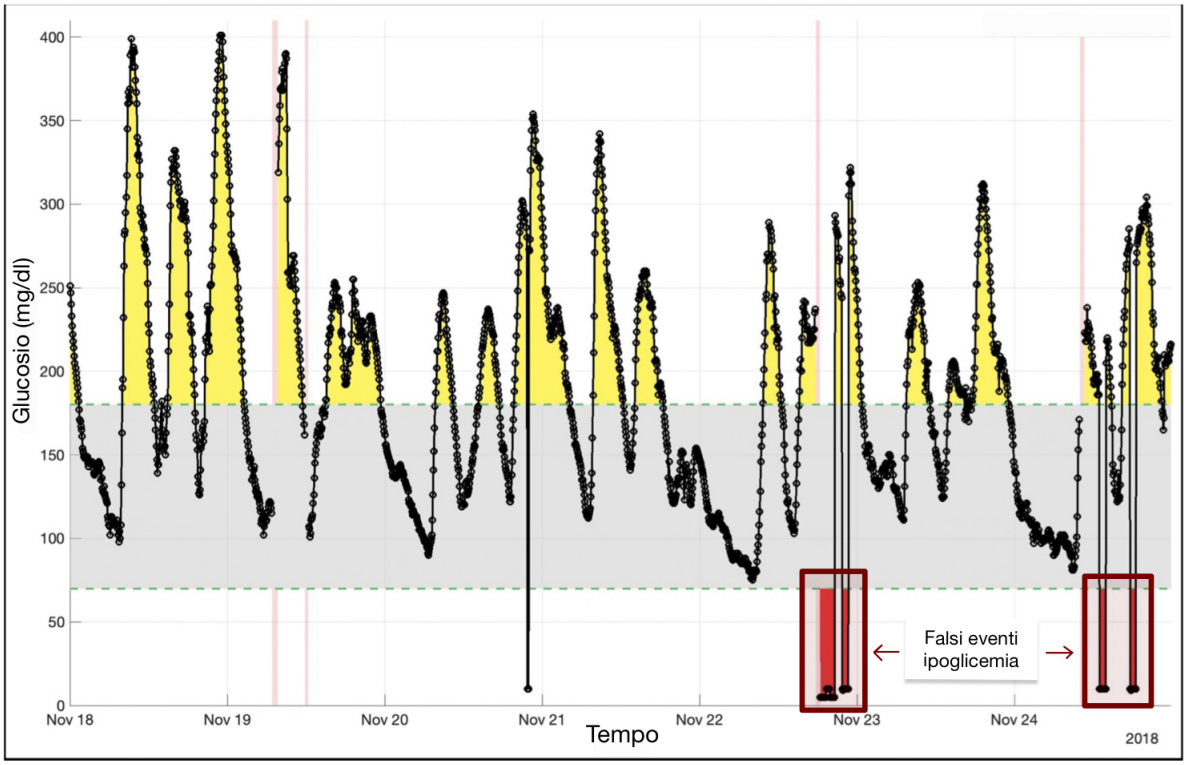


Figura 2.4: Esempio di traccia con falsi eventi di ipoglicemia prolungata

profili glicemici dei pazienti. Si è in particolare optato per una semplice interpolazione lineare, effettuata grazie al tool "Agata", su intervalli di dati "missing" di durata massima di 30 minuti, al fine di limitare al minimo la possibile alterazione del profilo glicemico; tutti i "gap" maggiori di questa lunghezza sono stati ignorati e lasciati inalterati.

2.2.3 Costruzione time-tables e partizionamento in settimane

Dopo aver completato le operazioni descritte nei precedenti paragrafi 2.2.1-2.2.2, è stato possibile creare dei dati in formato "timetable" (presente in Matlab per creare tabelle con valori temporali associati ad ogni riga) che rappresentassero l'intero profilo glicemico del paziente. È seguita successivamente una fase cosiddetta di "partitioning", nella quale all'interno del profilo glicemico del paziente sono stati isolati gruppi di dati settimanali; vista la presenza di dati "missing" (dovuti sia a mancanza di dati sia alle operazioni precedenti), si è ritenuto opportuno considerare valide solamente quelle settimane che presentassero almeno 6 giorni di dati effettivi "non missing" (un totale quindi di almeno 1728 campioni

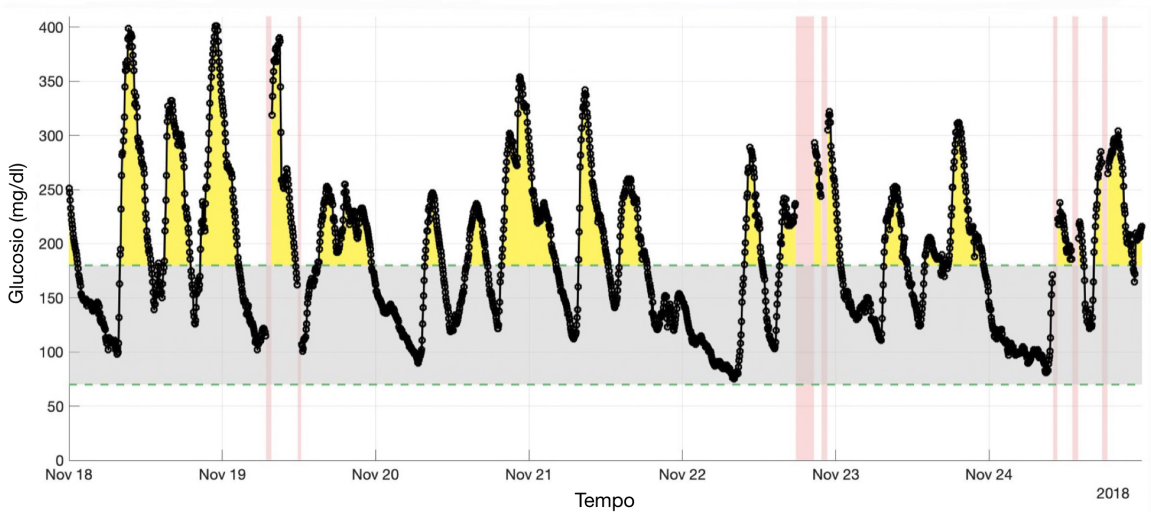


Figura 2.5: Esempio del risultato finale dopo l'eliminazione dei falsi eventi di ipoglicemia prolungata della figura 2.4

disponibili). A questo punto quindi ogni paziente risulta essere rappresentato da una serie di "timetables" con i dati delle singole settimane del paziente stesso. Un esempio di profilo glicemico settimanale ottenuto grazie a questa procedura è riportato in figura 2.6

2.2.4 Descrizione ed analisi del dataset ottenuto

Il dataset finale ottenuto comprende i dati di 108 pazienti, per un totale di 4675 profili settimanali con frequenza di campionamento pari a 5 minuti. La distribuzione del numero di settimane registrate di ogni paziente non è omogenea, come si può dedurre dalle figure 2.7-2.8. Sono presenti nel dataset quindi pazienti con un numero molto basso di settimane (inferiore alle 10), fino ad arrivare a pazienti con più di 120 settimane riportate. Dal grafico riportato in figura 2.9, si può notare che la maggior parte dei pazienti possiede fino ad un anno di monitoraggio, ma vi è inoltre una porzione consistente di pazienti del dataset, pari al 19%, che presenta i dati di due o più anni di monitoraggio. Appare quindi evidente l'esigenza di gestire con algoritmi appositi, in particolare appartenenti all'ambito del machine learning, la grande quantità di dati estratti dal dataset. Inoltre, al fine di ridurre la dimensionalità del dataset ed estrarre alcune proprietà (statistiche e non) delle tracce CGM, a questa fase è seguita la fase di estrazione delle features, discussa nel successivo capitolo.

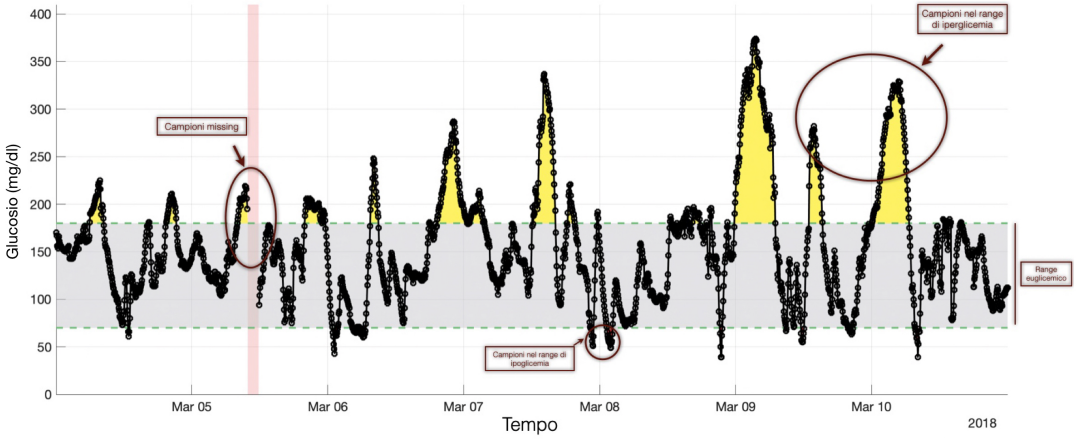


Figura 2.6: Esempio di un profilo glicemico settimanale dopo la fase di partizionamento

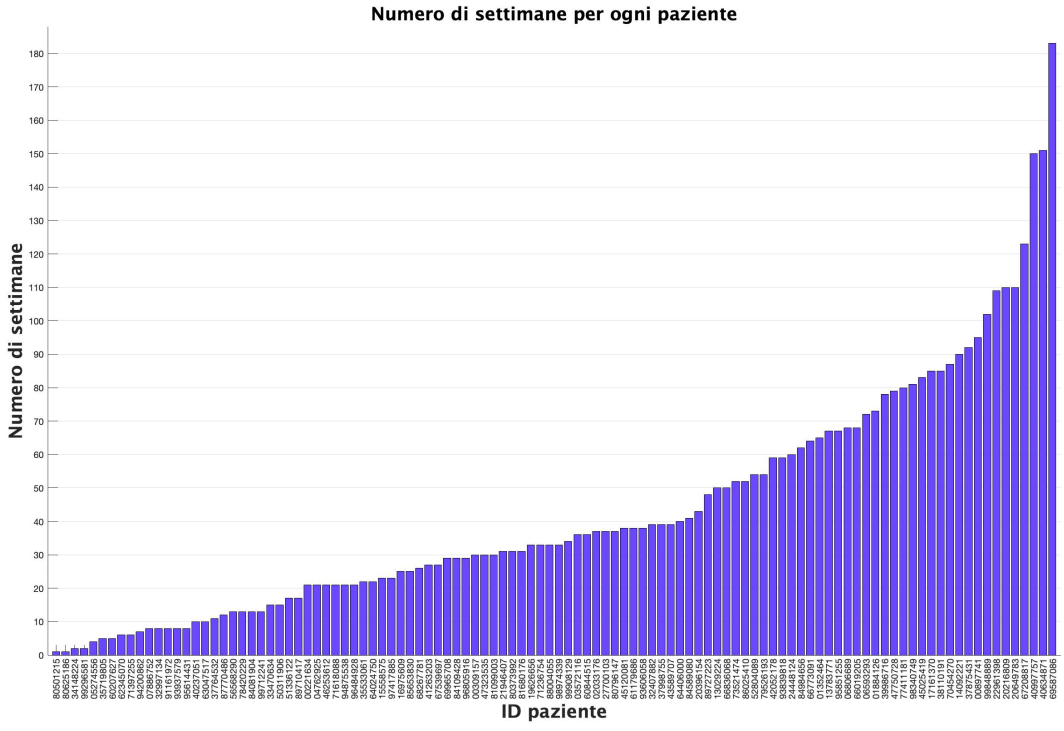


Figura 2.7: Numero di settimane per ogni paziente, identificato con il corrispondente ID

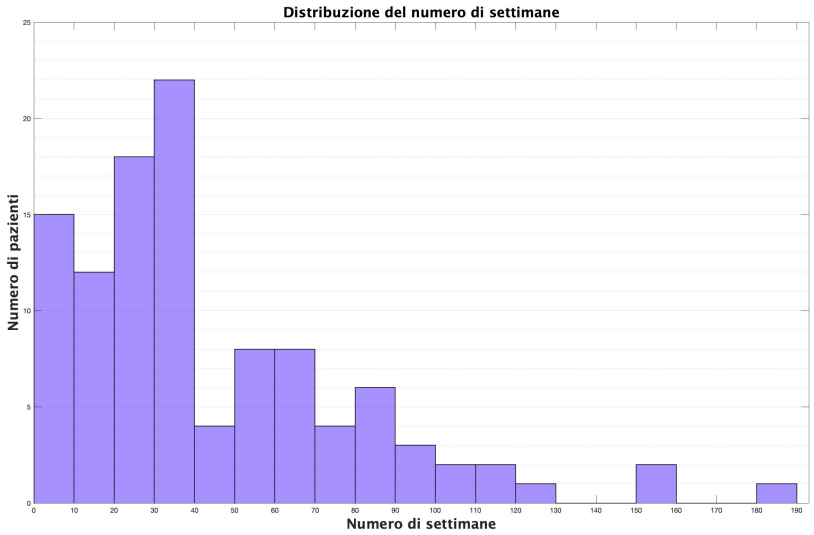


Figura 2.8: Distribuzione del numero di settimane nel dataset

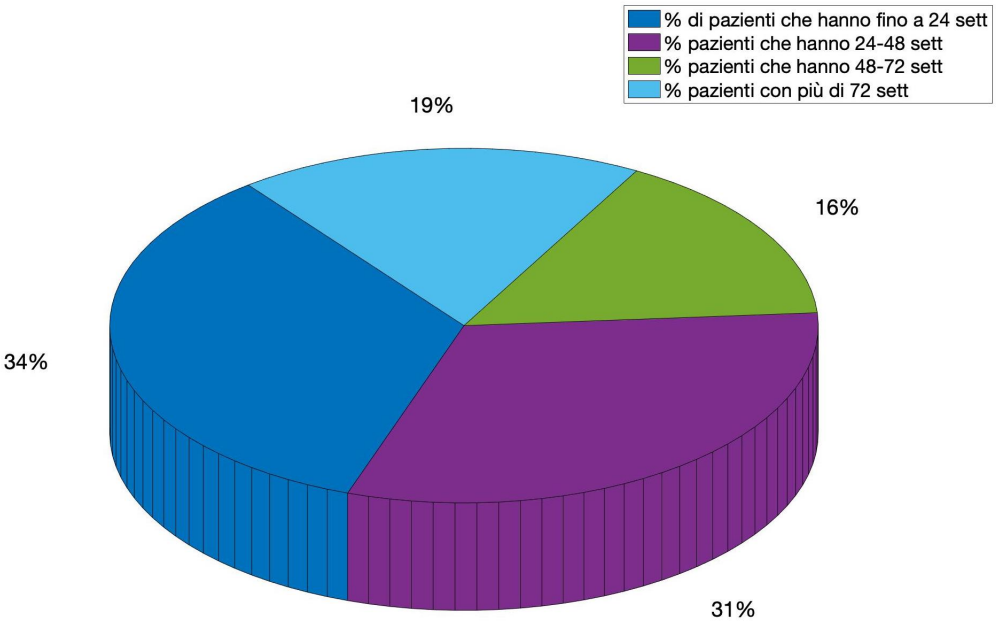


Figura 2.9: Grafico a torta con le percentuali di pazienti che presentano nel dataset rispettivamente fino a 24 settimane, da 24 a 48 settimane, da 48 a 72 settimane e più di 72 settimane

Capitolo 3

Estrazione delle features e creazione dei "clustering input datasets"

Nel seguente capitolo verrà descritta la fase di estrazione delle "features", ovvero di quelle metriche in grado di rappresentare determinate proprietà del segnale glicemico e di ridurre la dimensionalità dell'input; ne verranno in particolare brevemente descritte definizione e utilizzo in ambito di ricerca.
L'estrazione è avvenuta grazie all'utilizzo del tool "Agata" in ambiente Matlab su due tipi di dati: da un lato sull'insieme di tutti i profili glicemici settimanali di ogni paziente, dall'altro invece sui singoli profili settimanali. Ciò ha permesso di condurre parallelamente due tipi di indagine e quindi di clustering, già brevemente anticipate nel paragrafo 1.4 ed esposte con maggior dettaglio successivamente nel capitolo 4.
Il presente capitolo comincia quindi con una descrizione delle metriche scelte (3.1), per poi invece passare alla vera e propria fase di estrazione delle features, che permetterà di ricavare i due dataset di input (paragrafi 3.2.1-3.2.2) per la successiva fase di clustering.

3.1 Descrizione delle features

Le features sono state estratte, come anticipato, sui dati "globali" di ogni paziente e parallelamente sui dati dei singoli profili settimanali. La scelta delle features estratte si è basata su quelle metriche maggiormente utilizzate nella pratica clinica ed in letteratura per la caratterizzazione del segnale glicemico ed in generale per l'analisi delle condizioni cliniche del paziente e dell'andamento della terapia; sono state nello specifico calcolate in totale 42 features, riguardanti caratteristiche di

variabilità del segnale (paragrafo 3.1.1:), percentuali di tempo "spese" nei diversi range di valori glicemici (paragrafo 3.1.2), rischio (paragrafo 3.1.3), qualità del controllo del segnale (paragrafo 3.1.4) e del segnale stesso (paragrafo 3.1.5).

3.1.1 Features basate sulla variabilità glicemica

In questo paragrafo vengono descritte le metriche caratterizzanti la variabilità del segnale che sono state estratte all'interno della fase di estrazione delle features. Come affermato in [26], un'elevata variabilità del segnale glicemico risulta essere associata ad un peggior grado di controllo del segnale glicemico e ad un aumentato fattore di rischio dello sviluppo di complicanze dovute alla patologia.
Le principali features scelte ed estratte al fine di ottenere una valutazione circa la variabilità del segnale glicemico sono:

- "Area Under Curve" (mg/dl*min): il valore dell'area calcolata sotto la curva del profilo glicemico risulta essere un indice rappresentativo della magnitudo delle escursioni glicemiche e di valutazione dell'efficacia della terapia di controllo di iperglicemie postprandiali [27]- [28];

- "Control Variability Grid Analysis": questo tipo di analisi è uno strumento utilizzato in letteratura per valutare, attraverso una rappresentazione grafica, l'entità delle escursioni glicemiche e di conseguenza la qualità del controllo del soggetto [29]; nello specifico, per ogni paziente si è ricavato un grafico, all'interno del quale ogni profilo settimanale è rappresentato da un punto con due coordinate (asse X: valore minimo di glucosio, asse Y: valore massimo di glucosio): così facendo, si ottiene un grafico simile a quello in figura 3.1, dove è possibile individuare diversi range rappresentanti la qualità del controllo glicemico; nello specifico, il profilo settimanale con il controllo migliore sarà quello con distanza euclidea minore dall'origine degli assi (zona verde della figura 3.1), mentre quello con qualità del controllo peggiore si troverà all'estremo opposto (zona rossa figura 3.1); nelle altre regioni si possono trovare profili con mancato controllo del grado di ipoglicemia (zona "Lower D" della figura 3.1) o di iperglicemia (zona "Upper D" della figura 3.1). In particolare, il valore scelto per rappresentare l'esito della seguente analisi sui pazienti del dataset del presente lavoro di tesi è stata la distanza euclidea del profilo settimanale con distanza minima dall'origine degli assi (e quindi quello rappresentante la settimana con la miglior qualità del controllo tra i profili settimanali del paziente).

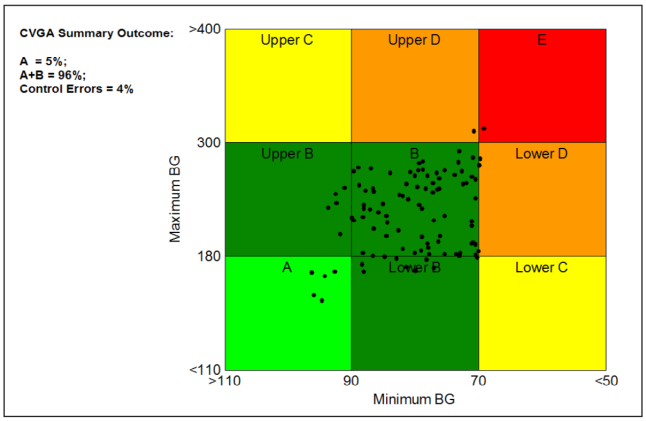


Figure 1. Control variability grid analysis: each point represents the extreme value of a patient over the considered time period.

Figura 3.1: Esempio di plot per analisi CVGA (fonte [29])

- Media, deviazione standard e mediana della concentrazione di glucosio: questi valori calcolano rispettivamente il valor medio (la cui distribuzione nel dataset di pazienti è riportata in 3.2), la deviazione standard e il valore mediano dei campioni del segnale CGM; nello specifico sono stati calcolati secondo le formule:

$$MEDIAglucosio = \frac{\sum_i CGM_i}{N} \tag{3.1}$$

$$SDglucosio = \frac{\sqrt{\sum_i (CGM_i - MEDIAglucosio)^2}}{N-1} \tag{3.2}$$

$$MEDIANAglucosio = CGM(\frac{N+1}{2}) \tag{3.3}$$

dove CGM_i è l'i-esimo campione della traccia CGM e N indica il numero totale di campioni d di glucosio;

- Coefficiente di variazione del profilo glicemico (%): calcola il coefficiente di variazione (CV) del profilo glicemico; in letteratura diversi studi, tra cui [30] e [31], hanno dimostrato che il valore del coefficiente di variazione è un buon parametro per la valutazione dell'entità delle escursioni glicemiche e stabilire dei limiti entro i quali tali escursioni possano essere considerate "stabili" o "instabili"; questa soglia limite tra le due classi di variabilità in [31] è stata definita attorno ad un valore di CV pari al 36%. La formula

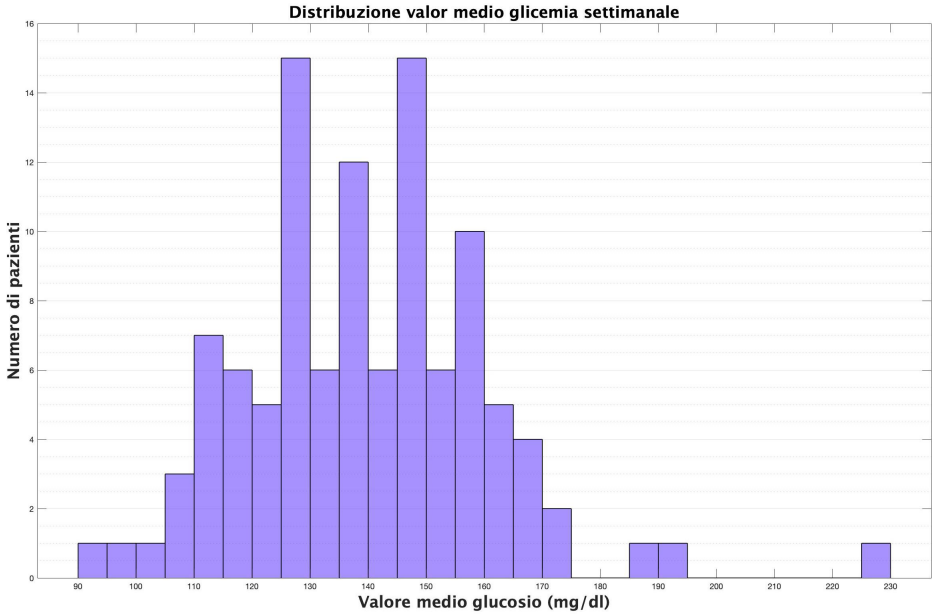


Figura 3.2: Distribuzione del valore medio di glucosio settimanale nel dataset di pazienti

utilizzata per il calcolo di questo parametro è:

$$CV(\%) = \frac{SDglucosio}{MEDIAglucosio} \tag{3.4}$$

dove SDglucosio e MEDIAglucosio sono definite rispettivamente in 3.2 e 3.1;

- Excursion Frequency Index: questo indice rappresenta il numero di escursioni glicemiche presenti nella traccia CGM che hanno ampiezza maggiore di 75 mg/dl; viene calcolato quindi come:

$$EFindex = \#\Delta_{>75mg/dl} \tag{3.5}$$

dove $\#\Delta_{>75mg/dl}$ indica il numero di escursioni (quindi la differenza tra due campioni di glicemia consecutivi) maggiori di 75mg/dl;

- Glucose Management Indicator (%): il Glucose Management Indicator (GMI) è un indice che nasce dall'esigenza di inserire come indicatore della variabilità glicemica la quantità di eHbA1C (emoglobina glicata stimata); la misura diretta infatti di questa grandezza spesso non coincide con la sua stima, ma è comunque stato dimostrato essere un importante indicatore della variabilità del profilo glicemico [32]; si è quindi deciso di sostituire il

termine eHbA1C con GMI, al fine di mantenere questa metrica nel "pool" di parametri utilizzati per la valutazione della terapia diabetica e renderne il significato maggiormente intuitivo; il GMI si ottiene tramite la formula calcolata in [32]:

$$GMI(\%) = 3.31 + 0.02392 * MEDIAglucosio \tag{3.6}$$

dove MEDIAglucosio è calcolata secondo la formula 3.1 ed espressa in mg/dl;

- Scarto interquartile della concentrazione di glucosio: fa parte degli indici di dispersione ed indica l'ampiezza della fascia di valori tra il terzo ed il primo quartile, e quindi la fascia che contiene la metà dei valori osservati;

- J index: è un indice di variabilità formulato in [33]: si propone di valutare la variabilità complessiva del profilo glicemico tenendo conto di due dei maggiori fattori che ne rappresentano una diretta influenza, ovvero media e deviazione standard del segnale glicemico; il calcolo dell'indice J è stato possibile adottando la seguente formula:

$$J = 0.001 * (MEDIAglucosio + SDglucosio)^2 \tag{3.7}$$

dove MEDIAglucosio è espressa in mg/dl;

- Range totale del profilo CGM: viene calcolato come la differenza tra il valore massimo assunto dal profilo glicemico e dal valore minimo 3.8; rappresenta quindi l'ampiezza totale del range di valori assunti dalla traccia CGM;

$$RANGEtot(mg/dl) = max(CGM_i) - min(CGM_i) \tag{3.8}$$

- Deviazione Standard della "glucose Rate Of Change": la "glucose rate of change" è definita in [34] da Clarke et al. e calcolata come la differenza tra il valore della glicemia ad un certo istante di tempo ed il valore registrato 15 minuti prima; indica quindi il tasso di variazione della glicemia, ovvero quanto velocemente varia il segnale glicemico nel tempo;

- Mean Amplitude of Glycemic Excursions (MAGE, mg/dl): questa features indica l'ampiezza media delle escursioni glicemiche e si è dimostrata essere utile per ricavare delle soglie che permettessero di distinguere soggetti sani

da soggetti diabetici stabili e non [35]; nello specifico, il valore di MAGE risulta compreso tra 22 e 60 mg/dl per i soggetti sani, tra 67 e 82 mg/dl per pazienti diabetici stabili ed infine tra 119 e 200 mg/dl per pazienti diabetici instabili; il suo valore totale nello specifico è dato dalla somma di due tipologie di escursioni, positive e negative, che permettono di ottenere MAGE+ e MAGE-, altri due indici inclusi tra le features estratte; in particolare quindi, MAGE+ è calcolato come l'ampiezza media dell'escursione glicemica tra due campioni di glucosio che differiscono tra loro più della deviazione standard del segnale, dove il primo campione è inferiore al secondo; viceversa, MAGE- rappresenta l'ampiezza media dell'escursione glicemica tra due campioni di glucosio che differiscono tra loro più della deviazione standard del segnale, dove però il primo campione è maggiore del secondo;

- Deviazione standard del "within-day means index": questa feature è sempre un indicatore di variabilità della traccia CGM; per ottenerne il valore, è necessario, calcolare la media dei livelli di glucosio per ogni profilo giornaliero, memorizzarne il valore all'interno di un vettore, e ricavare infine la deviazione standard del vettore stesso;

- Media del "within day sd index": anche questa feature è un indicatore del grado di variabilità della traccia; è ottenuto memorizzando all'interno di un vettore le deviazioni standard di ogni traccia giornaliera e calcolando infine la media del vettore stesso;

Nel paragrafo successivo invece, verranno descritte le metriche legate a caratteristiche legate a standard temporali del segnale glicemico.

3.1.2 Features basate sulla percentuale di tempo speso in determinati range glicemici

Come anticipato, in questo paragrafo verranno invece riportate e descritte le metriche temporali calcolate sulle tracce CGM.

- Percentuale di tempo in target (%): questo valore indica la percentuale di tempo in cui i campioni di glucosio si trovano all'interno del range euglicemico, quindi tra 70mg/dl e 180mg/dl; è stata ottenuta utilizzando la seguente formula:

$$\%Target = 100 * \frac{\sum CGM_{>70mg/dl\&\&<180mg/dl}}{N} \tag{3.9}$$

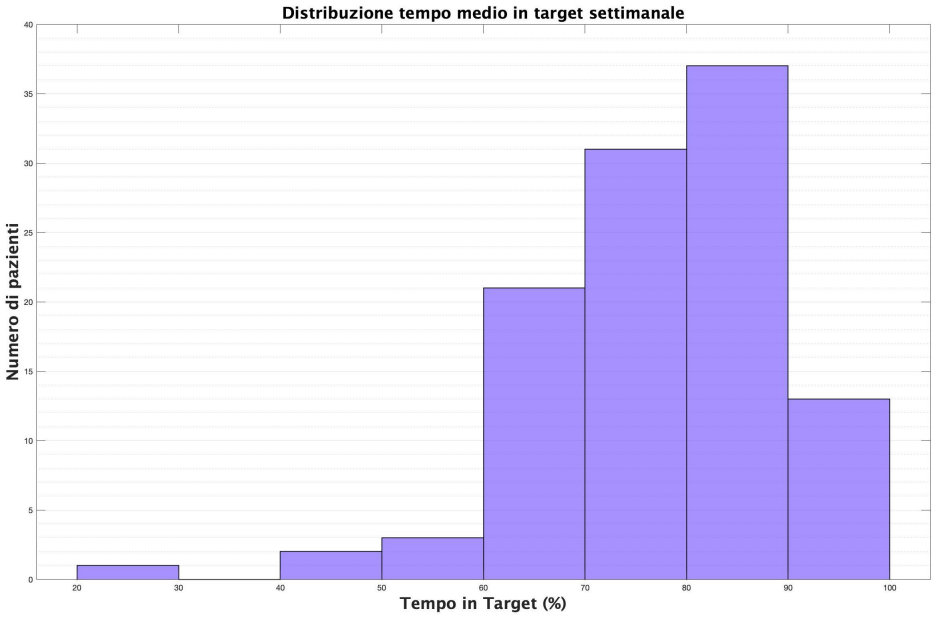


Figura 3.3: Distribuzione del tempo in target settimanale medio nel dataset di pazienti

In [10], viene suggerito di mantenere un valore percentuale di tempo in target maggiore del 70% per assicurarsi un controllo ottimale e sicuro del paziente; in figura 3.3 in particolare, è riportata la distribuzione dei valori medi settimanali di tempo in target dei pazienti (considerato quindi sull'insieme di tutte le settimane di ognuno); è possibile notare in particolare che una buona percentuale di pazienti rispetta la soglia minima del 70%.

- Percentuale di tempo in "tight target": metrica simile alla precedente, viene però ristretto il range all'intervallo 90-140 mg/dl; la formula che ne consegue quindi è:

$$\%TightTarget = 100 * \frac{\sum CGM_{>90mg/dl\&\&<140mg/dl}}{N} \tag{3.10}$$

- Percentuale di tempo in ipoglicemia: percentuale di tempo in cui i livelli di glicemia si trovano all'interno del range ipoglicemico, stabilito dall'International Consensus [10] come il range di valori inferiore ai 70mg/dl; inoltre, sempre in [10], si stabilisce che il valore ottenuto dovrebbe essere inferiore al 4% per garantire un buon controllo ed evitare le possibili complicanze già discusse nel capitolo 1; la formula per il calcolo di questa feature è la seguente:

$$\%Ipo = 100 * \frac{\sum CGM_{<70mg/dl}}{N} \tag{3.11}$$

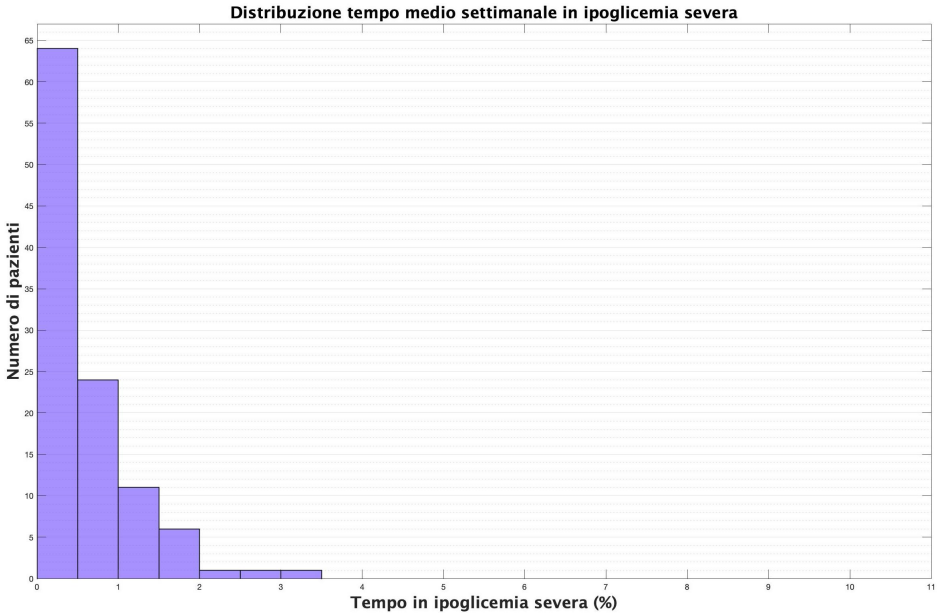


Figura 3.4: Distribuzione del tempo settimanale medio in ipoglicemia severa nel dataset di pazienti

- Percentuale di tempo in ipoglicemia severa: in questo caso si misura la percentuale di tempo in cui i livelli di glucosio si trovano al di sotto di un livello ancora più basso, in particolare al di sotto dei 54 mg/dl; questo range è definito in [10] come "range ipoglicemico di secondo livello" e, nel caso in cui la glicemia del paziente dovesse ritrovarsi all'interno di questo range, è necessario intervento medico immediato; come facilmente deducibile, la formula per il calcolo della feature è:

$$\%IpoSevera = 100 * \frac{\sum CGM_{<54mg/dl}}{N} \tag{3.12}$$

  In figura 3.4 è riportata la distribuzione del tempo settimanale medio in ipoglicemia severa dei pazienti del dataset: è possibile notare che la stragrande maggioranza dei pazienti ha un tempo inferiore all'1%, quindi questa metrica non sembra essere particolarmente critica per il dataset considerato.

- Percentuale di tempo in iperglicemia: percentuale di tempo in cui la glicemia si trova a valori superiori a 180 mg/dl e quindi all'interno della soglia iperglicemica; sempre in [10], si suggerisce di operare al fine di mantenere

questo valore al di sotto del 25%; la formula in questo caso diventa:

$$\%Iper = 100 * \frac{\sum CGM_{>180mg/dl}}{N} \tag{3.13}$$

- Percentuale di tempo in iperglicemia severa: analogamente per i range di ipoglicemia e ipoglicemia severa, anche nel range iperglicemico è stabilito un "range di secondo livello" più grave rispetto al precedente; nello specifico, l'iperglicemia severa viene stabilita quando i valori della glicemia superano i 250 mg/dl; la formula necessaria per il suo calcolo che ne consegue è:

$$\%IperSevera = 100 * \frac{\sum CGM_{>250mg/dl}}{N} \tag{3.14}$$

Nel paragrafo successivo invece vengono riportate delle features calcolate che riguardano diversi "indici di rischio" per il paziente.

3.1.3 Features basate sul rischio glicemico

Vengono di seguito riportate una serie di metriche indicanti diversi livelli di rischio per il paziente; nello specifico sono stati calcolati:

- "Blood Glucose Risk Index": abbreviato anche come "BGRI" e definito per la prima volta in [36]; la sua formulazione è dovuta alla considerazione espressa dagli autori secondo cui vi è una assimmetria nell'ampiezza dei range ipo-/iper-glicemico e euglicemico; nello specifico quest'ultimo non risulta centrato e simmetrico rispetto al centro dei possibili valori glicemici. Ciò che ne risulta, secondo gli autori, è che quando il profilo glicemico viene analizzato, le assunzioni statistiche delle metriche più comunemente utilizzate non vengono rispettate. Viene quindi proposta una trasformazione dei livelli di glicemia in scala logaritmica, definita secondo la seguente formula:

$$CGM_{i,log} = 1.509 * (log(CGM_i)^{1.084} - 5.381) \tag{3.15}$$

  Successivamente alla conversione in scala logaritmica dei valori di glucosio, viene calcolata una "funzione di rischio" di ogni valore definita come:

$$R_i = 10 * CGM_{i,log}^2 \tag{3.16}$$

A questo punto il range euglicemico è simmetrico sia all'interno del range di iperglicemia sia in quello ipoglicemico e centrato nel nuovo valore pari a 112.5 mg/dl; si è inoltre in grado di calcolare il "Low Blood Glucose Risk Index" ("LBGRI") ed il "High Blood Glucose Risk Index" ("HBGRI"), i quali rappresentano rispettivamente il numero e l'entità dei livelli bassi ed elevati di glucosio presenti nella traccia; per ogni campione di glicemia, possono essere infatti definiti:

$$LowR_i = \begin{cases} R_i, & \text{se } CGM_i < 112.5 \\ 0, & \text{altrimenti} \end{cases} \tag{3.17}$$

$$HighR_i = \begin{cases} R_i, & \text{se } CGM_i \geq 112.5 \\ 0, & \text{altrimenti} \end{cases} \tag{3.18}$$

Successivamente, su un'intera traccia CGM, è possibile definire LBGRI e HBGRI come la media dei $LowR_i$ e $HighR_i$ dell'intera traccia secondo le formule:

$$LBGRI = \frac{\sum_i LowR_i}{N} \tag{3.19}$$

$$HBGRI = \frac{\sum_i HighR_i}{N} \tag{3.20}$$

Sempre in [37], si è osservato che LBGRI risulta essere in indice del rischio di futuri episodi di ipoglicemia (sarà infatti elevato nei soggetti con livelli di glucosio prevalentemente bassi o eventi ipoglicemici particolarmente "aggressivi"); HBGRI invece, è stato dimostrato essere un indice dell'entità della variabilità del segnale glicemico (risulterà quindi elevato nei soggetti con alta variabilità glicemica). Sommando questi due fattori, gli autori hanno ottenuto il "Blood Glucose Risk Index", capace quindi di incorporare in un solo indice sia il rischio di incorrere in episodi ipoglicemici che di elevati livelli di variabilità glicemica, e quindi numero ed ampiezza delle escursioni glicemiche; BGRI quindi è definito come:

$$BGRI = HBGRI + LBGRI \tag{3.21}$$

Tutte e tre le metriche sono state inserite nel set di features estratte;

- "Average Daily Risk Range": questa feature (detta anche "ADRR") è stata definita per la prima volta in [37] e nasce dall'esigenza di avere un indice che

fosse equamente predicibile sia di escursioni glicemiche elevate (altamente correlate a iperglicemia) sia del rischio di incorrere in eventi di ipoglicemia; in [36] in particolare, al fine di validare questa metrica, è stata confrontata in termini di predicibilità sia di ipoglicemia che di iperglicemia, con le performance di altri indici utilizzati in letteratura: nello specifico, ADRR è quella che ha dimostrato una miglior ed equamente bilanciata sensitività nella predizione di ipoglicemia e di iperglicemia. Per calcolare ADRR, è necessario trasformare i dati di glucosio in scala glicemica tramite 3.15 e calcolare per ogni giornata gli indici $LowR_i$ e $HighR_i$ e di tutti i campioni giornalieri (secondo rispettivamente 3.17 e 3.18); il valore di ADRR associato a quella giornata sarà pari a:

$$ADRR = \frac{1}{N} \sum_i (LowR_i + HighR_i) \tag{3.22}$$

Come per BGRI, anche ADRR (essendo ugualmente combinazione di LBGRI e HBGRI) è risultato essere un indice efficiente in termini di predizione di future escursioni glicemiche nei range di ipo- ed iper-glicemia; nello specifico, sempre in [36], si sono ricavati diversi range di valori per ADRR rappresentanti diversi gradi di rischio: "rischio basso" se ADRR<20; "rischio moderato" per valori di ADRR compresi tra 20 e 40; "alto rischio" se ADRR>40;

- "Glycaemic Risk Assessment Diabetes Equation" (detto "GRADE score"): questa feature è stata formulata da N.R. Hill et al. in [38] e permette di ottenere un unico indice che rappresenta il rischio associato ad un profilo glicemico; viene determinato attraverso la definizione di una funzione di rischio associata a dei pesi ed utilizzando la formula:

$$GRADE_{score} = 425 * (log(log(\frac{CGM_i}{18})) + 0.16)^2 \tag{3.23}$$

Per un soggetto sano, $GRADE_{score}$ risulta essere minore di 5, mentre i soggetti diabetici possiedono $GRADE_{score}$ oltre a questa soglia; inoltre, è possibile anche ricavare la percentuale di contributo di ogni fascia glicemica rispetto al globale; possono infatti essere calcolati anche:

$$GRADE_{hypo}(\%) = 100 * \frac{\sum GRADE_{i,CGM<70mg/dl}}{\sum GRADE_i} \tag{3.24}$$

$$GRADE_{hyper}(\%) = 100 * \frac{\sum GRADE_{i,CGM>180mg/dl}}{\sum GRADE_i} \tag{3.25}$$

$$GRADE_{eu}(\%) = 100 * \frac{\sum GRADE_{i,70<CGM<180}}{\sum GRADE_i} \tag{3.26}$$

Grazie a 3.24-3.25-3.26 è possibile quindi determinare quale sia il range glicemico che contribuisce maggiormente al rischio globale espresso grazie a 3.23; ad esempio quindi se un paziente presenta $GRADE_{score}$=10 ($GRADE_{hypo}$=18%, $GRADE_{eu}$=80%,$GRADE_{hyper}$=2%) indica un rischio glicemico elevato, con contributo maggiore di rischio derivante da eventi ipoglicemici.

3.1.4 Features basate sulla quantificazione del controllo glicemico e sulla qualità del profilo glicemico

In questa sezione sono raggruppate invece metriche indicative della qualità del controllo del profilo glicemico del paziente, corrispondenti a quelle di seguito elencate:

- "M index": questo indice viene presentato in [39] per la prima volta nel 1965 e si propone come scopo quello di fornire un valore che permettesse di valutare la qualità del controllo glicemico o della terapia diabetica in atto; in [39] inoltre sono stati ricavati dei valori soglia per la definizione di diverse qualità di controllo: "good control" per M<18, "fair control" per 19<M<31, "bad control" per M>32; la formula per il calcolo di M è quella di seguito riportata:

$$M = mean(1000 * |log(\frac{CGM_i}{100})|^3) \tag{3.27}$$

- "Index of glycemic control" ("IGC"): definito in [51] come la somma tra "Hyperglycemia index" e "Hypoglycemia index", i quali sono definiti rispettivamente come la media pesata dei valori di glucosio nei range di iper- e ipo-glicemia; i pesi utilizzati nel calcolo rendono questo indice altamente flessibile e facilmente convertibile al "mimare" il comportamento di altre metriche, come MAGE o BGRI; presente tra i parametri definiti di "controllo" in quanto altamente correlato con la percentuale di tempo in target;

Per quanto riguarda invece la qualità della traccia CGM, è stata inserita tra le features anche la percentuale di campioni "missing" della traccia stessa.

3.1.5 Analisi del numero di eventi di ipoglicemia e iperglicemia settimanali per singolo soggetto

Infine, tra le 42 features calcolate sul dataset iniziale, sono stati calcolati anche numero mediano e durata media di eventi di ipo-/iper-glicemia e di ipoglicemia severa settimanali; nello specifico, come indicato in [10], un evento ipo- o iperglicemico è stato identificato a partire dall'istante in cui, per almeno 15 minuti consecutivi, si registrassero valori di glicemia appartenenti rispettivamente alla soglia ipo- o iper-glicemica; specularmente, la fine di un evento è stata decretata a partire dall'istante in cui fossero presenti valori appartenenti alla soglia euglicemica per più di 15 minuti consecutivi. Per quanto riguarda invece la definizione di evento di ipoglicemia severa, si è adottata la definizione sempre riportata in [10], nella quale viene definito come il mantenimento per più di 120 minuti consecutivi di livelli di glicemia inferiori ai 54 mg/dl. In figura 3.5, è riportata la distribuzione tramite boxplot del numero settimanale di eventi di ipoglicemia per paziente e dal grafico evince che sostanzialmente quasi la metà dei pazienti ha un numero mediano di eventi inferiore a 5, mentre la rimanente porzione da 10 fino ad oltre 15 nei casi più estremi.

3.2 Creazione dei "clustering input datasets"

In questa sezione verranno approfonditi i due dataset, ottenuti grazie alla precedente fase di estrazione delle features, che permetteranno successivamente di effettuare la procedura di stratificazione (appunto "clustering") dei pazienti e dei profili settimanali.

3.2.1 Dataset "paziente-specifico"

Tra gli obbiettivi perseguiti dal presente lavoro di tesi, vi è quello di individuare delle possibili sottocategorie di pazienti con diabete di tipo I; al fine di ottenere ciò quindi, l'estrazione delle features è stata effettuata sull'insieme di tutte le settimane registrate per ogni paziente. Ne consegue che ad ogni paziente risultano ora essere associati media, deviazione standard e valore mediano di tutte le features: in questo modo si sono potute ottenere delle "metriche globali" associate

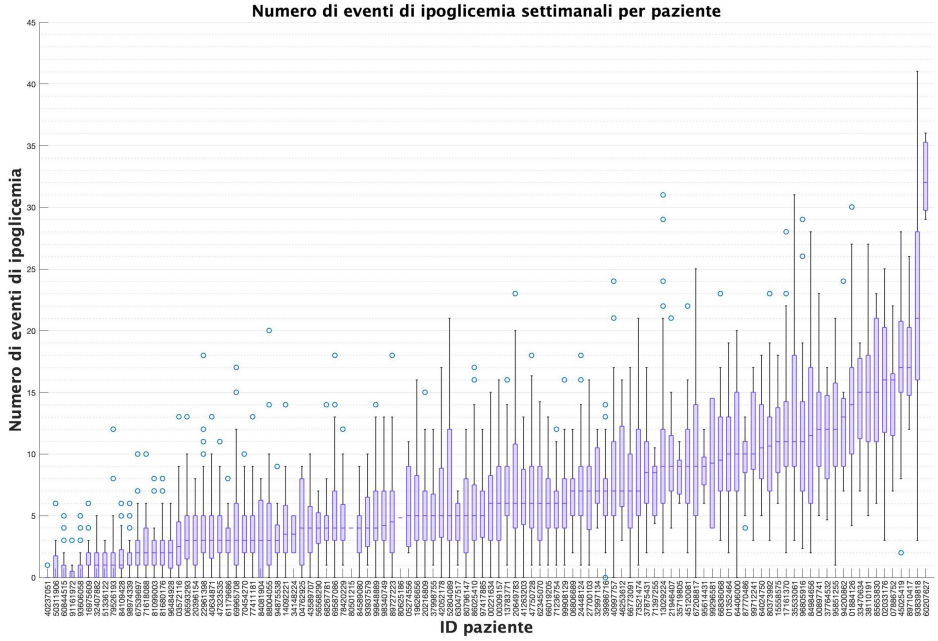


Figura 3.5: Boxplot numero di eventi ipoglicemici settimanali per ogni paziente

ad ogni soggetto, che ne permettessero quindi l'identificazione con delle caratteristiche che tenessero conto di tutte le sue settimane raccolte. A questo punto è stato creato il "clustering input dataset" necessario per il successivo algoritmo di stratificazione, ovvero una matrice di dimensioni (108,42): ogni riga della matrice quindi rappresenta un paziente, del quale sono riportati nelle 42 colonne i valori medi (mediani per il numero di eventi) delle features precedentemente calcolate.

3.2.2 Dataset "paziente-specifico"

Il secondo obbiettivo principale di questo lavoro di tesi è quello invece di ricercare dei possibili sottogruppi di dati glicemici settimanali con caratteristiche simili tra loro: questi permetterebbero infatti di ottenere dei "pattern settimanali" di glicemia ricorrenti, rappresentanti a loro volta di specifici pazienti o tipi di terapia. In questo caso quindi le features precedentemente descritte sono state calcolate singolarmente su ogni profilo glicemico settimanale; successivamente, anche in questo caso si è costruita la matrice input per il successivo algoritmo di clustering ("clustering input dataset"): qui la matrice ha dimensione (4675,42) e rappresenta quindi in ogni riga le 42 features ricavate su ogni settimana considerata singolarmente.

Effettuata la fase di estrazione delle feature, che ha permesso di ottenere due

matrici con le caratteristiche globali dei pazienti e delle singole settimane, è susseguita la vera e propria fase di stratificazione non supervisionata, esposta nel dettaglio nel capitolo successivo.

Capitolo 4

Metodologie di stratificazione non supervisionata

4.1 Introduzione al clustering

Nel seguente capitolo verranno presentati due dei principali metodi presenti in letteratura per effettuare operazioni di clustering. Gli algoritmi di clustering appartengono ad una delle due "macro aree" in cui si suddividono le tecniche di machine learning; in esso infatti possiamo distinguere due categorie di algoritmi, atti a risolvere altrettante tipologie di problematiche di classificazione: "supervised" (o supervisionata) e "unsupervised" (non supervisionata). Nel primo caso, l'obbiettivo dell'algoritmo è sostanzialmente quello di riuscire a distinguere oggetti appartenenti a classi distinte dei quali però sia disponibile a priori la classe di appartenenza; l'algoritmo quindi, attraverso una fase di "training" nella quale vengono proprio sfruttate queste informazioni a priori, è in grado di determinare i valori di determinati parametri e/o "regioni di decisione": queste permettano in futuro, una volta in cui venga richiesta la classificazione di un nuovo oggetto non appartenente al dataset sfruttato nella fase di training, di riuscire a categorizzarlo nella classe corretta. Algoritmi appartenenti a questa classe sono ad esempio le cosiddette "Support Vector Machines", reti neurali o metodi di regressione logistica [40].
Gli algoritmi invece di tipo "unsupervised", come quelli di stratificazione (o appunto clustering), si ripropongono di individuare, all'interno di un gruppo di oggetti ricevuti in input, dei sottoinsiemi (detti appunto "clusters") di oggetti che abbiamo caratteristiche o features simili tra loro; la sostanziale differenza con gli algoritmi di tipo "supervised" è che in questo caso la stratificazione avviene

in maniera automatica, basandosi su diverse misure in grado di valutare la somiglianza tra oggetti e soprattutto senza possedere a priori informazioni su possibili sottoclassificazioni degli oggetti stessi ma piuttosto analizzandone le proprietà. Il problema può essere formalizzato quindi attraverso la seguente definizione: dato un dataset di oggetti in input $X = \{x_1, x_2, ..x_N\}$, dove $x_j = \{x_{j,1}, x_{j,2}, ..x_{j,d}\} \in \Re^d$, l'obbiettivo della procedura di clustering è quello di dividere gli N oggetti in K clusters $\{C_1, C_2, ...C_k\}$ minimizzando (o massimizzando) un qualche criterio di somiglianza (o non somiglianza) tra gli oggetti; in [41] viene proposta anche la seguente definizione più informale: "un cluster è un'aggregazione di punti nello spazio dell'input tale che la distanza tra due punti appartenenti al cluster è minore della distanza di un qualunque punto del cluster ed un qualunque altro punto che non appartenga ad esso". Non essendoci però una definizione univoca e formalmente precisa del concetto di "clustering", non esiste un unico e universale criterio per ottenere la miglior soluzione al problema: piuttosto, esistono diversi criteri di clustering, ognuno dei quali impone una certa struttura agli oggetti in esame o ai cluster ricercati, e che risultano essere più adatti alla soluzione di un determinato problema (con una certa tipologia di dati) piuttosto che un altro. Motivo per cui, gli algoritmi di clustering trovano applicazione in un numero elevato di aree disciplinari [42], dalla bioinformatica all'astronomia, alle più attuali finanza o analisi comportamentale degli utenti dei social network.
In generale però, è possibile suddividere gli algoritmi di clustering in due macro categorie:

- Algoritmi gerarchici, i quali sono in grado di produrre vere e proprie "strutture gerarchiche" attraverso fusioni (o divisioni) consecutive di oggetti simili (o diversi) tra loro;

- Algoritmi di tipo "partitioning": questi algoritmi invece suddividono gli oggetti in base a misure di distanza ("k-means"), densità di oggetti nello spazio ("density-based scan") o adeguatezza ad un predefinito modello matematico ("self organizing maps");

Inoltre, la procedura di clustering può essere di due tipi: "hard clustering" (gli oggetti possono essere assegnati solamente ad un solo cluster) o "soft clustering" (nel quale invece gli oggetti possono appartenere a più cluster diversi tra loro). Nel presente lavoro di tesi, si sono adottati i due algoritmi più rappresentativi di entrambe le macro categorie, k-means e clustering gerarchico, le cui caratteristiche verranno ulteriormente approfondite nei prossimi paragrafi; a concludere il

capitolo poi, una breve descrizione della loro applicazione ai dataset ottenuti nel capitolo precedente.

4.2 L'algoritmo k-means

L'algoritmo k-means fa parte degli algoritmi di clustering di tipo "partitioning" ed in particolare di quella classe che, al fine di suddividere un dataset di oggetti in input in diversi sottogruppi, sfrutta misure di distanza tra essi; l'obbiettivo perseguito per la creazione di un cluster è semplicemente quello di minimizzare la distanza tra gli oggetti appartenenti al cluster stesso, rendendo quest'ultimo coerente al suo interno e differente dagli altri cluster creati. Per raggiungere questa condizione, l'approccio più comune è quello di ottimizzare un certo criterio attraverso una procedura iterativa di tipo "hill-climbing": partendo da una determinata partizione randomica iniziale, gli oggetti sono assegnati iterativamente ai diversi cluster in modo da ottimizzare il valore della funzione criterio; ogni partizione risulta quindi essere una perturbazione della precedente, la complessità algoritmica è di tipo combinatorio e solo una certa porzione delle possibili combinazioni può essere esplorata. Per questo motivo, la procedura di clustering tende a convergere ad un minimo locale, e non assoluto, della funzione criterio. La soluzione trovata al termine della convergenza dell'algoritmo quindi potrebbe non essere la migliore in termini assoluti.
La strategia di clustering più comunemente utilizzata è quella di ottenere una partizione degli oggetti tale che, dato un numero predefinito di clusters, miminizzi lo scarto quadrato o "varianza intracluster". Si suppone quindi che un dato insieme di N oggetti in d dimensioni $X = \{x_1, x_2, ..x_N\}$, dove l'oggetto j-esimo è tale che $x_j = \{x_{j,1}, x_{j,2}, ..x_{j,d}\} \in \Re^d$, venga partizionato in K clusters $\{C_1, C_2, ...C_k\}$ in modo tale che il cluster $C_{k-esimo}$ conti esattamente n_k oggetti e che ogni oggetto possa appartenere solamente ad un clusters; si avrà quindi che:

$$\sum_{k=1}^{K} n_k = N \tag{4.1}$$

Il centro di ogni cluster $m^{(k)}$, detto "centroide", sarà l'oggetto che avrà delle "coordinate" $\{x_{m,1}, x_{m,2}, ..x_{m,d}\}$ che avranno valore pari al valor medio di ogni coordinata di tutti gli oggetti appartenenti al k-esimo cluster; formalmente si

avrà:

$$m^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i^{(k)} \tag{4.2}$$

dove $x_i^{(k)}$ è l'i-esimo oggetto appartenente al cluster $C_{k-esimo}$. La varianza intracluster s_k^2 per il cluster $C_{k-esimo}$ sarà corrispondente alla somma delle distanze euclidee tra ogni oggetto appartenente a C_k e il suo centroide $m^{(k)}$, ovvero sarà pari a:

$$s_k^2 = \sum_{i=1}^{n_k} (x_i^{(k)} - m^{(k)})^T (x_i^{(k)} - m^{(k)}) \tag{4.3}$$

Lo scarto quadratico totale S_k^2 dell'intero dataset contenente i K clusters è la somma delle varianze intracluster di tutti i cluster e risulta pari a:

$$S_k^2 = \sum_{k=1}^{K} s_k^2 \tag{4.4}$$

L'obbiettivo quindi dell'intero processo di clustering sarà quello di trovare una partizione degli oggetti in esattamente K clusters che permetta di minimizzare S_k^2. Alternativamente, è possibile anche minimizzare il rapporto tra la varianza intracluster e la varianza tra i diversi cluster, ottimizzando quindi il rapporto definito come:

$$\frac{VAR_{within}}{VAR_{between}} = \frac{\frac{\sum_{k=1}^{K} s_k^2}{K}}{\frac{\sum_{k=1}^{K} n_k ||m_{tot} - m_k||^2}{(K-1)}} \tag{4.5}$$

dove m_{tot} è il centroide dell'intero dataset di oggetti. La partizione risultante dall'ottimizzazione di 4.4-4.5 sarà quella detta "a minima varianza". In base alle precedenti definizioni inoltre, si avrà che i clusters così formati avranno forma "sferica", con una serie di oggetti raggruppati attorno al centroide di ogni cluster, e la partizione a minima varianza sarà quella per cui le sfere saranno più compatte e separate possibile. Sono però possibili modifiche all'algoritmo e far sì che i cluster abbiano invece altre forme, come in [23] dove presentano forma ellissoidale.
Un algoritmo generico per costruire una procedura di clustering partizionale iterativa è quello riportato nei seguenti passaggi:

- Step 1: si seleziona una partizione iniziale randomica dei K clusters con i corrispondenti centroidi;

- Step 2: si genera una nuova partizione assegnando ogni oggetto al cluster con il centroide più vicino;
- Step 3: si determinano i nuovi centroidi dei nuovi cluster ottenuti;
- Step 4: si ripetono gli step 2 e 3 finchè non viene raggiunto un valore ottimo della funzione criterio (che potrebbe essere come precedentemente specificato la minimizzazione della varianza intracluster);
- Step 5: Si aggiusta il numero di cluster fondendo o dividendo i cluster ottenuti con varianza rispettivamente troppo poco o troppo elevata, oppure eliminando gli outlier;

Una possibile implementazione dell'algoritmo appena riportato è quella formulata da Lloyd nel 1982 [44]: N oggetti di dimensione D (aventi quindi D features) vengono inizialmente assegnati randomicamente a K clusters, dove K è un parametro definito a priori dall'utente; successivamente, finchè una condizione di convergenza o stop non venga verificata, eseguire iterativamente i seguenti passaggi:

- Step 1: calcolare il centroide di ogni cluster;
- Step 2: per ogni oggetto del dataset, calcolare la distanza dell'oggetto dai centroidi di tutti i cluster e assegnarlo al cluster con centroide più vicino; così facendo viene minimizzata la varianza intracluster;

La condizione di stop che permette di ottenere la configurazione finale dei K clusters può essere il raggiungimento di un numero massimo di iterazioni o la mancata modifica delle assegnazioni degli oggetti ai diversi cluster rispetto all'iterazione precedente. La complessità dell'algoritmo inoltre sarà $O(NKDT)$ dove N è il numero di oggetti e K di clusters, D è il numero di features degli oggetti ed infine T il numero di iterazioni. Essendo che i cluster vengono inizializzati nello step 1 in modo randomico, ogni esecuzione dell'algoritmo può portare a risultati diversi (che corrisponde a quanto espresso precedentemente, ovvero al raggiungimento di un minimo locale e non globale della funzione criterio); in particolare ciò risulta essere particolarmente vero per i dataset di oggetti nei quali i cluster non siano sepratati in maniera netta tra loro. Per minimizzare il margine di errore nel definire la composizione definitiva dei cluster, è possibile ad esempio eseguire l'algoritmo più volte con diverse condizioni iniziali e considerare come composizione finale quella ottenuta il maggior numero di volte.

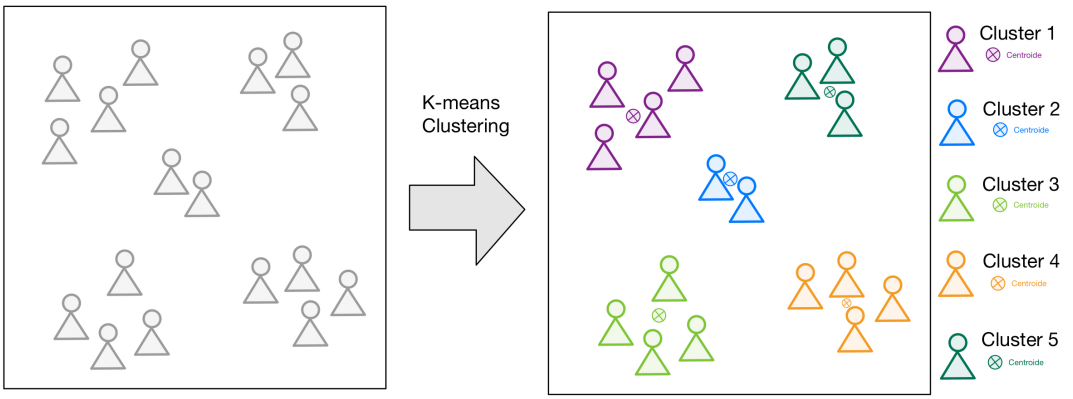


Figura 4.1: Schematizzazione funzionamento algoritmo di clustering applicato a dataset di pazienti

L'algoritmo k-means quindi potrebbe essere schematizzato e contestualizzato all'obbiettivo perseguito dalla presente tesi attraverso la figura 4.1: dato un dataset di pazienti, l'algoritmo di clustering, procedendo come finora descritto, dovrebbe essere in grado di riconoscere sottogruppi di pazienti con caratteristiche simile tra loro ed ottenere quindi i vari clusters, ognuno con il proprio centroide.
Come risulta facile intuire, uno dei parametri che è necessario definire per permettere la procedura di clustering con k-means è la definizione di distanza tra oggetti; nel prossimo paragrafo ne verranno descritte brevemente quelle maggiormente utilizzate.

4.2.1 Metriche di distanza

Come precedentemente riportato, al fine di determinare la composizione dei K cluster, è necessario definire il concetto di distanza tra elementi (o alternativamente di somiglianza). In base alla metrica selezionata, si otterranno diverse configurazioni di k-means che permettono di ottenere clusters con forma diversa. Di seguito quindi verranno riportare le più comuni, riportate in [42]:

- Distanza di Minkowski o norma L_p: la distanza tra due elementi è definita come:

$$D(x_i, x_j) = (\sum_{l=1}^{D} ||x_{il} - x_{jl}||^p)^{\frac{1}{p}} \tag{4.6}$$

  Il suo utilizzo richiede un passaggio di normalizzazione delle features nel caso in cui avessero unità di misura differente;

- Distanza Euclidea o norma L_2: caso particolare di 4.6 con p=2:

$$D(x_i, x_j) = (\sum_{l=1}^{D} ||x_{il} - x_{jl}||^2)^{\frac{1}{2}} \tag{4.7}$$

  L'utilizzo di questa metrica permette di ottenere cluster con forma sferica e risulta invariante nei confronti di traslazioni o rotazioni;

- Distanza City-Block o norma L_1: caso particolare di 4.6 con p=1:

$$D(x_i, x_j) = \sum_{l=1}^{D} ||x_{il} - x_{jl}|| \tag{4.8}$$

  Permette di ottenere cluster di forma rettangolare;

- Superior Distance o norma L∞: caso particolare di 4.6 con p=∞:

$$D(x_i, x_j) = |\max_{1<l<D} (x_{il} - x_{jl})| \tag{4.9}$$

- Distanza quadratica di Malanobis:

$$D(x_i, x_j) = (x_i - x_j)^T S^{-1} (x_i - x_j) \tag{4.10}$$

  dove S^{-1} indica la matrice inversa della matrice di covarianza intracluster, permette di ottenere cluster di forma ellissoidale;

- Correlazione di Pearson: misura il grado di correlazione tra due oggetti, largamente utilizzato nelle analisi di espressione genica:

$$D(x_i, x_j) = \frac{1 - r_{ij}}{2} \tag{4.11}$$

  dove

$$r_{ij} = \frac{\sum_{l=1}^{D} (x_{il} - m_i)(x_{jl} - m_j)}{\sqrt{\sum_{l=1}^{D} (x_{il} - m_i)^2 \sum_{l=1}^{D} (x_{jl} - m_j)^2}} \tag{4.12}$$

- Somiglianza del coseno ("cosine similarity"): metrica maggiormente utilizzata nel clustering dei documenti, non rileva la magnitudo delle differenze tra oggetti ma indipendente dal numero di features D:

$$S(x_i, x_j) = \cos(\alpha) = \frac{x_i^T x_j}{||x_i|| * ||x_j||} \tag{4.13}$$

Nel paragrafo successivo invece si procederà con la descrizione dei principali metodi per determinare il numero ottimo di cluster K con cui eseguire k-means, parametro imprescendibile per la riuscita del processo di clustering.

4.2.2 Scelta del numero ottimo di cluster

Come precedentemente specificato, al fine di ottenere la desiderata stratificazione degli oggetti del dataset utilizzando k-means, è necessario fornire come input dell'algoritmo (quindi a priori) il numero di cluster K. Questo numero K può essere determinato intuitivamente se la dimensionalità del problema non è eccessiva: ad esempio se gli oggetti sono in tre dimensioni, si potrebbe pensare di visualizzarli in uno spazio tridimensionale e determinare empiricamente il numero di cluster nel quale risultano rispettate il più possibile le funzioni criterio precedentemente descritte. Muovendoci però all'interno dell'ambito del machine learning, risulta difficile incappare in situazioni così fortunate, anzi, la dimensionalità del problema molto spesso è elevata ed è proprio in questi casi che risultano maggiormente utili tecniche e metodologie della precedentemente citata disciplina; spesso quindi l'utente non è in grado di visualizzare in maniera efficace gli oggetti del dataset e di conseguenza intuire empiricamente il numero ottimo di cluster. Per ovviare questo problema, si può procedere in due modi: da un lato si può ridurre la dimensionalità degli oggetti in input, applicando tecniche di compressione come ad esempio Principal Component Analysis, e procedendo poi con un'analisi qualitativa come quella precedentemente riportata; in alternativa, è possibile utilizzare dei metodi presenti in letteratura [46]- [47], di seguito ne vengono brevemente riportati i principali:

- "Silhouette Statistic": formulata per la prima volta nel 1990 da Kaufman e Rousseeuw [48]; questo tipo di analisi prevede di definire due quantità:

$$a(i) = \frac{1}{n_{C(i)}} \sum_{j \in C(i)} d_{ij} \tag{4.14}$$

$$b(i) = \frac{1}{n_{nearestC(i)}} \sum_{j \in nearestC(i)} d_{ij} \tag{4.15}$$

dove $a(i)$ è la distanza media dell'oggetto i-esimo dagli oggetti appartenenti al suo stesso cluster, mentre $b(i)$ è la distanza media dell'oggetto i-esimo dagli oggetti appartenenti al cluster più vicino; ciò che si ricerca è quindi minimizzare $a(i)$ e massimizzare invece $b(i)$. Si definisce "silhouette" $s(i)$ come:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4.16}$$

Si calcola quindi $s(i)$ per diversi valori del parametro K (numero di cluster) ed infine si determina il numero di cluster ottimo come K tale che:

$$K = argmax\{\frac{1}{N}\sum_{i=1}^{N} s(i)\} \tag{4.17}$$

Ovvero il numero in grado di massimizzare la media degli $s(i)$;

- "Gap Statistic": questa procedura nasce dalla considerazione secondo cui, per determinare il numero ottimo di cluster, è possibile osservare l'andamento della varianza intracluster in funzione del numero di cluster e selezionare quel valore K in corrispondenza del quale la curva diminuisce drasticamente la sua pendenza; la situazione quindi è simile a quella che può essere visualizzata in figura 4.2; come precedentemente specificato, nel caso in cui si utilizzi come metrica di distanza quella Euclidea, la varianza intracluster S_k^2 coincide con la somma delle distanze quadratiche tra gli oggetti del cluster ed è quindi definita come:

$$S_k^2 = \sum_{k=1}^{K} s_k^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k} (x_i^{(k)} - m^{(k)})^T (x_i^{(k)} - m^{(k)}) \tag{4.18}$$

Questo metodo prevede di confrontare, per diversi di K, il logaritmo di $S_k^2(log(S_k^2))$ con il suo valore atteso nella cosiddetta "null reference distribution" W_k^{null} , ovvero la distribuzione dei dati nel caso in cui appartenessero tutti allo stesso cluster (K=1); si calcola successivamente:

$$Gap(K) = log(W_k^{null}) - log(S_k^2) \tag{4.19}$$

L'obbiettivo è quello di massimizzare $Gap(k)$ e quindi scegliere il minimo

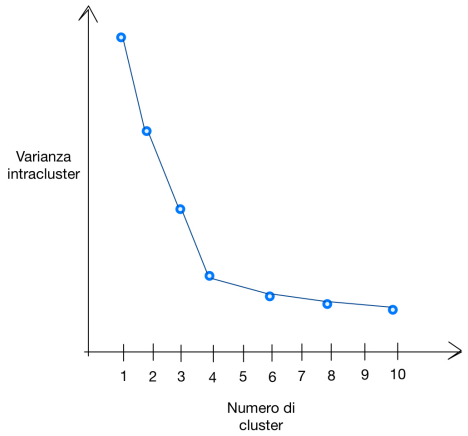


Figura 4.2: Andamento della varianza intracluster all'aumentare del numero di clusters

K per cui si abbia:

$$Gap(K) \geq Gap(K+1) - \sigma_{K+1} \tag{4.20}$$

dove σ_{K+1} è la precisione della stima di $log(W_k^{null})$;
Ciò che si ottiene quindi è il minimo numero di cluster che permetta di minimizzare la varianza intracluster;

- "Calinksi & Harabasz": uno dei metodi con maggior successo e consiste nella valutazione della relazione che intercorre tra la cosiddetta "between cluster scatter matrix" (BCSM) e la "within cluster scatter matrix" (WCSM). Più precisamente, BCSM è la somma delle distanze tra il centroide di ogni cluster ed il centro del dataset, pesate in base alle dimensioni dei cluster; WCSM è la somma delle distanze di ogni oggetto di ogni cluster dal centroide del cluster di appartenenza.
  Formalmente, si definisce il seguente rapporto:

$$\frac{trace(BCSM)}{trace(WCSM)} \frac{N-k}{k-1} \tag{4.21}$$

  Il valore di K ottimo sarà quello che massimizza 4.21;

Ricapitolando quindi, al fine di poter effettuare una procedura di clustering utilizzando k-means, è necessario scegliere sia la metrica che definisce la distanza tra oggetti sia il metodo con il quale stimare il numero ottimo di cluster K. L'ultimo parametro può risultare di difficile stima e le condizioni iniziali dell'algoritmo pos-

sono influenzare enormemente il risultato finale della procedura. Questo metodo però possiede una bassa complessità computazionale e risulta facilmente adattabile sia ad un clustering di tipo "hard" (dove gli oggetti possono appartenere ad un solo cluster) che ad uno "soft" (dove invece gli oggetti stratificati possono appartenere a più cluster). Nel paragrafo successivo invece, verrà approfondito un secondo tipo di clustering largamente utilizzato: il clustering di tipo gerarchico.

4.3 Il clustering gerarchico

Il clustering gerarchico è il secondo dei metodi di stratificazione utilizzato nel presente lavoro di tesi. Rispetto alla metodica precedentemente esposta (paragrafo 4.2), permette non solo di ottenere gruppi di oggetti con features simili, ma anche di rivelare una rigida struttura gerarchica che ne rappresenta legami e relazioni. Nello specifico, il clustering di tipo gerarchico può essere interpretato come una serie annidata e consecutiva di partizioni del dataset: questa tipologia quindi risulta essere una speciale sequenza di operazioni di partizionamento degli oggetti dell'insieme considerato. Caratteristica peculiare di questo algoritmo però è quella di tenere traccia di questa serie di partizioni e quindi dei legami intercorrenti tra gli oggetti. La struttura prodotta grazie a questo "tracciamento" è detta "dendrogram" e verrà successivamente approfondita nei paragrafi a seguire.
Altra caratteristica importante di questo metodo di stratificazione è rappresentata dal suo essere adattabile a due modalità con cui l'algoritmo può procedere, come riportato in figura 4.3: "agglomerativa" o "divisiva". Nella modalità "agglomerativa", l'algoritmo di clustering gerarchico inizia considerando ogni oggetto appartenente ad un cluster a sè stante (ovvero formato solo dall'oggetto stesso); procede poi unendo gradualmente oggetti simili tra loro in cluster di dimensione più elevata finchè non vengono raggruppati tutti in un unico grande cluster. Nella modalità "divisiva" invece, il processo risulta essere invertito: l'algoritmo quindi parte da una situazione in cui tutti gli oggetti sono racchiusi all'interno di un unico cluster e, attraverso una serie di divisioni consecutive, raggiunge la condizione iniziale della modalità "agglomerativa" in cui si hanno tanti cluster formati dai singoli oggetti del dataset. L'algoritmo di clustering gerarchico impone quindi una certa struttura "rigida" ai dati e per far ciò trasforma quella che è chiamata "matrice di prossimità" in una serie annidata di partizioni consecutive. La "matrice di prossimità" è una matrice quadrata (N, N) , con N pari al numero totale

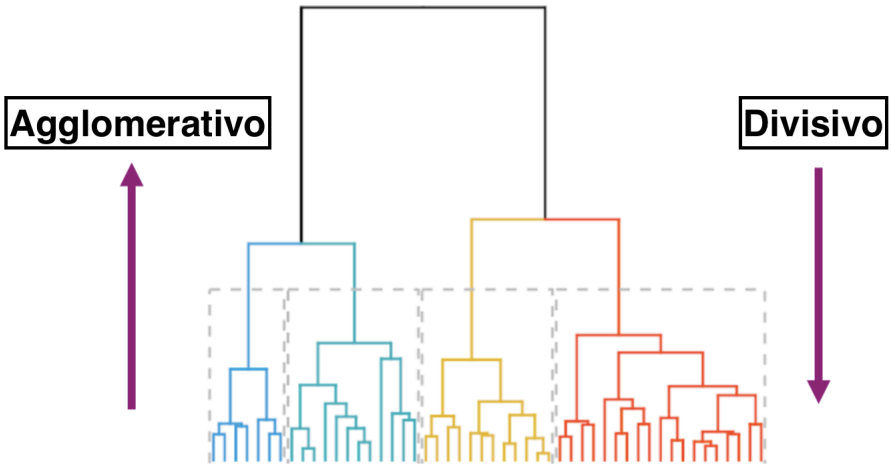


Figura 4.3: Clustering gerarchico: agglomerativo e divisivo (fonte [1.7] 51)

di oggetti nel dataset, che memorizza in ogni sua cella le distanze tra i diversi oggetti; compito dell'algoritmo di clustering è imporre come questa "matrice delle distanze" debba essere interpretata al fine di poter eseguire la serie di partizioni citata precedentemente. Inoltre, l'algoritmo deve anche definire come questa matrice debba essere aggiornata successivamente alla fusione o divisione delle entità presenti nel dataset: ciò corrisponde ad imporre quale sia il concetto di distanza adottato all'interno della procedura che viene detta anche di "linkage".
Il clustering gerarchico, ad esempio di tipo agglomerativo, si potrebbe riassumere nei seguenti passaggi:

- Step 1: l'algoritmo comincia con $K{=}N$ cluster singolari, contenenti quindi tutti un solo oggetto; viene inoltre costruita la matrice di prossimità, rappresentante le distanze tra gli oggetti del dataset (in base alla definizione di distanza adottata);

- Step 2: le entità più simili tra loro, e quindi quelle a minima distanza, vengono accorpate all'intero di uno stesso cluster; formalmente:

$$D(C_i, C_j) = \min_{1 \leq m,l \leq K} D(C_m, C_l) \tag{4.22}$$

  dove $D(\cdot, \cdot)$ è la funzione adottata per il calcolo della distanza tra oggetti, $m \neq l$; C_i e C_j saranno quindi i clusters che devono essere fusi insieme;

- Step 3: la matrice di prossimità viene aggiornata secondo diversi metodi (descritti più nel dettaglio nel paragrafo 4.3.2 a seguire), calcolando le

distanze tra il nuovo cluster C_{ij} e tutti gli altri presenti nel dataset;

- Step 4: si diminuisce K di 1 e si ripetono iterativamente gli step 2 e 3 finché non venga raggiunta la condizione per la quale tutti gli oggetti del dataset risultino essere raggruppati all'interno di un unico grande cluster.

La costruzione della matrice di prossimità e della precedentemente citata struttura gerarchica ("dendrogram"), elementi e caratteristiche fondamentali del clustering gerarchico, vengono quindi ora approfondite nel paragrafo successivo.

4.3.1 Costruzione della matrice di prossimità e dell'albero gerarchico

Al fine di rendere la procedura di clustering gerarchico, costituita da serie consecutive di partizioni o "fusioni" di oggetti simili tra loro, più facilmente comprensibile, è stata definita una "struttura grafica" in grado di riassumerne i passaggi. Come affermato nel paragrafo precedente, questa struttura è chiamata "dendrogram" ed è in grado di riassumere e catturare la "storia" dell'intero processo di clustering, nonché legami e relazioni tra gli oggetti del dataset, emergenti appunto dal processo stesso. Il "dendrogram" si presenta con una struttura "ad albero" (visibile in figura 4.3), nella quale i nodi rappresentano gli oggetti del dataset, le linee orizzontali i diversi cluster formatisi e l'altezza la misura di prossimità alla quale è avvenuta la fusione/divisione dei cluster; la "radice" dell'albero infine rappresenta il cluster dell'intero dataset, contenente quindi tutti gli oggetti considerati. Il primo passaggio necessario per la costruzione dell'albero gerarchico, consiste nella definizione della matrice di prossimità.
Come precedentemente ribadito, la matrice di prossimità è una matrice quadrata simmetrica di dimensione (N, N) , dove N rappresenta il numero totale di oggetti presenti nel dataset; la sua costruzione avviene calcolando tutte le distanze delle $N * \frac{(N-1)}{2}$ coppie di oggetti del dataset, una volta che sia stata definita la metrica di distanza o similitudine utilizzata nel processo di clustering. Determinata quindi la matrice delle distanze, la procedura di clustering, ad esempio gerarchico, si sviluppa nel seguente modo: si individua la coppia di oggetti a distanza minima (che possiedono quindi nella matrice di prossimità e nella "cella" corrispondente il numero più piccolo); questi oggetti verranno quindi verranno accorpati all'interno di uno stesso cluster e, di conseguenza, nell'albero gerarchico presenteranno, in corrispondenza della distanza alla quale si trovano, una linea orizzontale. A questo punto viene aggiornata la matrice di distanza attraverso l'utilizzo di uno dei

metodi esposti successivamente in 4.3.2, dipendenti dalla definizione di "distanza tra cluster" che l'utente decide di utilizzare. Una volta aggiornata la matrice, si ripete l'operazione di fusione delle entità a distanza minima tra loro e di aggiunta di linee orizzontali all'interno dell'albero gerarchico, fino a raggiungere l'accorpamento di tutti gli oggetti all'interno di un unico cluster e quindi la radice dell'albero stesso.
Verranno ora esposte le diverse metodologie di aggiornamento della matrice di prossimità.

4.3.2 Criteri di linkage e di aggiornamento della matrice di prossimità

Nel precedente paragrafo, è stata riportata l'esigenza di definire il concetto di distanza tra cluster al fine di poter calcolare la matrice di prossimità, di individuare elementi vicini tra loro e di poterla di conseguenza aggiornare ad ogni iterazione dell'algoritmo. Esistono una grande varietà di definizioni di distanza tra cluster (detti "criteri di linkage") ed in particolare tra un cluster C_l ed un nuovo cluster C_{ij} appena formato. È possibile però generalizzare questi concetti con l'unica formula di Lance e Williams [49]:

$$D(C_l,(C_i,C_j)) = \alpha_i D(C_l,C_i)+\alpha_j D(C_l,C_j)+\beta D(C_i,C_j)+\gamma|D(C_l,C_i)-D(C_l,C_j)| \tag{4.23}$$

dove $D(\cdot,\cdot)$ è sempre la funzione distanza scelta e α_i , α_j , β e γ sono i coefficienti che definiscono il tipo di funzione di distanza adottata. Ne verrano di seguito elencate alcune tipologie:

- "Single linkage" ("nearest neighbor"): in questo caso, la distanza tra due cluster è definita come la distanza minima tra ogni coppia di oggetti appartenenti ai diversi cluster; la 4.23 quindi diventa:

$$D(C_l,(C_i,C_j)) = min(D(C_l,C_i),D(C_l,C_j)) \tag{4.24}$$

  Ottenuta ponendo i parametri ai seguenti valori:(α_i,α_j)= $\frac{1}{2}$, β=0 e γ= $\frac{-1}{2}$;

- "Complete linkage" ("farthest neighbor"): l'opposto del single linkage, la distanza tra due oggetti è definita come la distanza maggiore tra le coppie di oggetti appartenenti ai cluster; 4.23 diventa quindi:

$$D(C_l,(C_i,C_j)) = max(D(C_l,C_i),D(C_l,C_j)) \tag{4.25}$$

e i parametri hanno valori pari a: $(\alpha_i,\alpha_j)= \frac{1}{2}$, β=0 e $\gamma= \frac{1}{2}$;

- "Unweighted Pair Group Method Average" (UPGMA): la distanza tra due cluster è definita come la distanza media tra coppie di oggetti appartenenti ai cluster; si avrà quindi:

$$D(C_l,(C_i,C_j)) = mean\{D(\forall c_l \in C_l, \forall c_{ij} \in C_{ij})\} \tag{4.26}$$

e i parametri assumono valori pari a: $\alpha_i = \frac{n_i}{n_i+n_j}$, $\alpha_j = \frac{n_j}{n_i+n_j}$, β=0 e γ=0 ((n_i, n_j) rappresentano il numero di elementi rispettivamente nei cluster C_i, C_j);
La media calcolata può anche essere una media pesata, dove i pesi sono il numero di elementi del cluster di appartenenza dell'oggetto; in questo caso la metrica viene detta "Weighted Pair Group Method Average" (WPGMA);

- "Unweighted Pair Group Method Centroid" (UPGMC): la distanza tra due cluster è definita come la distanza tra i centroidi dei due cluster, quindi si ha che:

$$D(C_{lm}, C_{ij}) = D(centroid(C_{lm}), centroid(C_{ij})) \tag{4.27}$$

dove i parametri assumono i valori: $\alpha_i = \frac{n_i}{n_i+n_j}$, $\alpha_j = \frac{n_j}{n_i+n_j}$, β=$\frac{-n_i n_j}{(n_i+n_j)^2}$ e γ=0;
anche in questo caso, si possono aggiungere dei pesi pari al numero di elementi che compongono i cluster ed ottenere quindi la misura detta "Weighted Pair Group Method Centroid" (WPGMC);

- "Ward Distance" o "Minimum Variance Distance": in questo caso si considera il generico cluster C_k con n_k elementi, del quale possiamo definire il centroide m_k e la varianza s_k^2 rispettivamente utilizzando le formule 4.2 e 4.3 (dividendo però s_k^2 per il fattore $(n_k - 1)$); all'iterazione numero t dell'algoritmo, è possibile calcola la somma delle varianze intracluster di tutti i cluster S_t^2 utilizzando una versione adattata di 4.4:

$$S_t^2 = \sum_{k=1}^{K-t} s_k^2 \tag{4.28}$$

In particolare, a mano a mano che t cresce, e vengono creati nuovi cluster più grandi fondendo assieme più elementi (e quindi K diminuisce), S_t^2 aumenta;

nello specifico si avrà un aumento pari a:

$$\Delta S^2_{t+1} = S^2_{t+1} - S^2_t \tag{4.29}$$

L'obbiettivo è quello di fondere quelle entità che vadano a minimizzare questo aumento, ovvero ΔS^2_{t+1} ; questo valore nello specifico viene chiamato anche "Ward Distance" e sarà la metrica utilizzata anche nella matrice di prossimità per calcolare le distanze tra cluster. In particolare, si può dimostrare che equivale alla quantità:

$$\Delta S^2_{t+1} = \frac{n_i n_j}{n_i + n_j} ||m_i - m_j||^2 \tag{4.30}$$

dove (n_k, m_k) indicano al solito numero di elementi e centroide del cluster k-esimo C_k ;

Una volta scelta quindi la definizione di distanza tra cluster, è possibile eseguire la procedura di clustering gerarchico, la quale produce in uscita, come precedentemente specificato, il "dendrogram"; da qui, è possibile ricavare i cluster in base al livello di osservazione dell'albero: è possibile tracciare ad esempio una linea orizzontale ad un qualche livello di altezza del dedrogram e il numero di punti di intersezione tra questa e i rami verticali dell'albero indicano il numero di cluster selezionato (osservando i raggruppamenti di oggetti sottostanti la linea orizzontale è possibile anche osservare i cluster formati). Nel prossimo paragrafo verrà brevemente descritto come determinare il numero ottimo di cluster nel caso del clustering di tipo gerarchico.

4.3.3 Scelta del numero ottimo di cluster

A differenza del clustering di tipo partizionale, il clustering gerarchico non presenta delle tecniche univoche e comprovate e per determinare il numero ottimo di cluster. In letteratura è presente qualche articolo, come [50], nel quale gli autori cercano di ottenere un qualche criterio decisionale che permetta di risolvere il problema, ma, come affermato dagli autori stessi, tutti i metodi fino ad ora proposti sono basati su assunzioni molto forti, che non permettono di rendere questi criteri adattabili per ogni tipo di dato input della stratificazione, rendendoli fortemente dipendenti dal contesto di applicazione. Per la determinazione quindi del numero di cluster possono ad esempio entrare in gioco conoscenze a priori sull'entità oggetto della procedura di stratificazione oppure è possibile determinare il numero

di cluster osservando qualitativamente i punti in cui l'albero gerarchico presenti distanze maggiori tra un braccio e l'altro in due iterazioni successive. Come anticipato nel precedente paragrafo, una volta selezionato il numero di cluster, per determinare la composizione dei diversi sottogruppi si traccia una linea orizzontale che "taglia" l'albero dove il numero di intersezioni con i rami sia pari al numero di cluster selezionato: i punti di intersezione quindi saranno proprio le "radici" dei vari cluster contenenti i sottogruppi di oggetti.

Riassumendo, il clustering gerarchico, rispetto a quello partizionale, presenta dei vantaggi ma anche degli svantaggi; risulta in particolare di semplice implementazione ed i risultati finali sono più facilmente comprensibili, vista la produzione di strutture grafiche intuitive e utilizzabili dall'utente stesso. Inoltre, per poter eseguire il processo di clustering, non è necessario fissare a priori il numero di cluster, ma piuttosto lo si può fissare a posteriori anche con l'ausilio dell'albero gerarchico. Questa tipologia di stratificazione però è computazionalmente più onerosa, non permette un clustering di tipo "soft" (dove gli oggetti possono appartenere a più cluster diversi) e tendenzialmente l'algoritmo tende ad accorpare nuovi elementi in cluster già esistenti piuttosto che crearne di nuovi, magari più esatti ed efficienti (problema denominato "chaining").

Nel prossimo paragrafo invece, vengono riportati alcuni indici di massima per la valutazione della performance della procedura di clustering.

4.4 Metriche di valutazione dei metodi di stratificazione

Le tecniche di machine learning di tipo supervised, avendo a disposizione le effettive e reali classi a cui gli oggetti appartengono, possiedono diversi indici per la valutazione delle prestazioni della procedura di stratificazione; è possibile infatti definire due tipi di risultato per l'assegnazione di un oggetto ad una sottoclasse piuttosto che ad un'altra: supponendo quindi di dover risolvere un problema di classificazione in due classi (rappresentate dalle etichette binarie "0" e "1", rispettivamente indicate come classe "positiva" e "negativa"), ogni assegnazione da parte dell'algoritmo di un oggetto ad una determinata "label", potrà essere valutata come "true positive/negative" (TP/TN) se l'oggetto è stato correttamente assegnato alla classe di appartenenza (classe "positiva"/"negativa") o "false

positive/negative" se invece viene erroneamente assegnato alla classe opposta. Calcolando quindi il numero di assegnazioni corrette ed errate, ed il rapporto che intercorre tra essi, è possibile ricavare diversi indici che permettano di valutare le performance dell'algoritmo. Esempi di questi indici sono ad esempio il "Rand Index", "F1 score" o le più diffuse "confusion matrices" ("matrici di contingenza") [43]. Nel caso però di tecniche di tipo "unsupervised", non avendo a disposizione a priori della conoscenza delle vere sottoclassi di appartenenza degli oggetti, risulta difficile se non impossibile il calcolo di questi indici. Di conseguenza, è necessario basarsi su metodi maggiormente empirici e meno formali: ad esempio, ispirandosi ad una procedura simile alla "gap statistics" riportata al paragrafo (4.2.2), è possibile eseguire l'algoritmo un numero elevato di volte e selezionare la soluzione che porta ad un valore di varianza intracluster minore di tutte le altre, permettendo in questo modo di massimizzare il livello di compattezza dei cluster ottenuti. Metodiche di questo tipo però risultano essere dipendenti e strettamente correlate al tipo di esperimento e di dati, oltre che essere "operatore-dipendenti".
La selezione del numero ottimo di cluster per algoritmi di tipo gerarchico e la formulazione di indici di valutazione della stratificazione formali ed universali rappresentano quindi ancora ad oggi delle sfide e problematiche aperte di difficile soluzione.

Nel prossimo paragrafo, verranno invece descritti metodi e parametri utilizzati per la stratificazione dei due input dataset dei pazienti e dei profili glicemici settimanali, precedentemente esposti nel paragrafo 3.2.

4.5 Stratificazione dei "clustering input datasets": metodi e parametri utilizzati

Come anticipato nel capitolo precedente, verranno ora riportati i metodi ed i parametri scelti per il clustering dei pazienti e dei profili glicemici settimanali.

4.5.1 Clustering dei pazienti

Grazie alla fase di estrazione delle features, riportata nel capitolo 3 ed eseguita su tutti i profili glicemici settimanali di ogni paziente, si sono potuti ottenere 108 vettori di dimensione (1,42) rappresentanti le features e quindi le caratteristiche

"globali" di ogni paziente; come descritto nel paragrafo (3.2.1) questi vettori sono stati memorizzati in una matrice (108,42), la quale rappresenta l'input dell'algoritmo di clustering. Nello specifico, in questa prima tipologia di stratificazione, si è deciso di utilizzare un algoritmo di tipo partizionale, k-means nello specifico, con distanza di tipo euclideo. Per la determinazione del numero di cluster, è stata utilizzata la "silhouette statistics", ispezionando un numero di cluster appartenente al range da 1 a 10 (questo al fine di semplificare la procedura di analisi dei risultati descritta successivamente nel capitolo 5): il numero ottimo di cluster selezionato è quindi quello in grado di massimizzare il valore della silhouette statistics ed in questo caso è risultato essere pari a 7 (valore della silhouette statistics pari a 0.7514); il risultato finale della procedura di clustering è stato ottenuto con un numero massimo di iterazioni pari a 10000 e dopo 2000 iterazioni dell'algoritmo. I cluster ottenuti alla fine della procedura sono quelli risultati avere una somma delle distanze tra gli oggetti contenuti ed il proprio centroide minima tra tutti quelli ottenuti nelle 2000 iterazioni.
Si è però in questo modo ottenuto un cluster in particolare formato da un solo paziente, contenente due soli profili glicemici settimanali; si è proceduto quindi applicando una tecnica di riduzione dimensionale (PCA in particolare) al fine di visualizzare la disposizione degli oggetti "paziente" nello spazio delle due componenti principali ed interpretare il risultato ottenuto. Come si può dedurre dalla figura 4.4, il cluster numero 6 composto da un solo paziente (il paziente numero 30) risulta molto probabilmente essere un outlier, com'era già facilmente intuibile. É stato quindi eliminato dal dataset di pazienti e la matrice input dell'algoritmo è quindi diventata di dimensione (107,42). È stata successivamente ripetuta la procedura di clustering partizionale mantenendo gli stessi parametri in ingresso precedentemente impostati; il numero di cluster ottenuto eliminando il paziente outlier è risultato essere pari a 6, corrispondente ad un valore della silhouette statistics pari a 0.7491. Al fine di visualizzare graficamente il risultato finale, è stata nuovamente applicata anche la tecnica PCA, il cui risultato è riportato in figura 4.5. Analisi ed interpretazione dei cluster ottenuti verranno discusse nel capitolo 6.

4.5.2 Clustering dei profili settimanali

Come precedentemente riportato al paragrafo (3.2.2), la matrice input del processo di clustering dei profili settimanali è stata ottenuta calcolando le 42 features su ognuna delle 4673 settimane ricavate dall'operazione di partizionamento dei

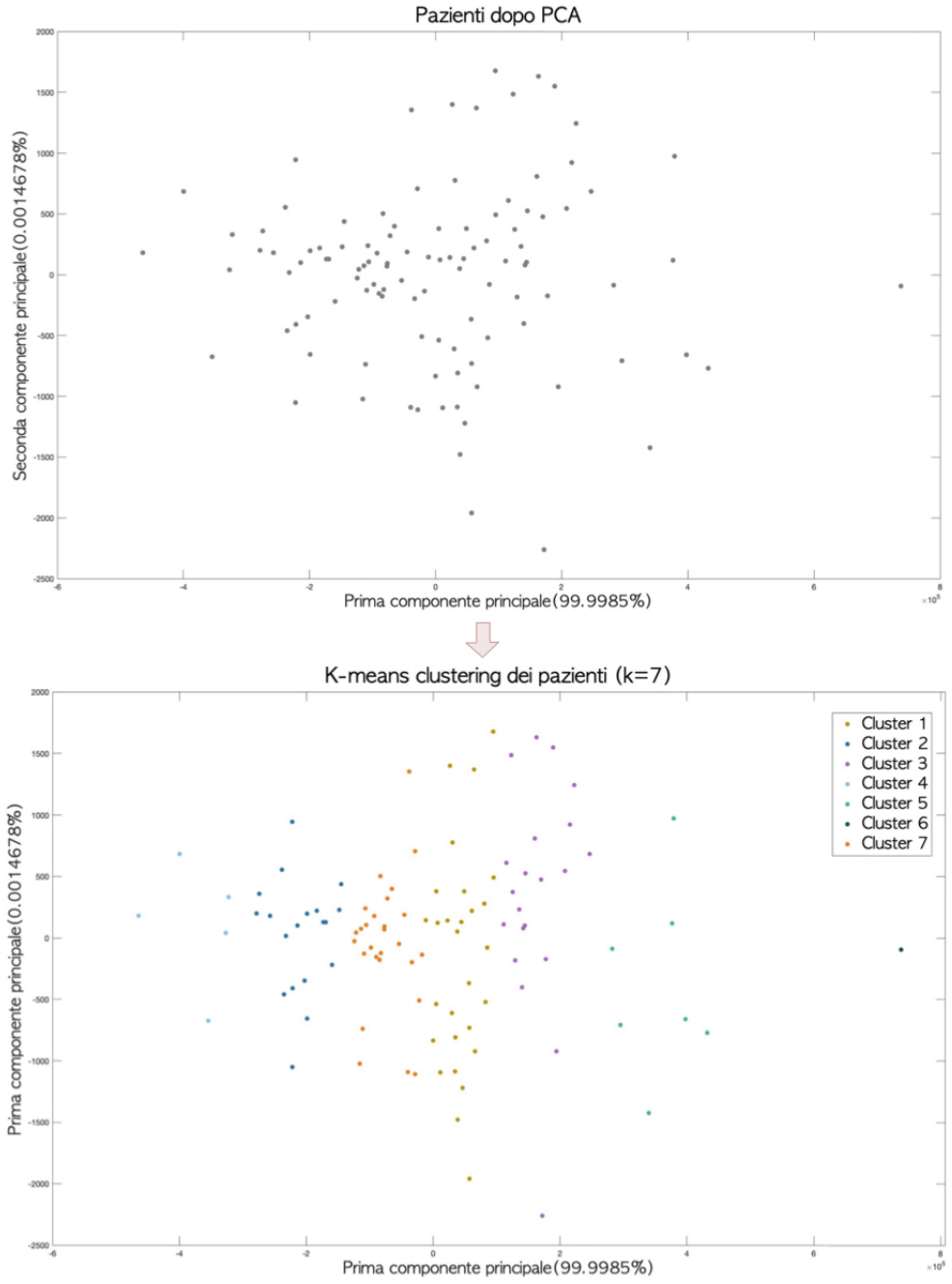


Figura 4.4: Pazienti nello spazio delle prime due componenti principali prima e dopo clustering k-means (k=7))

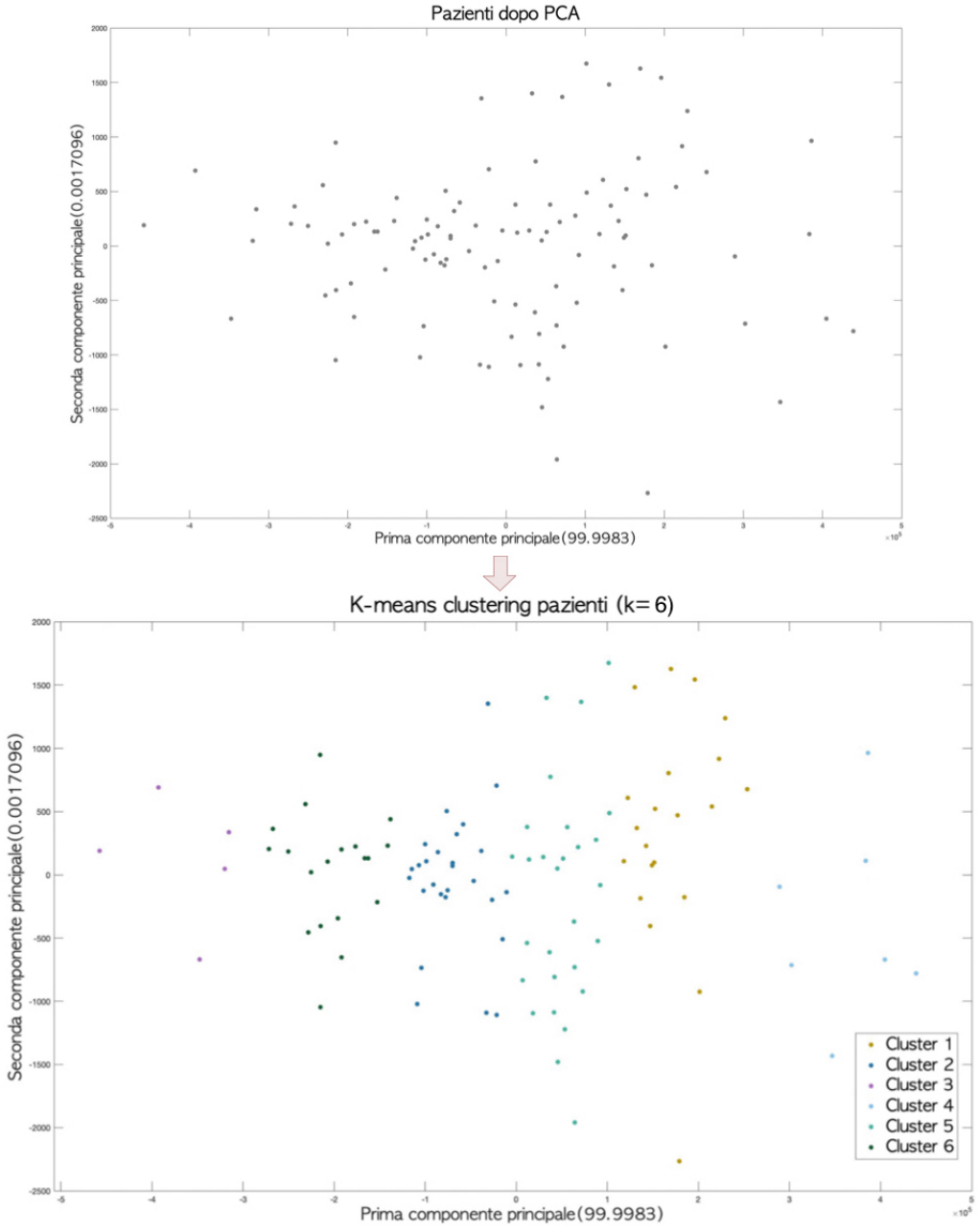


Figura 4.5: Pazienti nello spazio delle prime due componenti principali prima e dopo clustering k-means, successivamente a rimozione del paziente outlier (k=6))

dati (2.2.3); la matrice quindi come precedentemente descritto è risultata essere di dimensione (4673,42) e presenta quindi in ogni riga le features delle singole settimane, indipendentemente dal paziente a cui appartengono. Inizialmente si è testata una stratificazione con algoritmo k-means al fine di ottenere un risultato confrontabile con il precedente clustering dei pazienti e mantenere basso il carico computazionale, vista l'aumentata quantità di dati che l'algoritmo deve processare. Nonostante però l'aumento del numero di cluster ispezionati dal range 1-10 al range 1-20 e del massimo numero di iterazioni da 10^4 a 10^5, l'algoritmo non è riuscito a raggiungere un risultato soddisfacente: il numero di cluster ottenuto è risultato essere pari a 2, numero insufficiente per ottenere un'adeguata stratificazione dei profili settimanali. Si è quindi optato per un clustering di tipo gerarchico e la matrice di prossimità è stata inizializzata con distanza euclidea e aggiornata durante il processo di linkage con il criterio della "Ward distance". Per la scelta del numero di cluster si è deciso di procedere empiricamente osservando il dendrogram ottenuto e, al fine di mantenere una complessità tale da permettere un'analisi dei cluster efficace, si è optato per un numero di cluster pari a 8. Analisi e caratteristiche degli 8 cluster rappresentanti 8 sottogruppi di profili glicemici settimanali saranno riportati nel capitolo successivo.

Capitolo 5

Implementazione del processo di clustering nel database considerato

In questo capitolo verranno riportati i risultati del clustering sul dataset dei pazienti, per poi passare alla descrizione di quelli ottenuti sul dataset dei profili glicemici settimanali. Verranno quindi analizzati i cluster ottenuti, dandone una possibile interpretazione e analizzando punti di forza e limiti della stratificazione ottenuta.

5.1 Risultati clustering dei pazienti

In questo primo paragrafo verranno riportati i risultati della procedura di clustering sul dataset dei pazienti. Questo dataset, rappresentato da una matrice (107,42), è stato stratificato con algoritmo k-means, settato con i parametri riportati in (4.5.1). Nello specifico, verranno analizzati i cluster ottenuti in termini di distribuzione e valor medio delle features in ogni cluster di pazienti ottenuto tramite l'utilizzo di boxplot e tabelle, per poi riportarne una possibile interpretazione al paragrafo (5.1.2).

5.1.1 Descrizione clusters ottenuti

Come anticipato al paragrafo (4.5.1), i primi risultati sono stati ottenuti applicando l'algoritmo k-means (numero massimo di iterazioni 10^4, numero di iterazioni mediate per il risultato finale 2000) alla matrice input dei pazienti di dimensione (107,42); il numero di cluster ottimo è risultato essere 7, ovvero il valore che massimizza il valore della Silhouette nell'intervallo di valori di K da 1 a 10. È

Cluster	Numero di pazienti	Varianza intracluster
1	28	2.33e+10
2	19	2.80e+10
3	21	2.88e+10
4	5	1.41e+10
5	7	1.79e+10
6	1	0
7	27	2.97e+10

Tabella 5.1: Tabella con varianza intracluster e numero di pazienti per ogni cluster (k-means con k=7)

stato possibile inoltre ricavare anche la percentuale di varianza intracluster dei 7 cluster trovati, riportata in tabella 5.1, assieme al numero di pazienti componenti ogni cluster; è possibile notare, come anticipato nel capitolo precedente, che il cluster numero 6 è formato da un solo paziente, confermato da un valore della varianza intracluster pari a 0. Osservando inoltre le figure 4.4, che riportano rispettivamente la rappresentazione dei pazienti e dei 7 cluster ottenuti nello spazio delle due componenti principali, è possibile notare che il cluster numero 6 risulta chiaramente essere un outlier, trovandosi all'estremo destro del grafico. Si è quindi proceduto con la rimozione del paziente (e delle relative settimane) dal dataset e ripetuta la procedura di clustering con gli stessi parametri precedentemente utilizzati: l'input dell'algoritmo in questo secondo caso risultava quindi essere una matrice di dimensione (107,42).
L'analisi dei valori della Silhouette statistics ha permesso di individuare il numero di cluster ottimo, diventato pari a 6 e corrispondente ad un valore della statistica equivalente a 0.7491 (il valore precedente del clustering a 7 cluster era leggermente più elevato e pari a 0.7514). Inoltre, le varianze intracluster con il numero di componenti di ogni cluster sono riportate nella tabella 5.2: è possibile notare che, a parte i cluster 3 e 4, il numero di pazienti presenti in ogni cluster è abbastanza simile; non vi sono quindi sbilanciamenti importanti in termini di "cardinalità" della maggior parte dei gruppi ottenuti. In generale, i valori delle varianze intracluster sono elevati: questo potrebbe essere dovuto al fatto che, come si può notare dalla figura 4.5, dove sono riportati i cluster nello spazio della prima e seconda componente principale dell'input, i vari gruppi di pazienti non sono ben distinti e separati tra loro. L'algoritmo k-means quindi potrebbe non essere l'algoritmo migliore per stratificare i vettori rappresentanti i pazienti del dataset dato che, per definizione, k-means ha performance migliori proprio in

Cluster	Numero di pazienti	Varianza intracluster
1	21	2.88e+10
2	27	2.97e+10
3	5	1.41e+10
4	7	1.79e+10
5	28	2.33e+10
6	19	2.80e+10

Tabella 5.2: Tabella con varianza intracluster e numero di pazienti per ogni cluster (k-means con k=6)

quelle situazioni in cui le sottoclassi di oggetti siano ben distinte tra loro.
Successivamente sono stati ricavati i boxplot con le distribuzioni di tutte e 42 le features nei vari cluster di pazienti. Alcuni di questi sono ad esempio riportati nelle figure 5.1-5.2-5.3-5.4. Ciò che si può subito notare in maniera qualitativa è la presenza, nella maggior parte delle features, di una sorta di "pattern" ricorrente nella loro distribuzione, con una doppia tripletta di valori in scala decrescente, dal cluster 1 al 3 e poi dal 4 al 6; le distribuzioni quindi sembrano nella maggior parte delle volte essere distinte tra loro e avere poche porzioni in cui risultano essere sovrapposte. Emerge però la necessità di ulteriori test statistici, al fine di verificare però se queste distribuzioni siano effettivamente statisticamente diverse tra loro. Le differenze più marcate si possono apprezzare ad esempio nella figura 5.1, dove i valori dei boxplot dell'area sotto la curva di concentrazione di glucosio nei diversi cluster risultano essere molto diversi tra loro. Anche le distribuzioni dei valori medi di glucosio, riportate in figura 5.4 presentano differenze che sembrano essere significative tra un cluster e l'altro. Per visualizzare graficamente e globalmente i valori medi delle features nei vari cluster, è stato inoltre ricavato il cosiddetto "Parallel Coordinate Plot" in figura 5.5, dove si può apprezzare maggiormente la stratificazione ottenuta del dataset di pazienti. Al fine di analizzare ulteriormente nel dettaglio le proprietà statistiche delle features nei vari cluster ottenuti, è stata ricavata la tabella, visibile in figura 5.6, dove ogni riga rappresenta valor medio e deviazione standard della corrispondente features nei diversi cluster. Questa tabella ha permesso quindi di confrontare tra loro le caratteristiche dei cluster ottenuti: grazie al suo utilizzo nel prossimo paragrafo verranno riportate in particolare analisi e possibile interpretazione dei cluster stessi.

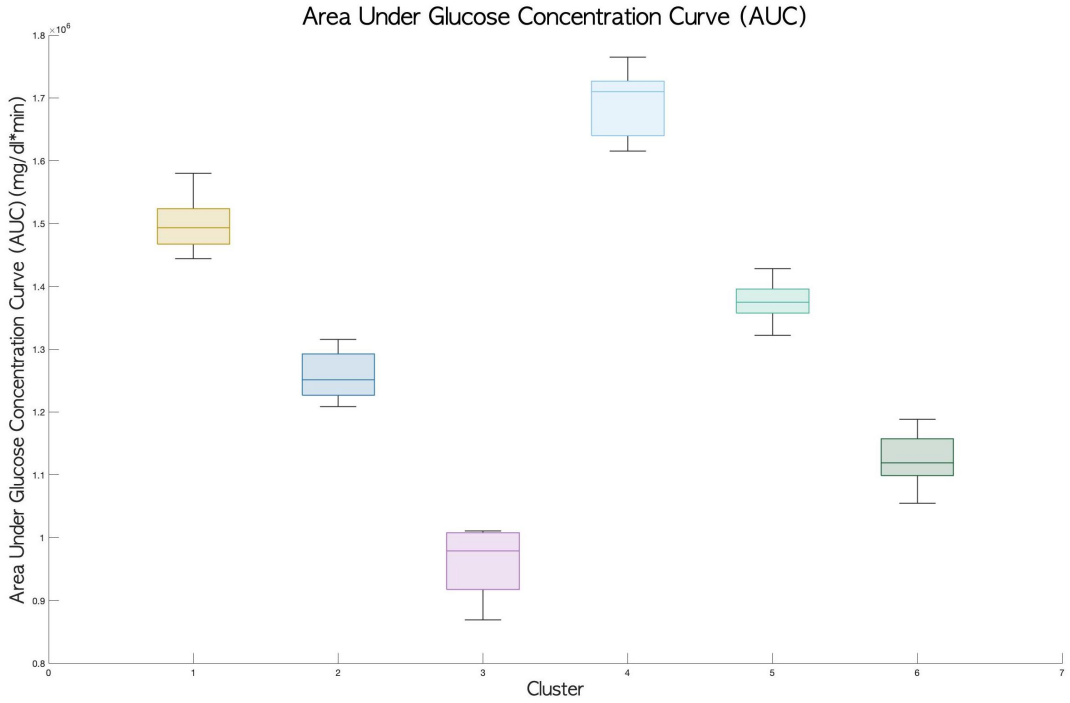


Figura 5.1: Boxplots valori area sotto la curva dei clusters

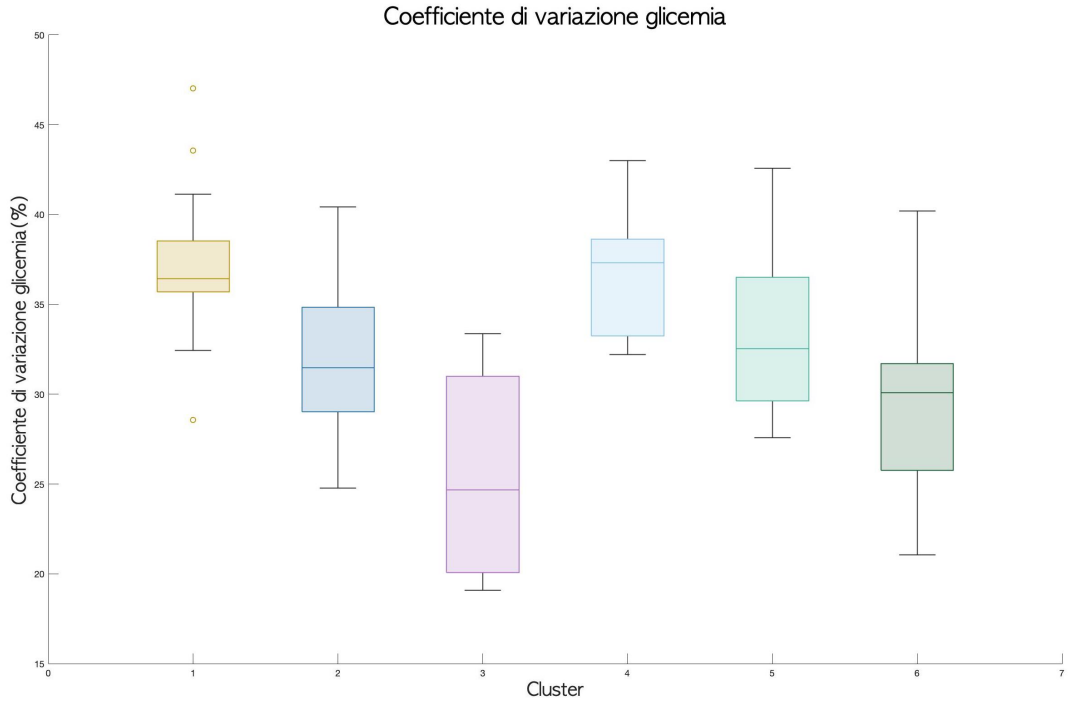


Figura 5.2: Boxplots valori coefficiente di variazione dei clusters

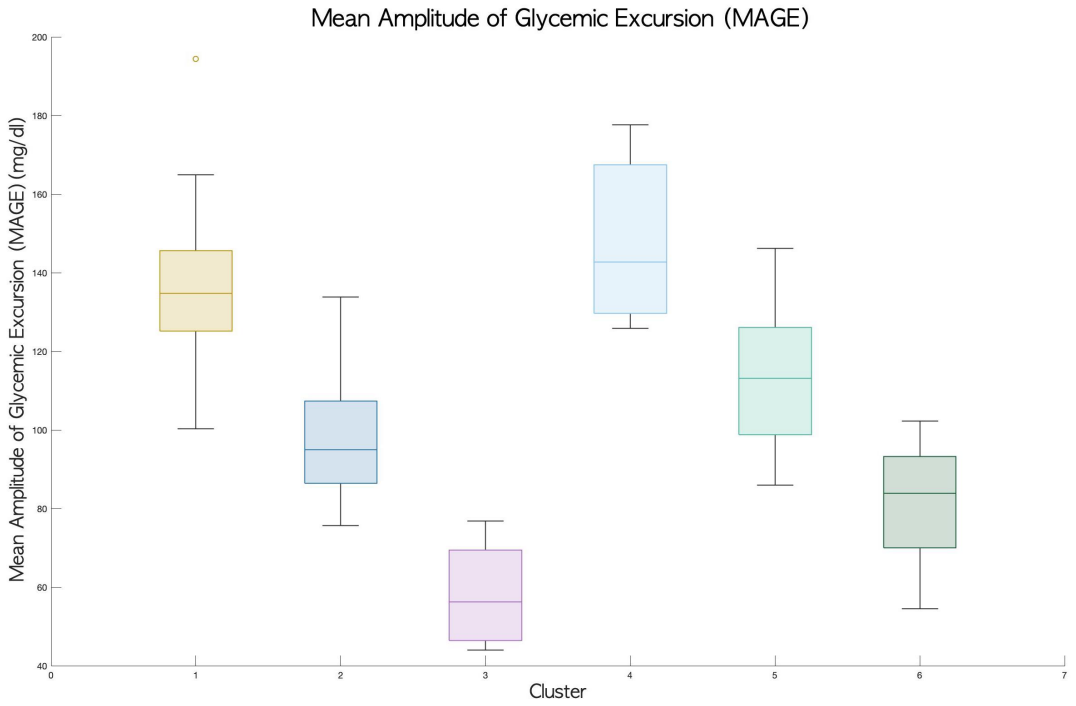


Figura 5.3: Boxplots valori MAGE dei clusters

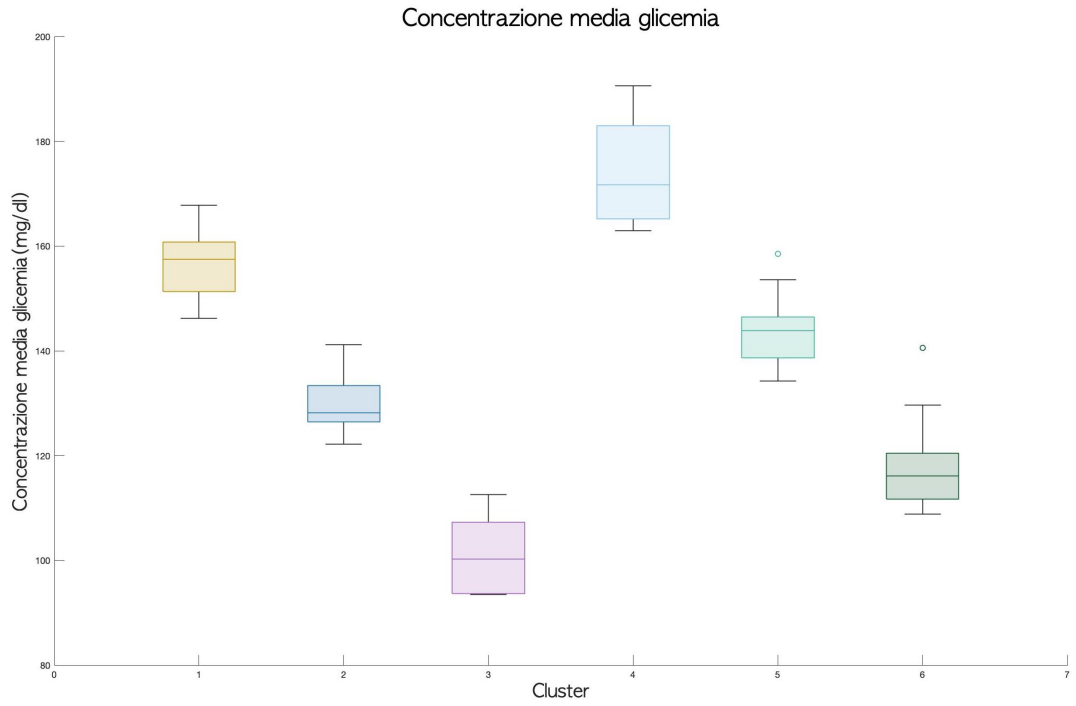


Figura 5.4: Boxplots valori medi glicemia settimanale dei clusters

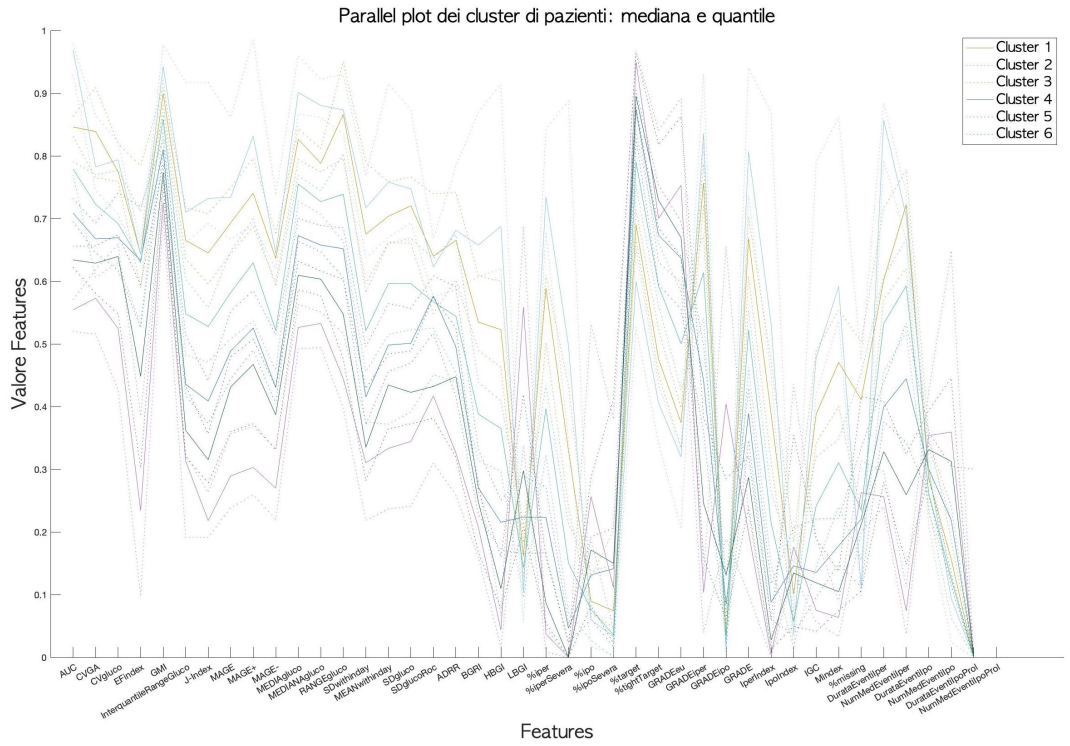


Figura 5.5: Parallel coordinates plot dei cluster di pazienti

5.1.2 Analisi ed interpretazione dei clusters ottenuti

Come precedentemente ribadito, utilizzando tabella riportata in figura 5.6, è possibile confrontare valore medio e deviazione standard delle varie features dei cluster formatisi grazie alla procedura di stratificazione. Nell'effettuare analisi ed interpretazione dei cluster, si sono confrontati i valori delle features non solo tra loro, ma anche con valori e range di riferimento presenti in letteratura. Verranno quindi di seguito elencate le caratteristiche di ognuno dei sei cluster e ciò che è emerso da suddetto confronto, cominciando in particolare da quei cluster che presentino valori più "estremi", per passare poi all'analisi dei rimanenti:

- Cluster 3: è composto solamente da 5 pazienti che presentano un valor medio di area sotto la curva più basso rispetto a tutti gli altri cluster (addirittura di un ordine di grandezza inferiore); come descritto nel capitolo 3, un valore basso in questa features indica dei profili glicemici a bassa variabilità. Infatti, anche tutti gli altri indici di variabilità del cluster sono tra i più bassi, come si può notare dai valor medi della deviazione standard della glicemia e della "glucose rate of change" o del numero di escursioni glicemiche maggiori di 75 mg/dl (Excursion Frequency index). Nello spe-

Features (media ± sd)	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5	CLUSTER 6
Numero di soggetti	21	27	5	7	28	19
Auc (Area Under Glucose)	1.4966e+06 (3.7918e+04)	1.2578e+06 (3.3773e+04)	9.5966e+05 (5.9330e+04)	1.6908e+06 (5.4577e+04)	1.3766e+06 (2.9382e+04)	1.1241e+06 (3.9438e+04)
CVGA	5.9290e+03 (906.4617)	4.8008e+03 (544.5644)	4.0520e+03 (489.1486)	5.8022e+03 (798.6262)	5.0421e+03 (898.7896)	4.4696e+03 (473.2263)
CV concentrazione glucosio	37.1109 (3.8724)	31.7848 (3.7358)	25.5428 (6.1557)	36.5556 (3.8644)	33.5495 (4.3237)	29.0875 (4.8101)
Excursion Frequency Index	5.0455 (1.0668)	4.4365 (1.0372)	1.8000 (1.1185)	4.9490 (0.7293)	4.5787 (0.9755)	3.3195 (1.2613)
Glucose Management Indicator	7.0626 (0.1395)	6.4090 (0.1160)	5.7287 (0.1941)	7.4741 (0.2533)	6.7488 (0.1438)	6.1204 (0.1877)
Range interquartile glicemia	78.4425 (11.5400)	52.2898 (8.6385)	31 (8.2462)	89.0613 (15.5608)	62.4540 (10.2961)	42.9719. (9.5869)
J index	46.6938 (5.6168)	29.5507 (3.7329)	16.1126 (2.9596)	57.0458 (9.4780)	37.3425 (4.6998)	23.2541 (3.8811)
MAGE	136.9606 (20.1668)	97.1728 (13.4801)	58.2165 (13.6917)	148.7345 (20.9409)	114.3909 (18.0935)	81.1786 (14.2431)
MAGE + Index	137.0964 (18.3493)	98.0659 (13.1248)	57.5459 (13.1805)	153.0201 (27.5510)	116.0726 (18.0335)	81.8823 (15.0254)
MAGE - Index	136.7345 (22.7279)	96.8217 (14.4518)	58.8901 (13.8472)	144.4090 (16.8418)	112.4763 (18.8419)	80.2390 (13.7037)
Concentrazione glucosio media	156.8816 (5.8329)	129.5562 (4.8511)	101.1142 (8.1166)	174.0852 (10.5893)	143.7620 (6.0109)	117.4921 (7.8472)
Concentrazione glucosio mediana	146.0476 (4.9998)	122.4722 (4.8359)	97.5000 (8.9022)	164.6429 (10.5068)	133.4821 (5.0580)	111.4211 (7.2443)
Range valori glucosio	308.4881 (35.4257)	236.2037 (27.7742)	163.2000 (23.1479)	317.7143 (25.9934)	267.8393 (38.5974)	197.8158 (33.2500)
Sd dei within-day means index	21.5141 (4.5417)	13.4038 (2.5591)	9.6494 (3.3076)	22.9111 (3.9723)	16.6947 (3.1653)	11.2361 (3.5461)
Media of within day Sd index	52.5348 (7.2365)	37.6961 (5.3272)	22.9776 (5.6638)	57.8226 (9.3534)	44.1058 (6.6582)	31.4220 (5.7994)
Deviazione standard glucosio	58.1064 (7.5272)	41.2544 (5.8628)	25.7399 (6.1648)	63.5660 (9.6657)	48.6808 (7.4319)	34.4567 (6.4471)
Sd rate of change glucosio	1.0664 (0.2149)	0.9036 (0.1381)	0.6241 (0.1593)	1.0582 (0.2330)	0.9244 (0.1591)	0.7478 (0.1681)
Average Daily Risk Range	43.2792 (8.6783)	31.6768 (7.1745)	21.6788 (6.9082)	44.6845 (9.5611)	34.9453 (8.6764)	27.3898 (8.1885)
Blood Glucose Risk Index	7.4387 (1.2393)	3.8654 (0.9705)	3.0877 (1.3075)	9.9835 (2.2306)	5.2744 (1.1609)	3.0773 (0.9807)
High Blood Glucose Index	6.6501 (1.2088)	2.7606 (0.7734)	0.5156 (0.3262)	9.4872 (2.0858)	4.5258 (1.0379)	1.5785 (0.7742)
Low Blood Glucose Index	0.7105 (0.2987)	1.0409 (0.4190)	2.4938 (1.2015)	0.4331 (0.2276)	0.6907 (0.3358)	1.4308 (0.6189)
% tempo range iper	29.8046 (3.8451)	11.6289 (3.6971)	1.5571 (1.1209)	40.5813 (6.5915)	20.1764 (4.8051)	5.9933 (4.2539)
% tempo range iper severa	7.6319 (3.3285)	1.1266 (0.9877)	0 (0)	13.2092 (6.4070)	3.6407 (2.1933)	0.2518 (0.4097)
% tempo range ipo	2.3794 (1.4454)	3.1959 (1.9511)	8.9070 (8.3325)	1.2208 (0.9124)	2.0264 (1.6451)	4.5650 (3.0600)
% tempo range ipo severa	0.4046 (0.4580)	0.5016 (0.4964)	0.8024 (1.2371)	0.1820 (0.2730)	0.3187 (0.4450)	0.7290 (0.7390)
% tempo in target	67.6448 (4.2923)	84.6581 (4.8348)	89.7342 (8.1823)	57.6245 (6.9553)	77.4334 (5.5747)	88.8972 (5.6269)
% tempo range tight target	36.1164 (3.7276)	51.1689 (7.0839)	54.3810 (10.6365)	29.3374 (3.9842)	45.7614 (5.6100)	57.2066 (9.3222)
GRADEeu	33.9041 (6.3900)	57.7857 (8.7018)	67.9183 (17.9180)	25.8336 (6.8009)	45.5999 (9.0390)	67.0317 (13.7832)
GRADEHyper	63.4421 (6.0217)	36.5777 (7.4956)	7.8388 (4.9891)	72.9500 (4.4680)	51.3440 (8.1178)	22.3385 (11.3581)
GRADEHypo	2.1167 (1.6093)	4.6591 (3.2374)	22.7121 (17.4074)	0.8925 (0.7533)	2.3159 (2.1845)	9.4345 (7.6172)
GRADE	8.1647 (0.7988)	4.7811 (0.7262)	2.1277 (0.8418)	10.2467 (1.4254)	6.3660 (0.8693)	3.5403 (0.9846)
Hyperglycemic Index	0.7861 (0.2541)	0.1938 (0.1071)	0.0137 (0.0139)	1.2328 (0.4698)	0.4261 (0.1801)	0.0765 (0.0636)
Hypoglycemic Index	0.1151 (0.1109)	0.1387 (0.1129)	0.2812 (0.3361)	0.0543 (0.0678)	0.0927 (0.1048)	0.2005 (0.1858)
Index of Glycemic Control	0.9123 (0.2932)	0.3473 (0.1902)	0.2958 (0.3316)	1.3031 (0.5073)	0.5529 (0.2536)	0.2897 (0.2190)
M value	18.3396 (4.1724)	7.1626 (2.3659)	2.4007 (1.3124)	26.5699 (7.4675)	11.7922 (3.2490)	4.3868 (2.0008)
% campioni missing	4.6875 (2.7792)	2.9321 (2.2123)	3.1276 (1.0251)	3.0435 (3.3285)	3.8431 (2.8607)	3.3130 (2.6651)
Durata media eventi iper	147.9313 (21.5896)	94.9860 (13.0563)	61.5833 (11.2515)	199.8124 (17.7176)	124.5967 (20.8461)	85.5982 (40.7762)
Num mediano eventi iper sett	19.2238 (3.5989)	11.9621 (3.5390)	2.1430 (1.2658)	19.1524 (1.2156)	15.7104 (3.1016)	6.1694 (3.0982)
Durata media eventi ipo	41.4821 (22.4366)	37.3504 (8.3259)	48.2375 (10.9808)	33.3690 (8.5063)	37.4362 (17.6560)	41.2174 (8.2364)
Num mediano eventi ipo sett	5.7409 (3.2789)	7.0644 (3.8534)	14.5586 (10.9739)	2.5727 (1.9033)	4.7524 (3.4698)	9.6379 (5.4408)
Durata media eventi ipo prol	145.8333 (52.3171)	140.3571 (63.3152)	0 (0)	300.0000 (113.3893)	151.2500 (54.9799)	140.0000 (58.8700)
Num mediano eventi ipo prol sett	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Figura 5.6: Tabella con valori medi e deviazione standard delle features in ogni cluster (sono evidenziati, per ogni feature, il valore più basso in verde e più alto in rosso)

cifico, il coefficiente di variazione della glicemia è inferiore alla soglia del 36% (indicata in [10])) consigliata al fine di contenere la variabilità del segnale; questa caratteristica è a maggior ragione dimostrata dal valor medio della "Mean Amplitude of Glycemic Excursion", particolarmente basso ed inferiore ai 60 mg/dl; secondo [35], un valore di MAGE inferiore a questa soglia infatti risulta essere tipico addirittura di soggetti sani. Valor medio e mediano della traccia CGM di questi pazienti risultano essere bassi, quello medio in particolare di poco superiore ai 100 mg/dl: il controllo glicemico di questo cluster di pazienti sembra essere quindi buono. Anche il valore di "M index" è buono ed inferiore a 18 (ancora, secondo [39], il controllo glicemico è considerato efficace). Dal valore di BGRI però si può notare che, nonostante HBGI sia basso (indicante appunto la bassa variabilità del segnale), si ha tra i più elevati valori di LBGI; come riportato in [36], un elevato valore di questo parametro indica un elevato rischio di ipoglicemie. Infatti questo rischio trova riscontro nelle percentuali di tempo speso dalla glicemia nei diversi range: la percentuale di tempo in target è la più alta tra tutti i cluster (maggiore della soglia minima del 70% indicata in [10]), il tempo in iperglicemia basso (minore del valore soglia massima del 25% [10]) ma il tempo in ipoglicemia è molto elevato, quasi due volte superiore rispetto al massimo valore ideale del 4% [10]: ciò si riflette anche nel numero mediano di eventi ipoglicemici settimanali, che è il più alto tra tutti i cluster. È possibile quindi interpretare questo cluster come il sottogruppo di pazienti nei quali la terapia permette di avere un buon controllo della variabilità glicemica, che risulta essere però probabilmente troppo "aggressivo", portando di conseguenza all'aumento del tempo in ipoglicemia e del numero di eventi ipoglicemici;

- Cluster 4: questo insieme sembra raggruppare i pazienti (7 in totale) sottoposti ad una terapia non particolarmente efficace nel ridurre la variabilità del segnale glicemico; infatti, il coefficiente di variazione si attesta attorno al 36%, il numero di escursioni glicemiche superiori ai 75 mg/dl è quasi tre volte tanto quello del cluster 3, valor medio e mediano più elevati di tutti i cluster e quasi all'interno della soglia iperglicemica, deviazione standard dei campioni di glicemia tre volte tanto del valore più basso tra i cluster; il MAGE inoltre è anch'esso molto alto ed all'interno della soglia dei soggetti diabetici instabili [35], quindi con un controllo della variabilità glicemica molto basso. Questa caratteristica è confermata anche dall'alto valore di

HBGI (circa 9.49). Anche il valore di ADRR è elevato e maggiore della soglia massima di 40, definendo quindi questi pazienti come pazienti ad elevato rischio [37]. Per quanto riguarda i valori invece delle percentuali di tempo speso nei possibili range glicemici, emerge che le percentuali riguardanti la fascia iperglicemica (ed anche di iperglicemia severa) sono oltre le soglie massime consigliate (la prima addirittura del 40%) e il numero mediano di eventi di iperglicemia settimanali pari a 19 (valore più alto tra tutti i cluster); il tempo in target medio è pari circa al 57%, molto inferiore alla soglia consigliata dall'International Consensus [10]. Dall'altro lato però il tempo trascorso dalla glicemia nella soglia ipoglicemica è il più basso dei cluster (e inferiore al 4%), come anche il numero mediano di eventi di ipoglicemia settimanali. Il valore dell'indice "M" risulta essere il più altro tra tutti i cluster, ma comunque nel range dei valori associati ad un controllo discreto [39]. Si può concludere che questo cluster rappresenti i pazienti con uno scarso controllo e forti escursioni glicemiche, una elevata percentuale di tempo trascorso in iperglicemia, di conseguenza un numero basso di eventi di ipoglicemia;

- Cluster 1: possiede il coefficiente di variazione e l' "excursion frequency index" più elevati di tutti i cluster; caratterizzato quindi una variabilità elevata, testimoniata anche dal valore più alto nella deviazione standard della "glucose Rate Of Change", ma un valore di MAGE in media inferiore a quello del cluster 4. Anche questo cluster, formato da 21 pazienti, ha un valore di ADRR maggiore di 40 (pazienti ad alto rischio) e un valore di HBGI elevato. Ha quindi una variabilità simile a quella (elevata) del cluster 4, ma migliora in alcuni parametri; il numero mediano di eventi di iperglicemia è lo stesso del cluster 4, ma la durata media degli eventi iperglicemici inferiore. Inoltre, il numero di eventi invece di ipoglicemia è decisamente inferiore a quello del cluster 3. Tutto questo si riflette in una percentuale di tempo in iperglicemia di poco superiore rispetto al valore soglia (circa del 30%), una percentuale di tempo speso in range di ipoglicemia inferiore alla soglia massima consigliata del 4%, ma una percentuale del tempo in target non sufficiente per garantire il superamento del valore soglia del 70%. Questo cluster quindi presenta valori vicini al cluster 4 (estremo caratterizzato da variabilità e iperglicemia elevate), riuscendo a migliorarne alcuni parametri ma non abbastanza da rientrare nei range stabiliti in [10];

- Cluster 2: formato da 27 pazienti, questo cluster si avvicina in media alle caratteristiche del cluster 1, migliorando ulteriormente alcune metriche: la maggior parte di queste risultano infatti avere dei valori intermedi tra i due cluster estremi (3 e 4); il coefficiente di variazione della concentrazione di glucosio ed il numero medio di escursioni glicemiche maggiori di 75 mg/dl rimangono ancora abbastanza elevati (pari rispettivamente a 31.8% e 4.44) ma ad esempio diminuiscono MAGE (97.17 mg/dl) e valor medio della glicemia settimanale (circa 129 mg/dl); questo valore di MAGE in particolare permette di collocare i soggetti appartenenti a questo cluster nella soglia tra i soggetti diabetici stabili ed instabili [35]; HBGI e LBGI presentano entrambi valori modesti, indicando che questo cluster rappresenta soggetti con rischio moderato sia in termini di variabilità glicemica che di ipoglicemia: questa considerazione è confermata anche dal valore di ADRR, pari a 31.67, collocando questi pazienti nel range di rischio moderato (valore di ADRR compreso tra 20 e 40 [37]). Nonostante ciò il valore dell'indice M inferiore a 18 indica un controllo buono, e le metriche di tempo rispettano tutte le indicazioni riportate in [10]: la percentuale di tempo in iperglicemia è minore del 25% (11.63%), in ipoglicemia del 4% (3.67%) ed il tempo in target molto buono (84%). Infine però il numero mediano di eventi ipoglicemici e iperglicemici rimane ancora elevato (rispettivamente 11.96 e 7.06) e non sembra essere sbilanciato nei confronti dell'uno o dell'altro;

- Cluster 5: cluster con il numero di pazienti più elevato, pari a 28. Se partendo dal cluster 4 e procedendo con 1 e 2 le caratteristiche dei cluster sembravano migliorare nel loro complesso, con il cluster 5 ci si avvicina nuovamente al cluster 4; pur rientrando nei range consigliati per alcune metriche, gli indici di variabilità sembrano essere più elevati rispetto al cluster 2: aumentano infatti il coefficiente di variazione, il MAGE (che si pone vicino alla soglia tra soggetti diabetici stabili e non), HBGI, i livelli medi di glicemia settimanali (143 mg/dl), l'indice globale di rischio per questi soggetti (ADRR pari a 34.95 ma ancora comunque minore di 40). Questo leggero peggioramento in termini di variabilità potrebbe essere la causa di un peggioramento della percentuale di tempo in target, che scende a circa il 77%, e di un aumento del tempo in iperglicemia (circa 20.2%), con conseguente aumento anche degli episodi di iperglicemia; il tempo in ipoglicemia invece rimane a valori buoni (circa del 2%, secondo miglior valore tra i clusters) ed anche il numero mediano di eventi di ipoglicemia è

relativamente basso;

- Cluster 6: questo cluster, formato da 19 pazienti, è quello che, tra tutti i cluster, sembra raggiungere in media un compromesso migliore tra grado di variabilità/iperglicemia e ipoglicemia: il coefficiente di variazione della glicemia è pari circa al 29% (secondo valore più basso tra i clusters), MAGE all'interno della soglia dei pazienti stabili (81.88 mg/dl), livelli medi glicemia settimanale attorno ai 117 mg/dl, valore più basso in assoluto tra i clusters di Blood Glucose Index; quest'ultima metrica in particolare, indica variabilità bassa (HBGI pari a 1.58) pur non elevando il rischio di ipoglicemia: LBGI infatti si attesta a valori attorno a 1.43 (valore dimezzato rispetto al cluster 3, estremo con elevata ipoglicemia). Per quanto riguarda le metriche temporali, il tempo in target è molto buono (circa 88.9%), il tempo in iperglicemia rispetta la soglia del 25%, il tempo in ipoglicemia invece si trova sul valore soglia massimo consigliato (4.57%) ma è comunque inferiore a quello del cluster 3. Il numero di eventi di iperglicemia sono contenuti, leggermente elevati quelli di ipoglicemia ma comunque inferiori al cluster 3;

La procedura di clustering ha quindi permesso di ottenere dei sottogruppi di pazienti con valori di features e caratteristiche che, per lo meno in media e per il livello qualitativo della precedentemente riportata analisi, sembrano essere diverse tra loro e rappresentare quindi diverse possibili casistiche di pazienti o esiti della terapia; sono stati infatti individuati cluster di pazienti con bassa variabilità ma alto tempo in ipoglicemia, o ridotta variabilità ma basso tempo in target, o con features indicanti un buon compromesso tra l'esigenza di mantenere bassa la variabilità del segnale e di incorrere in eventi di ipo- o iper-glicemia. Tutte queste considerazioni potrebbero quindi essere potenzialmente un ulteriore ausilio per la comprensione dei diversi tipi di conseguenze di determinate scelte terapeutiche o potrebbero essere indicanti della conferma dell'esistenza di pazienti con caratteristiche e/o sensibilità individuali e specifiche nei confronti della terapia stessa. Ulteriori analisi quindi sarebbero necessarie per la conferma e la validità statistica delle precedenti affermazioni, ma il risultato sembra andare nella direzione sperata per il raggiungimento di uno degli obbiettivi che si propone il presente lavoro di tesi.

5.2 Risultati clustering dei profili settimanali

Nel paragrafo seguente, verranno invece riportati i risultati ottenuti sul dataset dei profili glicemici settimanali, ottenuto grazie alla procedura descritta precedentemente al paragrafo 3.2.2. L'input dell'algoritmo di clustering, in questo caso di tipo gerarchico agglomerativo, è quindi costituito da una matrice (4673,42); la matrice di prossimità è stata inizializzata con distanza di tipo euclidea ed aggiornata successivamente durante il processo di linkage con il criterio della "Ward distance", al fine di dare priorità alla minimizzazione della varianza intracluster all'aumentare del numero di iterazioni dell'algoritmo. L'albero gerarchico (dendrogram) ottenuto al completamento della procedura è quello riportato in figura 5.7: dalla sua ispezione, e al fine di ottenere un numero di cluster che permettessero di mantenere bassa la complessità dell'analisi (che come nel caso precedente dovrà essere al momento solamente di tipo qualitativo) e di avere un numero di clusters confrontabile con il caso precedente esposto in 5.1. I cluster ottenuti sono quindi 8 e visualizzabili nei punti di intersezione della linea orizzontale in figura 5.7 e il dendrogram stesso.
La procedura di descrizione ed analisi dei clusters è la medesima di 5.1: verranno riportati i boxplot con le distribuzioni delle features nei clusters dei profili glicemici settimanali e ne verranno analizzate ed interpretate le principali proprietà statistiche.

5.2.1 Descrizione clusters settimanali ottenuti

Utilizzando un algoritmo di tipo gerarchico ed osservando il dendrogram ottenuto, sono stati ottenuti 8 clusters di profili glicemici settimanali. Le settimane si sono distribuite nei diversi clusters secondo le percentuali riportate in figura 5.8: i cluster più piccoli risultano essere i cluster 1,2 e 4, mentre la percentuale di settimane rimanenti si distribuisce abbastanza uniformemente negli altri clusters. Vista l'elevata quantità di profili glicemici settimanali, è stato ricavato il "parallel plot" in figura 5.9 al fine di visualizzare in maniera più efficace la stratificazione ottenuta. Soprattutto nelle prime features, la stratificazione è particolarmente evidente, poi diventa invece più confusa e sfumata.
Come nel caso precedente di presentazione dei cluster di pazienti ottenuti, anche qui sono stati ricavati i boxplot con le distribuzioni delle features nei vari cluster di profili settimanali. A differenza però del clustering effettuato sui profili glicemici "globali" dei pazienti, e non settimanali, alcune features non sembrano

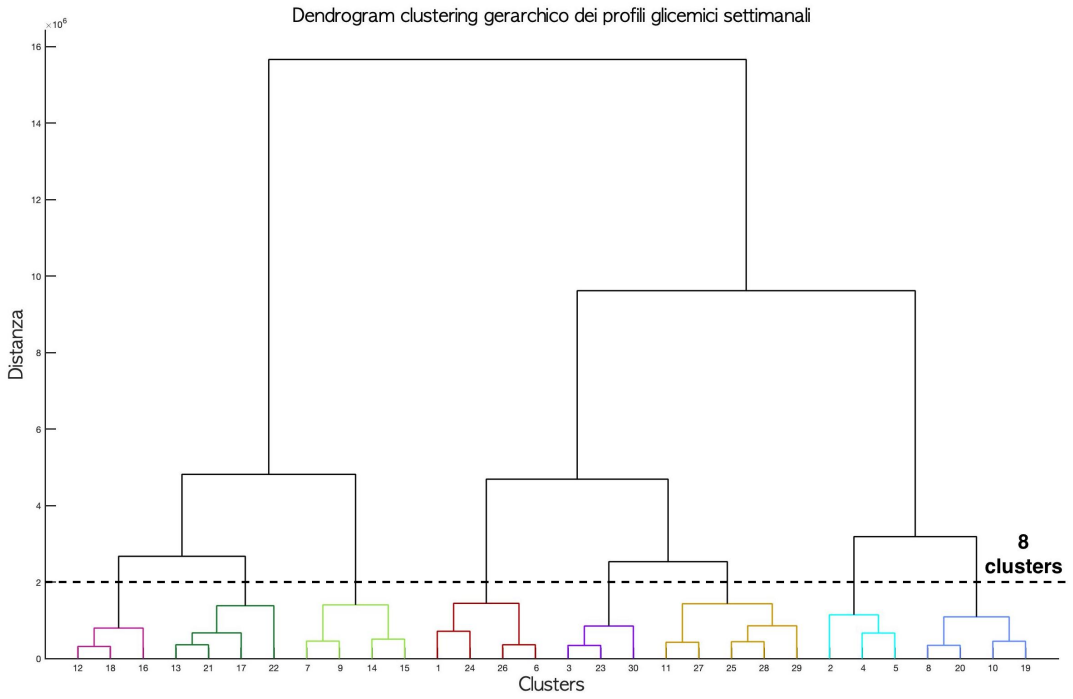


Figura 5.7: Dendrogram dei profili glicemici settimanali con evidenziati in diversi colori gli 8 clusters identificati

presentare delle differenze significative tra loro; ad esempio i coefficienti di variazione glicemica (figura 5.10) hanno valori mediani poco diversi tra loro e le distribuzioni dei valori di questa feature negli 8 cluster sembrano avere porzioni sovrapponibili tra loro; la stessa tendenza la si può ritrovare anche nei boxplot della deviazione standard della "glucose rate of change" (riportata in appendice B) o nella percentuale di tempo in ipoglicemia (figura 5.11) o durata di eventi di ipoglicemia (presente in appendice B); sono però presenti comunque delle distribuzioni di features che sembrano avere, sempre qualitativamente, delle differenze significative: i valori dell'area sotto la curva glicemica ad esempio (figura 5.12), la concentrazione media della glucosio durante la settimana (figura 5.13), o features relative a tempo trascorso in iperglicemia o numero mediano di eventi di iperglicemia settimanali (riportati in appendice B). Al fine però di esaminare maggiormente nel dettaglio queste differenze e ipotizzarne qualitativamente il significato, come nel clustering dei pazienti, è stata ricavata una tabella con valori medi e deviazione standard di ogni feature in ognuno degli 8 cluster ottenuti, consultabile in figura 5.14. Nel paragrafo successivo quindi ne verrà riportata l'analisi e la possibile interpretazione.

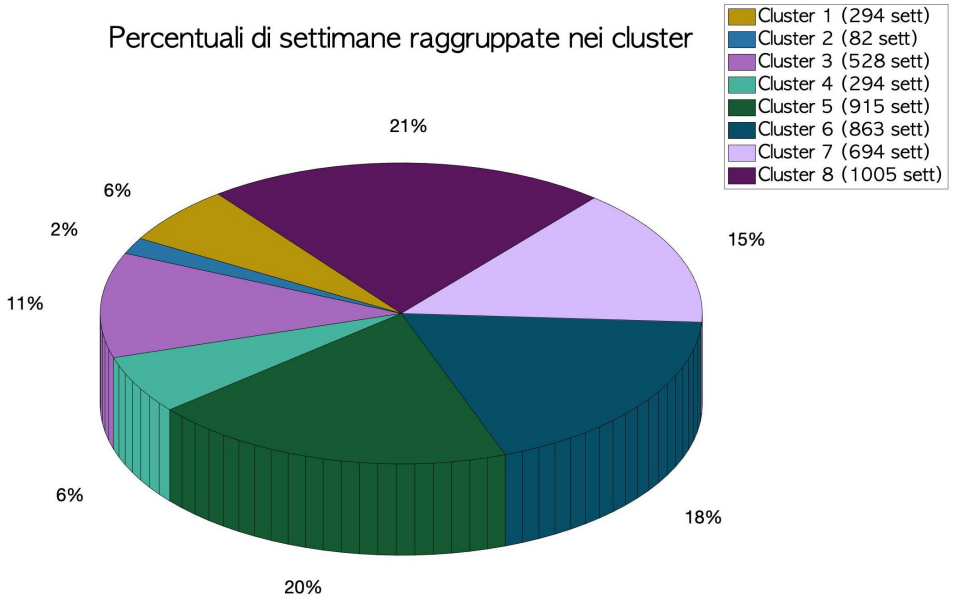


Figura 5.8: Percentuali di settimane sul totale raggruppate in ognuno degli 8 clusters

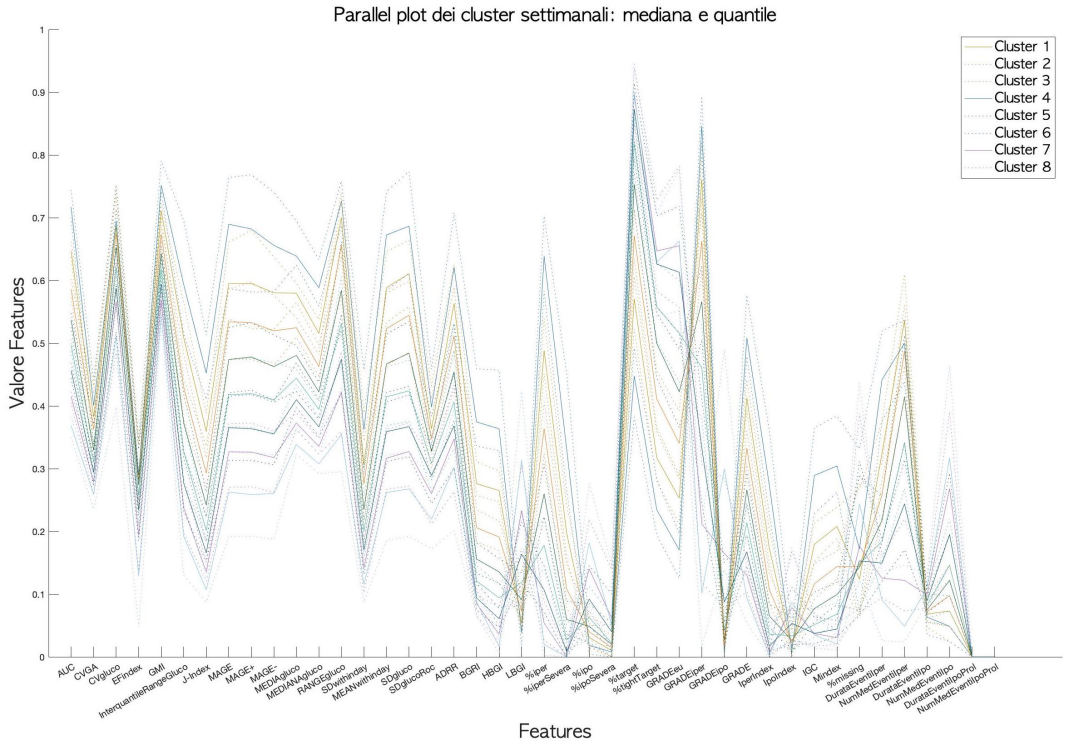


Figura 5.9: Parallel Coordinates Plot dei valori medi delle features nei clusters

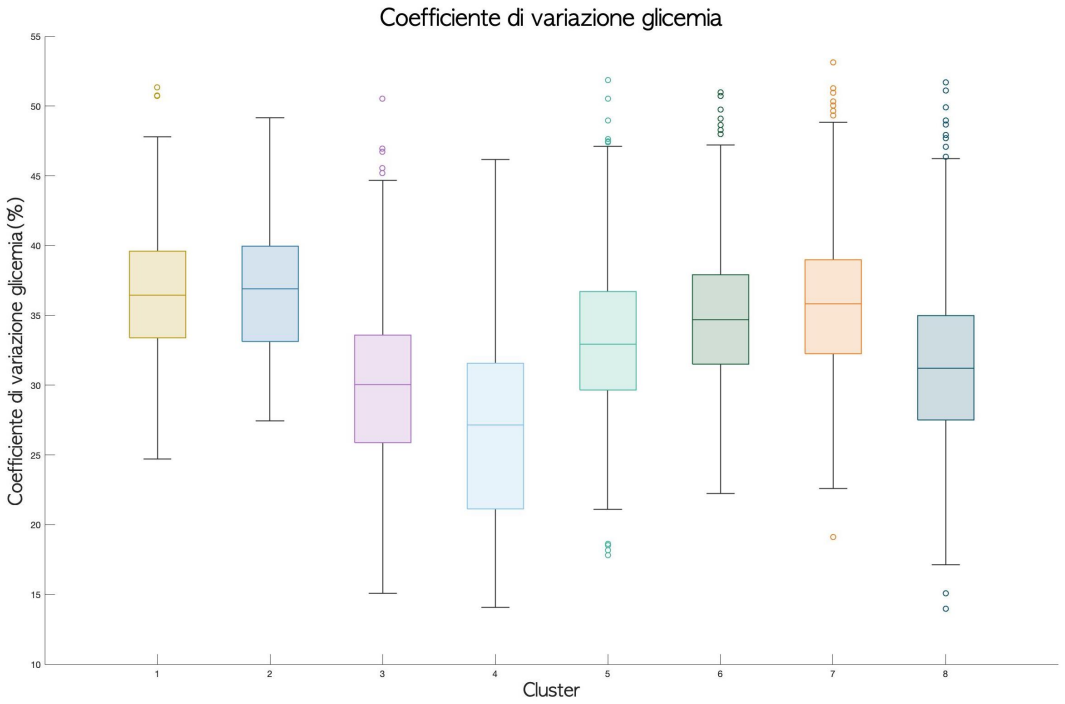


Figura 5.10: Boxplot valori coefficiente di variazione nei clusters settimanali

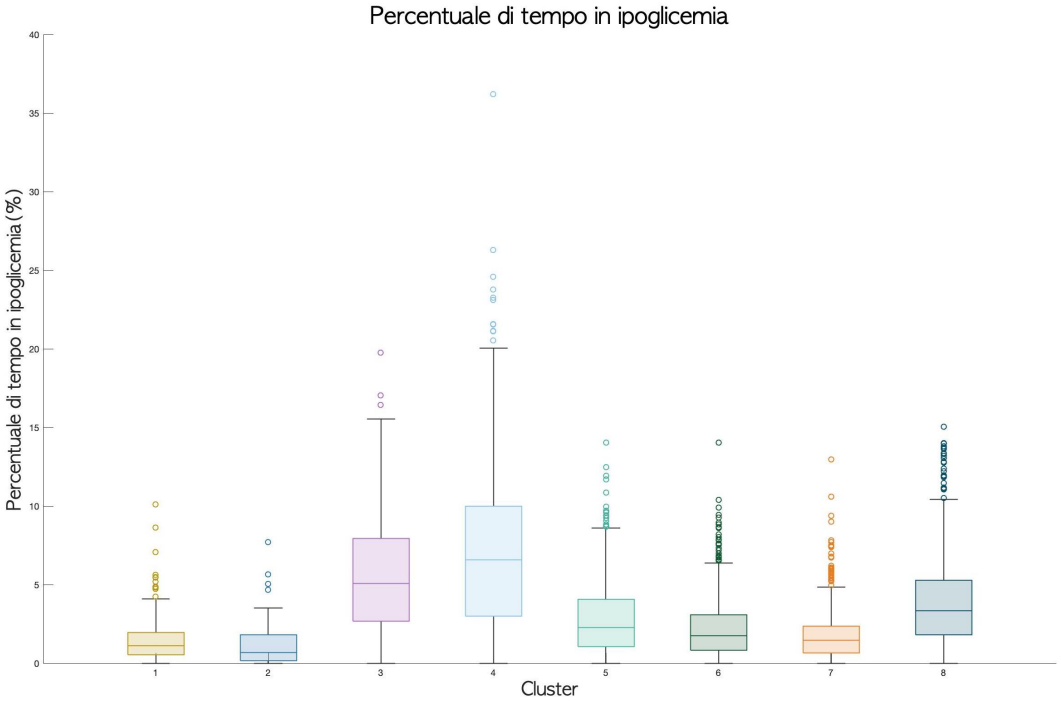


Figura 5.11: Boxplot tempo in ipoglicemia nei clusters settimanali

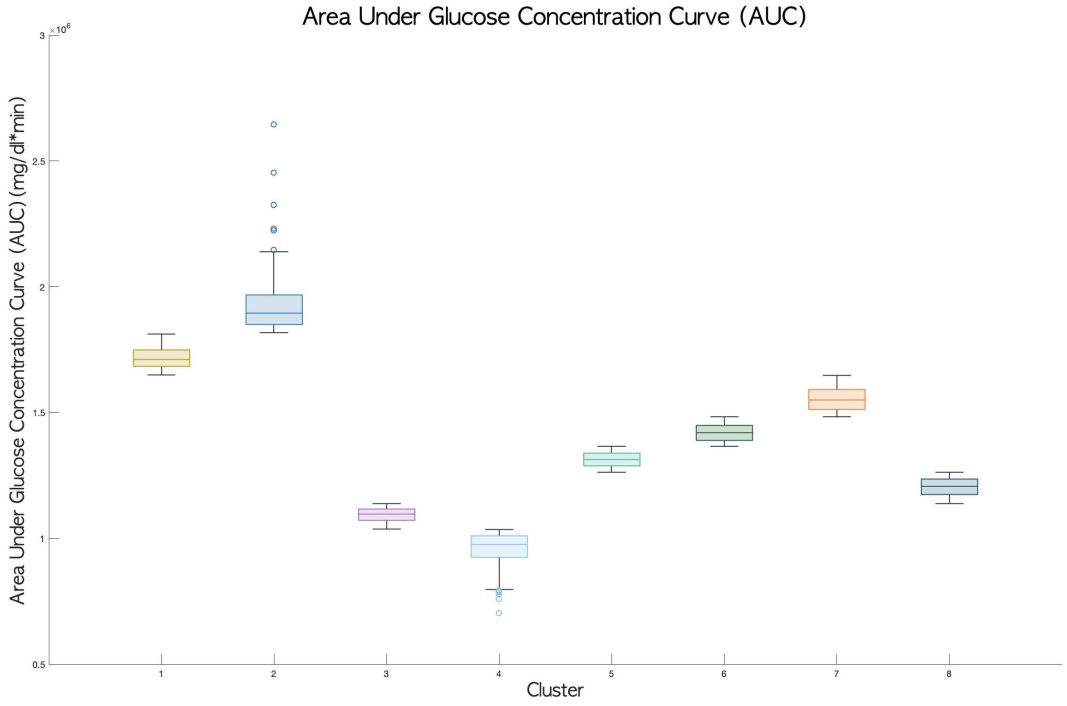


Figura 5.12: Boxplot valori area sotto la curva nei clusters settimanali

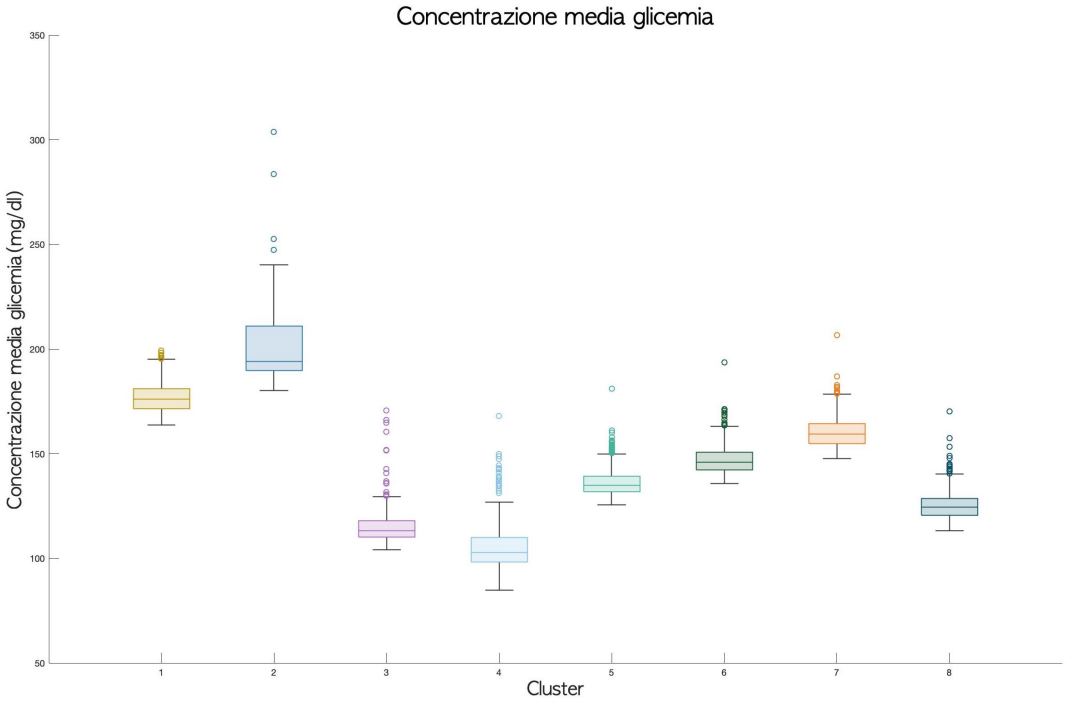


Figura 5.13: Boxplot valore glicemia medio nei clusters settimanali

Features (media ± sd)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Numero di sett	294	82	528	294	915	863	694	1005
Auc (Area Under Glucose)	1717795,3883 (41127,1636)	1941440,9558 (143206,0567)	1093631,4425 (28440,6969)	956096,7159 (67856,7746)	1313769,6416 (29693,2244)	1420701,9545 (33666,9343)	1554209,5197 (45910,8160)	1204587,5669 (35981,3552)
CVGA	6204,5123 (1392,2477)	6292,1334 (1063,9659)	4525,0853 (793,2166)	4174,5423 (749,2903)	5050,8442 (997,8845)	5356,8147 (1146,3131)	5802,9062 (1211,6451)	4809,2425 (915,0368)
CV concentrazione gluco	36,6528 (4,7538)	36,7522 (4,7243)	29,8879 (5,8005)	26,7971 (7,0685)	33,3029 (5,3325)	34,8799 (4,8527)	35,9914 (5,2178)	31,4321 (5,5177)
Excursion Frequency Index	5,4302 (1,4983)	5,5697 (1,8330)	3,4680 (1,4676)	2,4120 (1,6375)	4,5761 (1,2590)	4,9317 (1,3066)	5,2201 (1,4583)	4,1114 (1,3580)
Glucose Management Indicator	7,5415 (0,1679)	8,1444 (0,5051)	6,0573 (0,1888)	5,8362 (0,3086)	6,5654 (0,1506)	6,8250 (0,1587)	7,1406 (0,1676)	6,3037 (0,1577)
Range interquartile	90,5571 (16,6742)	109,5657 (24,5466)	42,3013 (10,3769)	35,3759 (13,6143)	58,0546 (11,7818)	66,8862 (12,1128)	76,7659 (13,9711)	49,3245 (11,0750)
J index	58,6719 (6,9788)	77,3034 (18,3669)	22,5019 (4,6132)	18,5163 (6,4631)	33,1129 (4,7728)	39,4584 (5,0060)	47,6390 (6,0427)	27,2765 (4,4771)
MAGE	152,1350 (23,8485)	172,8047 (28,9649)	81,9715 (18,8184)	67,7671 (24,3098)	107,3869 (20,7890)	120,8338 (19,9309)	135,2155 (22,7235)	93,4726 (19,9800)
MAGE + Index	154,4872 (27,6818)	174,8446 (32,7273)	82,9604 (18,9425)	68,1589 (25,1134)	108,5405 (21,6538)	122,7378 (20,9415)	136,8588 (24,1580)	94,2468 (20,3683)
MAGE - Index	149,7828 (23,0454)	170,7648 (28,6796)	80,9825 (19,4077)	67,3752 (24,0670)	106,2334 (21,3739)	118,9298 (20,8971)	133,5723 (23,6619)	92,6983 (20,4756)
Glicemia media	176,9026 (7,0173)	202,1077 (21,1155)	114,8541 (7,8920)	105,6098 (12,9025)	136,0967 (6,2973)	146,9468 (6,6345)	160,1420 (7,0069)	125,1534 (6,5936)
Glicemia mediana	167,1863 (9,2286)	195,0289 (25,7024)	108,8091 (7,7261)	100,8367 (11,6363)	127,6412 (6,8937)	137,2264 (7,9192)	149,9224 (8,5842)	118,0201 (6,4773)
Range glucosio	326,2608 (35,3947)	338,0736 (24,4664)	204,7020 (49,5255)	174,6192 (51,5168)	256,4622 (47,8211)	280,5972 (45,6286)	306,5496 (43,2541)	231,6813 (48,0532)
Sd del within-day means index	24,2510 (9,1773)	29,4542 (12,2210)	11,5585 (5,1810)	9,9927 (5,3811)	15,9203 (6,3461)	18,3521 (6,6583)	21,4664 (8,0700)	13,5578 (5,5098)
Media del within day Sd index	58,8837 (8,8348)	66,8999 (11,0434)	31,4077 (6,9302)	26,1315 (9,1917)	41,3091 (7,3294)	46,5980 (7,1519)	52,1341 (8,3119)	36,0029 (7,2197)
SD glicemia	64,9097 (9,3863)	74,2954 (12,2315)	34,4819 (7,8937)	28,7556 (9,9500)	45,4263 (8,1994)	51,3095 (7,8785)	57,6954 (9,1203)	39,4799 (8,0338)
SD rate of change glicemia	1,1400 (0,2887)	1,2547 (0,3859)	0,7845 (0,2068)	0,6817 (0,2173)	0,9201 (0,1953)	0,9877 (0,2131)	1,0620 (0,2551)	0,8625 (0,1951)
Average Daily Risk Range	46,4501 (9,4427)	51,9384 (10,2731)	28,5462 (9,4133)	25,2118 (9,9438)	33,7963 (9,1743)	37,4998 (9,2839)	41,7988 (9,8270)	30,9358 (9,3182)
Blood Glucose Risk Index	10,2172 (1,6074)	14,9202 (4,5330)	3,1209 (1,1686)	3,2188 (1,2649)	4,4495 (1,2686)	5,7181 (1,2992)	7,5353 (1,4645)	3,5182 (1,2087)
High Blood Glucose Index	9,7817 (1,4709)	14,5565 (4,4283)	1,4456 (0,9183)	0,9328 (1,1344)	3,5551 (1,0391)	5,0151 (1,1243)	6,9807 (1,3225)	2,3163 (0,9263)
Low Blood Glucose Index	0,4356 (0,2965)	0,3637 (0,3421)	1,6753 (0,6755)	2,2861 (1,0372)	0,8944 (0,4590)	0,7030 (0,3840)	0,5546 (0,3456)	1,2019 (0,5469)
% tempo range iper	42,3591 (5,2570)	55,9704 (8,9020)	5,4478 (4,6287)	3,5150 (5,2430)	15,6999 (5,2026)	22,8177 (5,3624)	31,2047 (5,5546)	9,5955 (4,8254)
% tempo range iper severa	14,0416 (4,7483)	25,8341 (10,6043)	0,5230 (1,1744)	0,3778 (1,3547)	2,4036 (2,2409)	4,5181 (2,7401)	7,9643 (3,7430)	1,1206 (1,5120)
% tempo range ipo	1,4710 (1,4041)	1,1961 (1,4260)	5,5686 (3,6638)	7,4246 (5,5470)	2,8475 (2,2891)	2,2027 (1,8451)	1,8015 (1,6170)	3,8973 (2,7961)
% tempo range ipo severa	0,2702 (0,4839)	0,3604 (0,7299)	0,9562 (1,0289)	1,1302 (1,5573)	0,4937 (0,7015)	0,4071 (0,6100)	0,3448 (0,5487)	0,7149 (0,9014)
% tempo range target	56,1699 (5,6960)	42,8334 (9,2152)	88,9836 (6,4099)	89,0603 (7,3301)	81,4526 (6,3867)	74,9795 (6,1517)	66,9938 (6,0575)	86,5072 (6,4175)
% tempo range tight target	27,4602 (4,8442)	20,0551 (5,8745)	56,8061 (9,7332)	53,8053 (10,6877)	48,7499 (7,8670)	42,8555 (6,7300)	35,5225 (5,9402)	54,4786 (8,9976)
GRADEeu	25,0661 (5,3814)	15,9933 (5,3657)	66,8455 (15,5386)	65,7382 (16,7879)	52,3785 (11,8656)	42,9007 (9,1614)	34,4792 (7,7644)	61,5191 (13,7714)
GRADEiper	73,8798 (4,9751)	83,0971 (4,9572)	21,3354 (12,6617)	13,5435 (14,4717)	43,9512 (10,6889)	54,7035 (8,3845)	63,8855 (7,2537)	31,9672 (11,8862)
GRADEipo	1,0541 (1,2894)	0,9096 (1,4995)	11,8191 (8,4900)	20,7183 (14,5443)	3,6703 (3,4913)	2,3958 (2,4438)	1,6353 (1,8813)	6,5237 (5,1842)
GRADE	10,5604 (0,9541)	13,7063 (2,6414)	3,3352 (1,0671)	2,6043 (1,4549)	5,5273 (0,9239)	6,8247 (0,9393)	8,4538 (0,9604)	4,3184 (0,9838)
Hyperglycemic Index	1,2991 (0,3412)	2,2814 (1,0228)	0,0822 (0,1139)	0,0544 (0,1248)	0,2985 (0,1761)	0,5002 (0,2060)	0,8082 (0,2766)	0,1574 (0,1312)
Hypoglycemic Index	0,0724 (0,1110)	0,0970 (0,2070)	0,2580 (0,2419)	0,3130 (0,3662)	0,1329 (0,1614)	0,1074 (0,1400)	0,0908 (0,1301)	0,1890 (0,2073)
Index of Glycemic Control	1,3715 (0,3972)	2,3784 (1,1247)	0,3402 (0,2892)	0,3674 (0,3854)	0,4314 (0,2765)	0,6077 (0,2901)	0,8989 (0,3392)	0,3464 (0,2863)
Mr value	27,4961 (5,4235)	43,9057 (17,0372)	4,2630 (2,6534)	3,3695 (3,0055)	9,2166 (3,2554)	13,1796 (3,6113)	18,9129 (4,5302)	6,1460 (2,8166)
% campioni missing	3,5149 (3,1997)	4,4026 (3,9110)	4,6106 (3,8948)	5,7102 (4,0632)	3,9482 (3,3881)	3,8177 (3,2652)	3,5581 (2,9408)	4,1657 (3,6915)
Durata media eventi iper	195,5453 (46,3654)	271,8049 (69,4025)	78,1020 (44,1776)	55,3962 (45,4821)	110,8764 (30,5074)	131,3978 (33,2570)	157,1130 (41,2237)	92,7180 (32,4291)
Num mediano eventi iper	21,9724 (4,7111)	20,5590 (3,8642)	6,3699 (4,2874)	4,0061 (4,6415)	13,9781 (4,1381)	17,4075 (4,4451)	20,1332 (4,6296)	10,0453 (4,2688)
Durata media eventi ipo	30,2657 (17,9809)	32,2030 (33,4366)	42,9717 (15,3011)	47,0348 (17,4736)	36,4121 (18,3954)	34,4156 (22,9917)	33,1489 (23,6380)	39,8265 (16,2072)
Num mediano eventi ipo	3,6906 (3,0025)	2,3670 (2,2150)	11,3259 (6,4328)	13,6462 (8,6421)	6,4616 (4,6081)	5,3107 (3,9240)	4,3378 (3,3852)	8,4699 (5,2807)
Num mediano eventi prol ipo	0,0068 (0,0824)	0,0244 (0,1553)	0,0133 (0,1145)	0,0272 (0,1828)	0,0077 (0,0872)	0,0046 (0,0680)	0,0043 (0,0849)	0,0129 (0,1131)
Durata media prol ipo	160 (28,28)	300 (106,0660)	148,5714 (33,2559)	144.6429 (26.6313)	160,7143 (55,1837)	162,5000 (75,1110)	178,7500 (44,1942)	146,5385 (22,1157)

Figura 5.14: Tabella valori medi e deviazione standard delle features negli 8 clusters (in verde valore più basso ed in rosso valore più alto tra i clusters)

5.2.2 Analisi ed interpretazione dei clusters settimanali ottenuti

Come anticipato nel paragrafo precedente, verrà di seguito analizzata la tabella in figura 5.14, al fine di poter dare un'interpretazione qualitativa e preliminare dei vari cluster di profili glicemici settimanali ottenuti. Anche in questo caso nello specifico, i valori medi delle features sono stati confrontati con le soglie ed i valori consigliati come ottimali e sicuri presenti in letteratura. L'analisi riguarderà prima i due cluster con i valori medi di features più estremi (uno verso l'alto, il cluster 2, l'altro verso il basso, cluster 4), per poi passare invece ai cluster con valori di features intermedi.

- Cluster 2: come anticipato, è il cluster delle settimane che presentano, per la grande maggioranza delle 42 features, i valori medi più elevati ed è formato da un numero di settimane relativamente piccolo (82). Presenta un elevato valore di area sotto la curva, addirittura un ordine di grandezza superiore rispetto al valore delle stessa feature nel cluster "estremo inferiore" (cluster 4); il coefficiente di variazione dei campioni glicemici non si discosta di molto però da quello degli altri cluster, risulta comunque essere superiore al 36% (valore soglia massima consigliata); il range interquantile della curva glicemica è marcatamente più elevato rispetto al valore corrispondente in tutti gli altri cluster, ed il livello medio di glicemia nella settimana è anch'esso decisamente elevato, oltre la soglia minima dell'iperglicemia (risulta essere infatti pari a circa 202 mg/dl). Anche il MAGE è elevato e superiore alla soglia dei 119 mg/dl, rendendolo quindi indicante di profili glicemici settimanali con escursioni glicemiche elevate, tipici di soggetti diabetici instabili. Il valore di ADRR è superiore a 40 (pari a circa 51.94), etichettando queste come settimane a rischio elevato. Il valore di BGRI è anch'esso il più alto tra tutti i cluster, con HBGI elevato indicante alta variabilità glicemica e nonostante LBGI abbia il valore più basso di tutti i cluster; questa tendenza all'elevata variabilità si riflette anche nelle metriche di tempo speso nei vari range glicemici: il tempo in iperglicemia è elevato e superiore alla soglia massima consigliata, il tempo in target basso e decisamente inferiore al 70%, mentre il tempo in ipoglicemia è il più basso in assoluto (circa 1.2%). Inoltre, il numero mediano di eventi iperglicemici, come la loro durata, è tra i più elevati (circa 20.6), mentre il numero di eventi di ipoglicemia risulta essere il più basso.

Da questa analisi preliminare, si può supporre che questo cluster contenga quei profili glicemici settimanali che un elevato tempo in iperglicemia, tanto da avere una media glicemica oltre la soglia dei 180 mg/dl, con un controllo quindi del profilo glicemico e terapia inefficaci.

- Cluster 4: formato da più settimane, 294, ha un ordine di grandezza inferiore nel valor medio di area sotto la curva glicemica rispetto al cluster 2 ed anche un valore di coefficiente di variazione inferiore al 36% (pari a circa il 26.8%); anche MAGE è il più basso, appena sopra la soglia caratteristica di pazienti diabetici stabili. La media glicemica settimanale è buona (circa 105.61 mg/dl), il valore di ADRR inferiore a 40 ma di poco superiore a 20 (indicante pazienti a rischio moderato ma possiamo dire tendente al basso); il valore di BGRI è tra i più bassi, grazie ad un valore di HBGI altrettanto piccolo (ulteriore conferma di profili settimanali con variablità contenuta), nonostante l'alto valore di LBGI (alto rischio di ipoglicemia); similmente al cluster 3 ottenuto nel clustering dei pazienti, anche qui il tempo in iperglicemia è basso, il tempo in target molto buono (circa 89%), ma tempo in ipoglicemia e numero di eventi ipoglicemici decisamente elevati.
  Le settimane che formano questo cluster quindi potrebbero essere quelle settimane in cui il controllo è eccessivo, portando i livelli di glicemia a variare molto poco ma rischiando spesso di attraversare la soglia ipoglicemica.

- Cluster 1: formato da 294 settimane e con caratteristiche che si avvicinano a quelle del cluster 2 ("estremo iperglicemia"); alcune features risultano migliorate, come il diminuito valore di MAGE (da circa 172.8 a 152.14 mg/dl) o del livello medio della glicemia (176 mg/dl), ancora però elevati ed oltre le soglie considerate di stabilità. ADRR è ancora elevato e superiore a 40 (settimane ad alto rischio), come anche il valore di BGRI. Anche il tempo trascorso in iperglicemia è superiore alla soglia massima consigliata (e pari a 42.36%, con il numero di eventi iperglicemici più elevato tra i cluster) ed il tempo in target inferiore alla soglia minima (56.17%); il tempo in ipoglicemia invece rispetta i range consigliati ed anche il numero di eventi ipoglicemici contenuto.

- Cluster 3: possiede caratteristiche più vicine a quelle del cluster 4: la CV è bassa (29.89%), il MAGE all'interno del range tipico dei pazienti diabetici stabili, il livello medio del glucosio relativamente basso (114.85 mg/dl); ADRR leggermente più elevato del cluster 4, ma vista l'elevata variabilità

all'interno dei dati, questa differenza potrebbe essere trascurabile; tempo in target ed in iperglicemia rispettano di gran lunga le soglie consigliate (rispettivamente 88.98% e 5.45%), ancora elevato invece il tempo medio in ipoglicemia (5.57%); il controllo è buono (M value minore di 18) e sono mantenute i parametri "buoni" del cluster 4, ma ancora non è risolto il problema relativo al tempo in ipoglicemia ed all'elevato numero di eventi ipoglicemici.

- Cluster 5-Cluster 6: rispettivamente di 915 e 863 settimane, hanno caratteristiche abbastanza simili tra loro e si avvicinano a quelle del cluster 1, ma risultano ulteriormente migliorate; la maggior parte infatti delle metriche che avevano valore elevato nel cluster 1, che in particolare impattavano negativamente sulla qualità del controllo glicemico e della terapia del paziente, sono diminuite nel loro valor medio (CV, MAGE, media glicemica, ADRR); da notare in particolare che ora le metriche di tempo rientrano nei range consigliati: il tempo in iperglicemia è notevolmente diminuito ed è aumentato quello in target, senza però impattare negativamente la percentuale di tempo in ipoglicemia, che rimane anch'essa all'interno dei range consigliati. In termini però di numero mediano di eventi, sia di iperglicemia che di ipoglicemia, i valori sono ancora elevati, ma comunque migliori rispetto ai cluster 1-2 (per quelli iperglicemici) e ai cluster 3-4 (per quelli invece ipoglicemici).

- Cluster 7: contiene 694 settimane, ha caratteristiche intermedie tra il cluster 1 ed il cluster 6; non rappresenta quindi settimane con una qualità del controllo e della terapia particolarmente buona (basti guardare ad esempio il valore medio del glucosio, elevato e quasi alla soglia dell'iperglicemia); il tempo in target quindi si abbassa rispetto al cluster 6, aumenta il tempo in iperglicemia e quindi gli eventi iperglicemici; contenuto però il tempo in ipoglicemia e relativi eventi ipoglicemici.

- Cluster 8: contiene più settimane in assoluto (1005) e potremmo dire che è quello che presenta il compromesso migliore tra le caratteristiche dei due cluster estremi 2-4; i valori rientrano quasi tutti nei range consigliati (metriche di tempo e variabilità tra tutte), la media glicemica non è particolarmente elevata ed il controllo buono (indice M inferiore a 18); unica nota negativa, ma non eccessivamente, il numero di eventi di ipo- e iper-glicemia è ancora non trascurabile (livelli mediani rispettivamente pari a circa 10 e

8) e probabilmente per questo ADRR si ritrova nel range indicante profili glicemici a rischio moderato; rispetto agli altri cluster però non si nota uno sbilanciamento verso l'iper- o l'ipo-glicemia, nemmeno nel numero mediano di eventi settimanali; proprio per questo motivo, rispetto agli altri cluster, è stato giudicato come il cluster in grado di ottenere il trade-off migliore tra i due cluster più estremi.

La procedura di clustering gerarchico ha quindi permesso di ottenere dei sottogruppi settimanali con caratteristiche che, come precedentemente ribadito, in questa prima analisi qualitativa sembrano essere simili tra loro (all'interno dello stesso sottogruppo) e diverse dai profili glicemici appartenenti a cluster diversi. Rispetto al clustering dei pazienti, queste differenze tra cluster in alcune feature sono risultate meno evidenti o potrebbero essere considerate trascurabili, vista l'alta varianza dei dati, mentre in altre sembrano essere maggiormente evidenti e significative. Anche i risultati di questa seconda procedura di stratificazione, avvenuta utilizzando solamente dati e features estratte dal segnale di monitoraggio in continua della glicemia, sembrano quindi andare nella direzione sperata per il raggiungimento degli obbiettivi del presente lavoro di tesi. Ulteriori analisi e test di validazione delle precedenti affermazioni risultano essere però necessari, come già ribadito anche per la procedura di clustering dei pazienti in 5.1. Nello specifico, possibili miglioramenti e sviluppi futuri verranno approfonditi nei capitoli a seguire, con un focus prima su possibili applicazioni dei risultati ottenuti (6).

Capitolo 6

Due possibili applicazioni delle metodologie di stratificazione sviluppate

In questo capitolo verranno approfondite e proposte alcune possibili applicazioni dei risultati ottenuti precedentemente con la procedura di clustering. Nello specifico, vengono riportati possibili utilizzi della stratificazione dei profili glicemici settimanali, quindi della seconda tipologia di clustering effettuata, attraverso due esempi applicativi (6.1-6.2). Il fondamento logico che ha permesso di elaborarli trova radici nell'algoritmo di clustering gerarchico stesso: esso infatti raggruppa ad ogni iterazione profili CGM settimanali con features simili tra loro; a partire da questa considerazione quindi, una volta ricavati ed interpretati i clusters, si potrebbe pensare di utilizzare quanto ottenuto per verificare l'andamento nel tempo della terapia o delle condizioni del paziente o quante delle sue settimane appartengano ad un cluster con determinate caratteristiche piuttosto che ad un altro. A partire da quest'ultima considerazione inoltre, si potrebbe ipotizzare che se un paziente possiede una percentuale elevata di settimane raggruppate tutte nello stesso cluster (che quindi presentano per la maggior parte del tempo caratteristiche simili tra loro) potrebbe essere un buon candidato per essere maggiormente prevedibile rispetto ad un altro che possiede invece profili settimanali appartenenti, in proporzioni più o meno omogenee, a clusters diversi (che presentano quindi tra loro poche caratteristiche in comune). Da qui nasce quindi l'idea di verificare se effettivamente esistano pazienti maggiormente "precibili" rispetto ad altri, indentificati con la verifica della presenza o meno di un cluster che racchiuda una elevata percentuale di profili glicemici settimanali del paziente stesso (chiama-

to "cluster dominante"). Questa ipotesi, se adeguatamente verificata, potrebbe permettere di migliorare le prestazioni degli algoritmi di predizione utilizzati per il miglioramento della gestione dei boli insulinici; questi algoritmi nacquero per sfruttare l'informazione sulla "storia passata" del segnale glicemico, disponibile grazie all'avvento di sensori CGM, ma ancora ad oggi presentano dei margini di errore non trascurabili: in letteratura ci si è spesso interrogati sul migliorarne le prestazioni, ma ancora poco si è indagato sulla possibilità che non siano gli algoritmi a non avere buone performance ma piuttosto che esistano pazienti che siano più adatti agli stessi.
Al fine di verificare quindi se ci siano caratteristiche ricorrenti nel tempo o pazienti maggiormente precidibili, è stata costruita la tabella riportata in figura 6.1, chiamata "tabella di predicibilità", dove ogni riga corrisponde ad un paziente e riporta, in ogni sua cella, la percentuale di settimane del paziente che sono state raggruppate nel corrispondente cluster.
Nel seguente capitolo verrà quindi prima riportato un esempio di utilizzo della tabella per l'analisi dell'andamento dei cluster nel tempo (6.1), e successivamente per una prima indagine sulla verifica dell'ipotesi secondo cui pazienti con "cluster dominante" siano più predicibili rispetto ad altri (6.2).

6.1 Analisi dell'andamento dei cluster nel tempo

Come anticipato nell'introduzione del capitolo, questo primo paragrafo vuole riportare un possibile esempio di utilizzo dei risultati ottenuti tramite la procedura di stratificazione dei profili glicemici settimanali, i cui risultati sono riportati in 5.2. Successivamente all'interpretazione dei clusters ottenuti, è nata l'ipotesi di poter utilizzare quanto appreso per l'analisi dell'andamento nel tempo della terapia di un paziente, delle sue condizioni o per l'individuazione di problematiche o caratteristiche ricorrenti nel tempo. La grande disponibilità di dati CGM in questo senso potrebbe quindi permettere di eseguire analisi nel lungo termine, piuttosto che nel breve termine, come avviene nei controlli di routine.
Sono stati quindi individuati, attraverso l'ispezione della tabella in figura 6.1, quei pazienti che presentassero un cluster contenente almeno il 40% delle settimane registrate ("cluster dominante"); successivamente è stato ricavato un grafico che riporta nel tempo l'andamento dei cluster settimanali. In figura 6.2 ad esempio, è riportata la cronologia dei cluster per il paziente 36; nello specifico, più del 50% delle settimane del paziente considerato sono state raggruppate, grazie all'algorit-

	Cluster 1 (% sett)	Cluster 2 (% sett)	Cluster 3 (% sett)	Cluster 4 (% sett)	Cluster 5 (% sett)	Cluster 6 (% sett)	Cluster 7 (% sett)	Cluster 8 (% sett)
Paziente 1	4,76	0,00	0,00	0,00	38,10	19,05	4,76	33,33
Paziente 2	10,00	3,33	3,33	3,33	10,00	36,67	33,33	0,00
Paziente 3	0,00	0,00	25,26	23,16	12,63	1,05	0,00	37,89
Paziente 4	0,00	0,00	30,77	7,69	3,08	0,00	0,00	58,46
Paziente 5	0,00	0,00	19,18	2,74	23,29	2,74	2,74	49,32
Paziente 6	0,00	0,00	2,70	2,70	40,54	5,41	2,70	45,95
Paziente 7	5,56	8,33	0,00	0,00	38,89	19,44	19,44	8,33
Paziente 8	19,05	0,00	0,00	0,00	23,81	23,81	28,57	4,76
Paziente 9	0,00	0,00	0,00	0,00	25,00	25,00	50,00	0,00
Paziente 10	0,00	0,00	11,11	2,78	18,06	15,28	0,00	52,78
Paziente 11	0,00	0,00	8,82	1,47	16,18	0,00	1,47	72,06
Paziente 12	0,00	0,00	25,00	50,00	0,00	0,00	0,00	25,00
Paziente 13	0,00	0,00	4,00	10,00	38,00	30,00	16,00	2,00
Paziente 14	0,00	0,00	13,43	0,00	32,84	2,99	0,00	50,75
Paziente 15	1,11	0,00	6,67	0,00	32,22	10,00	2,22	47,78
Paziente 16	0,00	0,00	0,00	0,00	52,17	30,43	13,04	4,35
Paziente 17	0,00	0,00	0,00	0,00	28,00	44,00	8,00	20,00
Paziente 18	0,00	0,00	16,47	7,06	14,12	1,18	0,00	61,18
Paziente 19	12,12	3,03	6,06	0,00	9,09	18,18	36,36	15,15
Paziente 20	2,73	0,00	0,00	0,00	18,18	46,36	29,09	3,64
Paziente 21	0,00	0,00	0,00	0,00	13,95	55,81	27,91	2,33
Paziente 22	19,09	9,09	1,82	1,82	21,82	21,82	18,18	6,36
Paziente 23	0,00	0,00	3,23	0,00	16,13	32,26	25,81	22,58
Paziente 24	5,50	0,92	1,83	0,92	26,61	41,28	20,18	2,75
Paziente 25	0,00	0,00	6,67	3,33	35,00	10,00	0,00	45,00
Paziente 26	0,00	0,00	67,57	18,92	2,70	0,00	0,00	10,81
Paziente 27	0,00	0,00	0,00	2,56	43,59	46,15	5,13	2,56
Paziente 28	0,00	0,00	0,00	0,00	75,00	0,00	12,50	12,50
Paziente 29	20,00	0,00	6,67	0,00	13,33	20,00	33,33	6,67
Paziente 30	0,00	0,00	22,73	4,55	40,91	0,00	4,55	27,27
Paziente 31	0,00	0,00	0,00	0,00	20,00	40,00	40,00	0,00
Paziente 32	27,27	0,00	0,00	0,00	18,18	18,18	36,36	0,00
Paziente 33	0,00	0,00	22,83	1,09	21,74	10,87	0,00	43,48
Paziente 34	10,26	0,00	0,00	0,00	2,56	41,03	43,59	2,56
Paziente 35	0,00	0,00	52,94	22,35	4,71	0,00	0,00	20,00
Paziente 36	0,00	0,00	5,13	1,28	43,59	8,97	0,00	41,03
Paziente 37	0,00	0,00	0,00	0,00	0,00	40,00	60,00	0,00
Paziente 38	0,66	0,00	0,66	0,00	30,46	36,42	23,18	8,61
Paziente 39	5,33	0,00	2,67	0,00	28,67	34,67	16,67	12,00
Paziente 40	0,00	0,00	14,81	3,70	29,63	7,41	3,70	40,74
Paziente 41	0,00	0,00	5,08	1,69	27,12	22,03	8,47	35,59
Paziente 42	43,59	20,51	0,00	0,00	2,56	7,69	25,64	0,00
Paziente 43	0,00	0,00	55,42	37,35	0,00	0,00	0,00	7,23
Paziente 44	0,00	0,00	5,26	2,63	18,42	36,84	5,26	31,58
Paziente 45	0,00	0,00	19,05	0,00	28,57	4,76	0,00	47,62
Paziente 46	0,00	0,00	3,33	3,33	36,67	6,67	0,00	50,00
Paziente 47	1,27	0,00	0,00	0,00	20,25	40,51	30,38	7,59
Paziente 48	40,00	46,67	0,00	0,00	0,00	0,00	13,33	0,00
Paziente 49	5,88	0,00	0,00	0,00	35,29	35,29	17,65	5,88
Paziente 50	25,93	35,19	0,00	0,00	1,85	9,26	22,22	5,56
Paziente 51	0,00	0,00	53,85	7,69	0,00	0,00	0,00	38,46
Paziente 52	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00
Paziente 53	41,67	13,89	2,78	0,00	0,00	2,78	38,89	0,00
Paziente 54	0,00	0,00	0,00	0,00	52,63	28,95	0,00	18,42
Paziente 55	0,00	0,00	0,00	33,33	16,67	0,00	0,00	50,00
Paziente 56	10,00	0,00	0,00	0,00	20,00	30,00	10,00	30,00
Paziente 57	4,55	0,00	0,00	0,00	31,82	22,73	22,73	18,18
Paziente 58	0,00	0,00	5,00	0,00	40,00	25,00	2,50	27,50
Paziente 59	10,29	0,00	0,00	0,00	22,06	27,94	33,82	5,88
Paziente 60	1,56	0,00	0,00	1,56	21,88	50,00	20,31	4,69
Paziente 61	0,00	0,00	68,00	6,00	0,00	0,00	0,00	26,00
Paziente 62	2,44	0,81	21,95	3,25	12,20	11,38	5,69	42,28
Paziente 63	0,00	0,00	0,00	0,00	33,33	22,22	11,11	33,33
Paziente 64	0,00	0,00	38,46	0,00	23,08	0,00	0,00	38,46
Paziente 65	30,05	3,83	0,00	0,55	4,37	19,13	40,44	1,64
Paziente 66	0,00	0,00	51,72	24,14	0,00	0,00	0,00	24,14
Paziente 67	34,48	6,90	1,15	0,00	5,75	9,20	40,23	2,30
Paziente 68	15,15	3,03	0,00	0,00	18,18	33,33	30,30	0,00
Paziente 69	0,00	0,00	0,00	0,00	16,67	16,67	66,67	0,00
Paziente 70	0,00	0,00	4,76	4,76	33,33	23,81	9,52	23,81
Paziente 71	5,77	0,00	0,00	1,92	15,38	30,77	25,00	21,15
Paziente 72	33,75	1,25	0,00	1,25	2,50	11,25	50,00	0,00
Paziente 73	0,00	0,00	7,69	7,69	30,77	0,00	0,00	53,85
Paziente 74	0,00	0,00	1,85	1,85	29,63	42,59	9,26	14,81
Paziente 75	0,00	0,00	9,68	0,00	29,03	9,68	0,00	51,61
Paziente 76	0,00	0,00	0,00	0,00	0,00	0,00	100,00	0,00
Paziente 77	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00
Paziente 78	21,62	0,00	2,70	0,00	8,11	21,62	40,54	5,41
Paziente 79	0,00	0,00	3,33	0,00	33,33	30,00	20,00	13,33
Paziente 80	0,00	0,00	3,23	3,23	38,71	25,81	0,00	29,03
Paziente 81	0,00	0,00	61,54	7,69	0,00	0,00	0,00	30,77
Paziente 82	10,34	10,34	0,00	3,45	6,90	13,79	48,28	6,90
Paziente 83	21,95	0,00	0,00	0,00	7,32	21,95	46,34	2,44
Paziente 84	0,00	0,00	3,23	96,77	0,00	0,00	0,00	0,00
Paziente 85	0,00	0,00	44,00	32,00	0,00	0,00	0,00	24,00
Paziente 86	7,69	0,00	3,85	0,00	15,38	26,92	34,62	11,54
Paziente 87	0,00	0,00	50,00	0,00	0,00	0,00	0,00	50,00
Paziente 88	0,00	0,00	18,18	81,82	0,00	0,00	0,00	0,00
Paziente 89	0,00	0,00	23,53	58,82	0,00	0,00	0,00	17,65
Paziente 90	0,00	0,00	31,25	27,08	14,58	0,00	0,00	27,08
Paziente 91	0,00	0,00	0,00	0,00	12,50	75,00	12,50	0,00
Paziente 92	5,26	0,00	2,63	2,63	28,95	28,95	23,68	7,89
Paziente 93	0,00	0,00	28,81	27,12	10,17	3,39	0,00	30,51
Paziente 94	0,00	0,00	12,50	0,00	37,50	25,00	12,50	12,50
Paziente 95	0,00	0,00	28,57	0,00	14,29	42,86	0,00	14,29
Paziente 96	33,33	28,57	0,00	0,00	0,00	14,29	23,81	0,00
Paziente 97	0,00	0,00	0,00	0,00	37,50	12,50	37,50	12,50
Paziente 98	0,00	0,00	34,33	1,49	16,42	1,49	0,00	46,27
Paziente 99	4,76	0,00	0,00	0,00	52,38	23,81	9,52	9,52
Paziente 100	6,90	0,00	3,45	0,00	27,59	24,14	34,48	3,45
Paziente 101	26,09	0,00	0,00	0,00	34,78	17,39	21,74	0,00
Paziente 102	0,00	0,00	43,21	7,41	7,41	0,00	0,00	41,98
Paziente 103	0,00	0,00	0,00	0,00	48,48	12,12	3,03	36,36
Paziente 104	0,00	0,00	0,00	50,00	0,00	0,00	0,00	50,00
Paziente 105	0,00	0,00	7,69	0,00	30,77	7,69	0,00	53,85
Paziente 106	0,00	0,00	0,98	0,00	33,33	38,24	19,61	7,84
Paziente 107	2,94	0,00	5,88	0,00	23,53	23,53	17,65	26,47

Figura 6.1: Tabella di predicibilità: ogni riga rappresenta un paziente e ogni cella la percentuale di settimane del paziente raggruppate nel corrispondente cluster

mo di clustering, nel cluster numero 3; vi è poi un 22% di settimane accorpate nel cluster 4, un ulteriore 20% nel cluster 8 ed infine le poche restanti nel 5. La maggior parte delle settimane del paziente quindi appartiene al cluster 3, che in 5.2 si è visto avere caratteristiche simili al cluster "estremo dell'ipoglicemia" (cluster 4), del quale però ne ha migliorato alcuni aspetti: il cluster 3 infatti riesce a diminuire di due punti percentuale il tempo speso in ipoglicemia e quindi a diminuire leggermente il numero di eventi ipoglicemici. Questo cluster è anche la condizione iniziale nella quale si trova il paziente (almeno all'interno del dataset di dati qui disponibili), e si può notare come, con l'avanzare delle settimane, spesso il profilo glicemico del paziente si sposti in maniera ricorrente verso il cluster 4; questo potrebbe essere indice di una terapia non particolarmente efficace e forse troppo "aggressiva", in quanto il cluster 4 come visto registra features peggiori del 3 ed è decisamente sbilanciato verso il range ipoglicemico. Questa informazione quindi di un ritorno ricorrente verso condizioni di sbilanciamento verso l'ipoglicemia potrebbe essere sfruttata per migliorare la gestione dei livelli di glicemia del paziente e sarebbe probabilmente difficile da ricavare con i controlli di routine, nei quali solitamente vengono esaminati i dati solamente delle due settimane precedenti: se la visita ad esempio avvenisse alla settimana numero 20, esaminando quindi i dati delle settimane 18-19, sarebbe difficile constatare che ci sono state 5 settimane nelle precedenti in cui questa piccola tendenza verso il range ipoglicemico (tipica appunto del cluster 3) si sia estremizzata. Dall'altro lato invece, verso la fine del monitoraggio, le oscillazioni sembrano spostarsi quasi stabilmente verso il cluster 8, che si è ricavato essere il cluster con il compromesso migliore tra i due individuati come estremi; se c'è stato quindi un aggiustamento di terapia, ad esempio, questo potrebbe essere l'indice di un miglioramento delle condizioni del paziente. In pazienti che non presentano un cluster dominante, questo tipo di analisi risulta di gran lunga più difficile, in quanto le settimane tendono ad avere caratteristiche differenti tra loro e non sarebbe possibile individuare dei "pattern" ricorrenti nel tempo. Risulta importante infine notare che, questo tipo di analisi è stata ricavata utilizzato solamente dati provenienti da sensore CGM e che magari, in un prossimo futuro, potrebbe essere effettuata in maniera automatica direttamente dal sensore stesso finchè risulta in funzione, magari con annessi "suggerimenti comportamentali" per "passare" da un cluster verso l'altro. É importante nuovamente ribadire che la presente è solamente un'analisi preliminare e per ora solamente concettuale, ma che potrebbe evolvere in futuro in strumenti applicabili alla pratica clinica o magari quotidiana del paziente stesso.

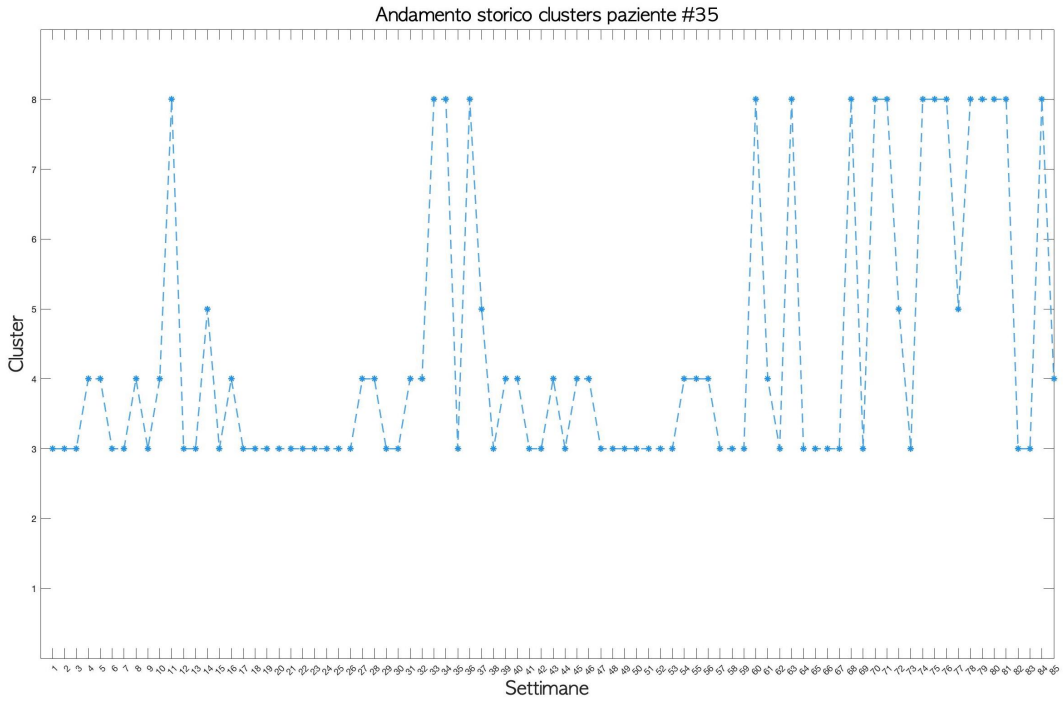


Figura 6.2: Andamento clusters settimanali nel tempo del paziente 35

6.2 Il concetto di paziente predicibile e cluster dominante

In questa sezione verrà invece approfondito e maggiormente indagato il concetto di "paziente predicibile" anticipato nell'introduzione del presente capitolo.
Basandosi sulla definizione e sull'algoritmo di clustering stesso, si è ipotizzato che, se un paziente possiede un numero elevato di settimane che sono state raggruppate nello stesso cluster, il suo profilo glicemico probabilmente presenterà delle caratteristiche ricorrenti nel tempo e simili tra loro; questo paziente quindi potrebbe essere più facilmente prevedibile rispetto ad un altro per il quale non sia stato individuato il già nominato "cluster dominante", e di conseguenza beneficiare maggiormente dall'utilizzo di un algoritmo di predizione. Questi algoritmi nello specifico sono nati in corrispondenza dell'avvento del sensore CGM in quanto, per la prima volta, si poteva disporre dell'andamento (quasi continuo) della curva glicemica del paziente: l'obbiettivo è quello di sfruttarli per predire con un certo anticipo ad esempio l'avvenire di eventi di ipoglicemia o delle ancor più gravi ipoglicemie notturne. Il margine di errore di predizione di questi algoritmi è ad oggi ancora non trascurabile ed è difficile per ora individuare se dovuto all'algoritmo in sè o magari al paziente stesso. L'utilizzo del clustering per indivi-

Gruppo di pazienti	Numero di pazienti
Non predicibili	38
Debolmente predicibili	33
Moderatamente predicibili	23
Altamente predicibili	13

Tabella 6.1: Tabella con numero di pazienti individuato per ogni tipologia di paziente: "non predicibile", "debolmente predicibile", "moderatamente predicibile", "altamente predicibile"

duare pazienti maggiormente predicibili potrebbe essere un valido strumento per migliorare le prestazioni di questi algoritmi e di conseguenza il controllo glicemico del paziente. Ciò che segue quindi vuole essere un esperimento preliminare per verificare l'esistenza o meno di pazienti predicibili, che potrebbe nel futuro svilupparsi in un'analisi maggiormente robusta e comprovata.

6.2.1 Applicazione algoritmo di predizione e analisi risultati ottenuti

Al fine di verificare, come già specificato per lo meno qualitativamente, l'esistenza di pazienti maggiormente predicibili, si è ipotizzato di applicare un algoritmo di predizione ai profili glicemici specifici gruppi di pazienti ricavati dal dataset precedentemente definito in 3.2; sono stati quindi formati gruppi di pazienti "predicibili" e di pazienti "non predicibili" attraverso l'utilizzo della tabella riportata nel paragrafo precedente in figura 6.1. Sono stati scelti nello specifico quattro range percentuali per definire il grado di possibile predicibilità del paziente stesso:

- Pazienti "non predicibili": soggetti che non possiedono alcun cluster contenente almeno il 40% delle settimane del paziente;
- Pazienti "debolmente predicibili" o "lightly predictable": soggetti con almeno un cluster che contenesse 40-49% delle settimane;
- Pazienti "moderatamente predicibili" o "moderately predictable": soggetti con cluster dominante con 50-59% delle settimane
- Pazienti "altamente predicibili" o "highly predictable": soggetti nei quali il cluster dominante contiene più del 60% delle settimane;

In questo modo sono stati ricavati quattro gruppi di pazienti, le cui cardinalità sono riportate in tabella 6.1. Ai dataset appena descritti è stato applicato un

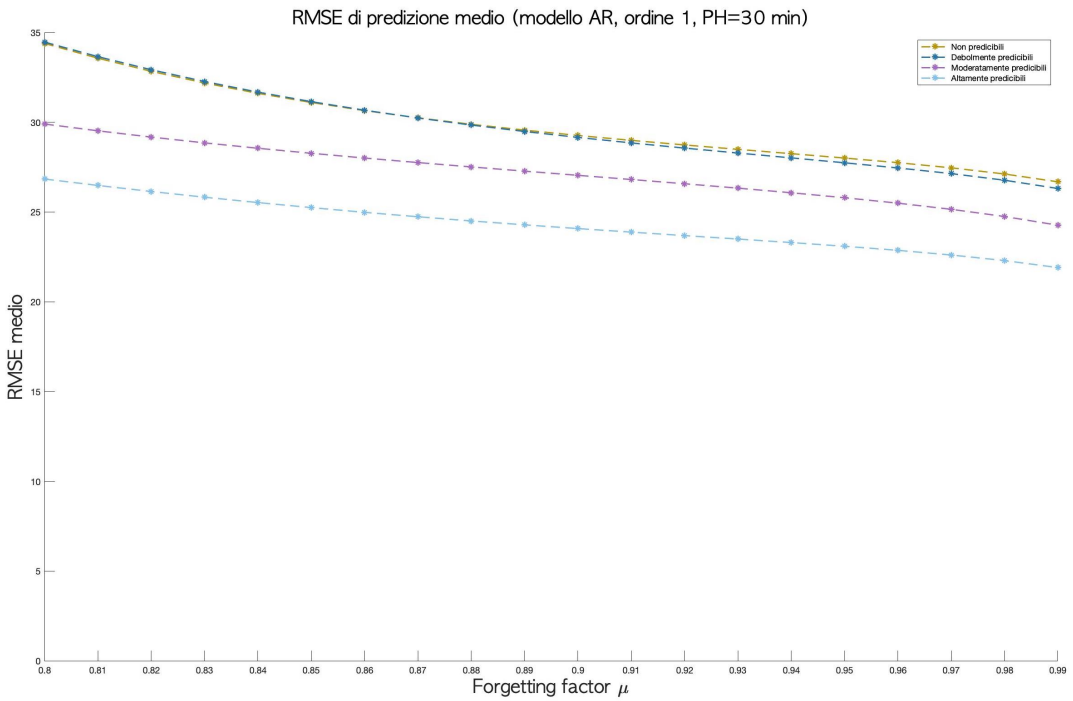


Figura 6.3: Valor medio RMSE nelle quattro classi di pazienti al variare del forgetting factor

algoritmo di predizione di tipo AR ("Auto-Regressive"), di ordine 1 e orizzonte di predizione di 30 minuti, i cui parametri sono stati identificati tramite l'utlizzo della tecnica denominata "Recursive Least Squares". I risultati delle predizioni sono stati valutati in termine di "Root Mean Square Error (RMSE)" e ritardo di predizione al variare del "forgetting factor" μ (parametro che stabilisce la quantità di campioni della storia passata vengono utilizzati dal modello per ricavare la predizione) e sono stati riportati nelle tabelle 6.2-6.3. Inoltre, in figura 6.3, è riportato il valor medio dell'errore di predizione nelle quattro classi di pazienti al variare del forgetting factor. Da una prima analisi qualitativa dei risultati è quindi emerso che:

- in media la predizione sui profili glicemici di pazienti predicibili (debolmente, moderatamente e altamente) ha registrato un valore di RMSE inferiore rispetto a quella effettuata sui pazienti non predicibili; questa differenza si mantiene anche al variare del forgetting factor μ ed è più elevata all'aumentare del "grado di predicibilità"; la differenza di prestazioni è infatti particolarmente marcata tra pazienti non predicibili e moderatamente/altamente predicibili, mentre, tenendo conto anche della varianza non trascurabile dei dati, l'errore di predizione si può considerare pressochè identico tra pazienti non predicibili e debolmente predicibili. Da questi dati preliminari, sembra

emergere che, più la percentuale di settimane raggruppate nello stesso cluster è elevata, migliori risultano essere le prestazioni in termini di errore di predizione;

- il ritardo di predizione invece è simile per tutti i gruppi, anche se in media i pazienti altamente predicibili presentano un valore minore degli altri; in tutti e quattro i gruppi però il ritardo risulta comunque inferiore all'orizzonte di predizione (indice quindi di un buon "guadagno" di predizione in tutti i tipi di paziente esaminati);

Da questa prima analisi, al momento solamente qualitativa, è quindi emerso che i dati raccolti sembrano confermare l'intuizione secondo cui, pazienti con una elevata percentuale di settimane raggruppate nello stesso cluster, possano essere maggiormente predicibili rispetto ad altri. Per validare però l'ipotesi, indagata per ora come anticipato in maniera sommaria e preliminare, sono necessarie ulteriori analisi e test statistici. Si può affermare in ogni caso che i risultati ottenuti sono positivi e sicuramente incoraggianti.
Nel prossimo capitolo quindi verranno riportati alcuni esempi di possibili sviluppi futuri del presente lavoro di tesi, assieme ad alcune conclusioni elaborate invece all'interno del contesto qui riportato.

Forgetting factor	Tipo paziente	RMSE	Ritardo di predizione (min)
μ=0.80	NP	34.39 (11.01)	19.34 (6.80)
	DP	34.46 (11.62)	17.88 (5.00)
	MP	29.90 (7.20)	20 (3.69)
	AP	26.84 (10.29)	16.92 (5.60)
μ=0.81	NP	33.57 (10.28)	19.74 (6.67)
	DP	33.65 (10.55)	18.48 (5.23)
	MP	29.52 (7.04)	20.22 (3.53)
	AP	26.48 (9.77)	17.31 (5.99)
μ=0.82	NP	32.84 (9.67)	20.66 (5.95)
	DP	32.92 (9.67)	18.79 (5.16)
	MP	29.17 (6.90)	20.43 (3.67)
	AP	26.14 (9.30)	17.31 (5.99)
μ=0.83	NP	32.19 (9.17)	20.92 (6.02)
	DP	32.27 (8.94)	19.24 (5.47)
	MP	28.85 (6.79)	20.43 (3.67)
	AP	25.82 (8.87)	18.08 (5.22)
μ=0.84	NP	31.61 (8.75)	21.32 (5.89)
	DP	31.68 (8.35)	20.15 (5.79)
	MP	28.55 (6.69)	21.09 (3.36)
	AP	25.53 (8.50)	18.08 (5.22)
μ=0.85	NP	31.10 (8.41)	21.58 (5.59)
	DP	31.15 (7.89)	20.76 (5.47)
	MP	28.27 (6.60)	21.30 (3.10)
	AP	25.25 (8.16)	19.23 (4.94)
μ=0.86	NP	30.65 (8.13)	21.58 (5.59)
	DP	30.67 (7.51)	21.06 (5.12)
	MP	28.01 (6.52)	21.30 (3.10)
	AP	24.98 (7.87)	19.23 (4.94)
μ=0.87	NP	30.25 (7.90)	21.84 (5.25)
	DP	30.24 (7.22)	21.36 (4.55)
	MP	27.75 (6.45)	22.17 (2.95)
	AP	24.73 (7.62)	19.62 (5.19)
μ=0.88	NP	29.89 (7.70)	21.97 (5.27)
	DP	29.84 (6.98)	21.52 (4.59)
	MP	27.51 (6.38)	22.61 (2.97)
	AP	24.50 (7.40)	20 (4.56)
μ=0.89	NP	29.56 (7.53)	21.97 (5.27)
	DP	29.49 (6.79)	21.52 (4.59)
	MP	27.28 (6.32)	23.70 (3.44)
	AP	24.28 (7.22)	20.38 (4.32)
μ=0.90	NP	29.26 (7.39)	22.63 (5.29)
	DP	29.16 (6.64)	22.27 (4.35)
	MP	27.04 (6.25)	23.91 (3.36)
	AP	24.07 (7.07)	20.77 (4.00)

Tabella 6.2: Tabella con prestazioni algoritmo di predizione al variare del forgetting factor (valori da 0.8 a 0.9) (NP=paziente "non predicibile", DP= paziente "debolmente predicibile", MP= paziente "moderatamente predicibile", AP= paziente "altamente predicibile")

Forgetting factor	Tipo paziente	RMSE	Ritardo di predizione (min)
μ=0.91	NP	28.99 (7.26)	23.55 (4.92)
	DP	28.85 (6.50)	22.56 (4.35)
	MP	26.81 (6.18)	24.56 (2.98)
	AP	23.88 (6.95)	21.92 (3.84)
μ=0.92	NP	28.74 (7.15)	24.34 (4.96)
	DP	28.56 (6.39)	23.03 (4.50)
	MP	26.57 (6.12)	24.78 (2.81)
	AP	23.69 (6.85)	22.69 (3.88)
μ=0.93	NP	28.49 (7.06)	24.47 (4.90)
	DP	28.29 (6.29)	23.48 (4.59)
	MP	26.33 (6.05)	24.78 (2.81)
	AP	23.49 (6.78)	23.46 (3.76)
μ=0.94	NP	28.25 (6.97)	25 (4.65)
	DP	28.02 (6.19)	24.24 (4.35)
	MP	26.07 (5.97)	25.65 (2.29)
	AP	23.30 (6.72)	23.85 (4.16)
μ=0.95	NP	28 (6.88)	25.66 (4.67)
	DP	27.74 (6.10)	24.55 (4.40)
	MP	25.80 (5.89)	26.52 (2.79)
	AP	23.09 (6.68)	25 (3.54)
μ=0.96	NP	27.75 (6.79)	25.92 (4.63)
	DP	27.46 (6.00)	25.30 (4.32)
	MP	25.49 (5.78)	26.95 (2.92)
	AP	22.86 (6.64)	25.38 (3.80)
μ=0.97	NP	27.46 (6.69)	26.71 (4.54)
	DP	27.14 (5.90)	26.36 (3.81)
	MP	25.15 (5.69)	27.61 (2.97)
	AP	22.61 (6.59)	25.38 (3.80)
μ=0.98	NP	27.12 (6.58)	27.11 (4.60)
	DP	26.77 (5.79)	26.97 (3.52)
	MP	24.75 (5.57)	28.26 (2.43)
	AP	22.29 (6.54)	25.77 (4.00)
μ=0.99	NP	26.68 (6.44)	27.76 (4.30)
	DP	26.31 (5.68)	27.42 (3.56)
	MP	24.26 (5.43)	28.26 (2.43)
	AP	21.90 (6.46)	26.54 (3.76)

Tabella 6.3: Tabella con prestazioni algoritmo di predizione al variare del forgetting factor (valori da 0.91 a 0.99) (NP=paziente "non predicibile", DP= paziente "debolmente predicibile", MP= paziente "moderatamente predicibile", AP= paziente "altamente predicibile")

Capitolo 7

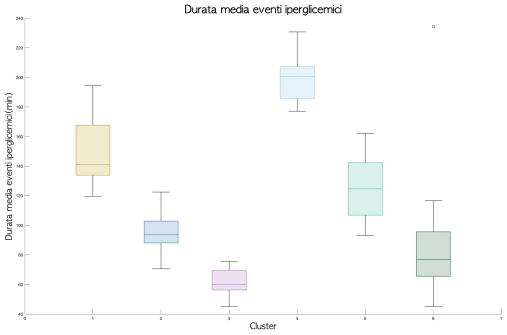
Conclusioni e possibili sviluppi futuri

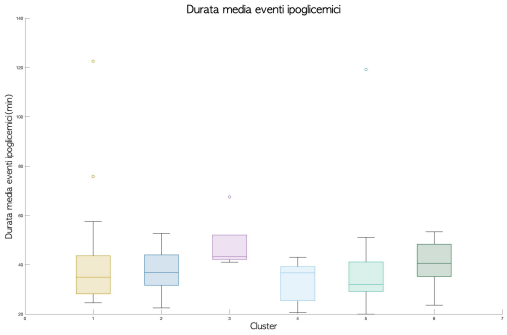
L'obbiettivo del presente lavoro di tesi era quello di coniugare tecniche ed intenti di due discipline, machine learning e precision medicine, al fine di indagare su possibili metodologie per il miglioramento della condizione clinica e della terapia di pazienti con diabete di tipo I. Nello specifico, sono state utilizzate due diverse tecniche di stratificazione non supervisionata per verificare l'esistenza di gruppi di pazienti con caratteristiche simili tra loro; a differenza però di altri lavori presenti in letteratura, come quelli riportati da [21] a [24], si è deciso qui di sfruttare solamente i dati provenienti da un sensore di monitoraggio in continua della glicemia di pazienti con diabete di tipo I, quindi lunghe serie temporali di campioni di glucosio, senza nessun'altra informazione a priori sui soggetti stessi (come peso, età, sesso o altri parametri clinici). La conferma dell'esistenza di "sottocategorie" di pazienti diabetici (e non solo) potrebbe contribuire alla realizzazione di uno dei propositi principali della "medicina di precisione", ovvero lo studio e l'elaborazione di terapie sempre più personalizzate ed ottimizzate sulle caratteristiche del paziente stesso.
La procedura di clustering è stata qui applicata sia su dati CGM rappresentanti "profili globali" dei pazienti (che considerassero quindi complessivamente tutte le settimane di ciascuno di essi), sia su profili glicemici settimanali, al fine di, come già ribadito, individuare possibili sottogruppi di pazienti o pattern glicemici settimanali con caratteristiche pressochè equivalenti tra loro. I risultati ottenuti ci hanno permesso effettivamente di isolare dei raggruppamenti di pazienti e profili settimanali, dei quali sono state riportate le caratteristiche e delle prime speculazioni sulla loro interpretazione. Sono però necessari ulteriori studi su diversi aspetti dei clusters ottenuti; studi che potrebbero ad esempio indagare sulla loro riproducibilità in altri dataset o correlazione con determinate caratteristiche

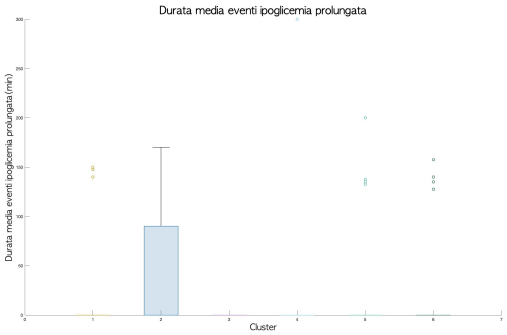
cliniche dei pazienti (ad esempio con future comorbidità), permettendo quindi di conferire ulteriore completezza ed autorevolezza, nonchè validità statistica, ad analisi simili a quella qui riportata. Inoltre, sarebbe opportuno anche indagare ulteriormente sulle prestazioni di algoritmi di "soft clustering": come è stato riportato nel corso della presente tesi infatti, i cluster ottenuti avevano varianza elevata, probabilmente per il fatto che i profili dei pazienti non formavano insiemi disgiunti nello spazio delle features; di conseguenza, algoritmi nei quali i soggetti possono appartenere a più di un solo cluster, potrebbero fornire risultati maggiormente soddisfacenti. Nel complesso comunque, i risultati di questa analisi preliminare ed al momento qualitativa, senza quindi alcuna prova di tipo statistico, sono buoni e possono essere l'incipit di nuovi sviluppi futuri.
Allo stesso modo, anche il concetto di "paziente predicibile", elaborato proprio grazie al fondamento logico che sta alla base della procedura di stratificazione ed ai risultati ottenuti grazie ad esso, sembra essere promettente, visti anche gli esiti positivi delle ultime analisi riportate alla fine del capitolo precedente. Anche in questo caso però, è necessario verificare se ciò che si è ottenuto è riproducibile e robusto anche con dati provenienti da altri dataset.
In conclusione, il presente studio ha conseguito risultati positivi e potrebbe contribuire all'esplorazione di nuovi orizzonti della medicina che, usufruendo trasversalmente di competenze appartenenti ad altri ambiti disciplinari, sarà sempre più in grado di migliorare le condizioni cliniche, e di vita, dei pazienti affetti da diabete.

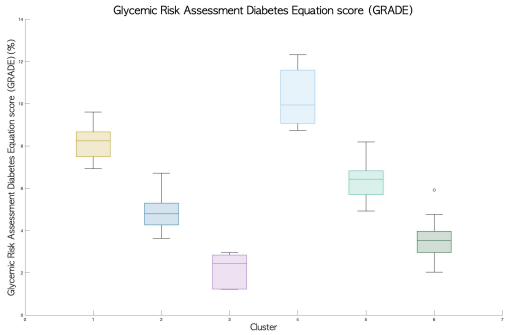
Appendice A

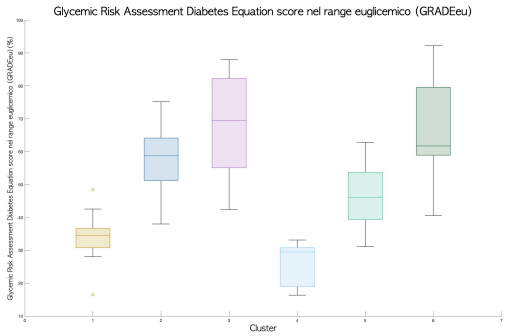
Boxplot risultati clustering dei pazienti

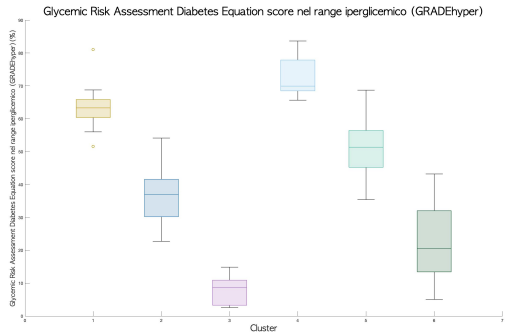

Durata media eventi iperglicemici
Durata media eventi iperglicemici(min)
Cluster

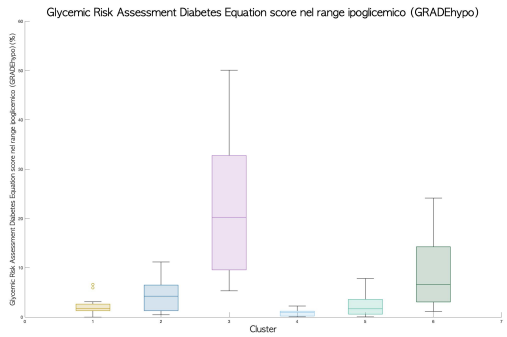

Durata media eventi ipoglicemici
Durata media eventi ipoglicemici(min)
Cluster

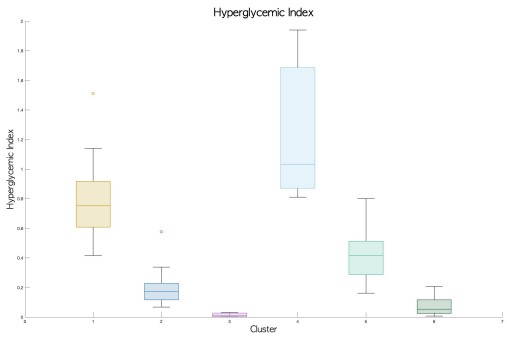

Durata media eventi ipoglicemia prolungata
Durata media eventi ipoglicemia prolungata(min)
Cluster

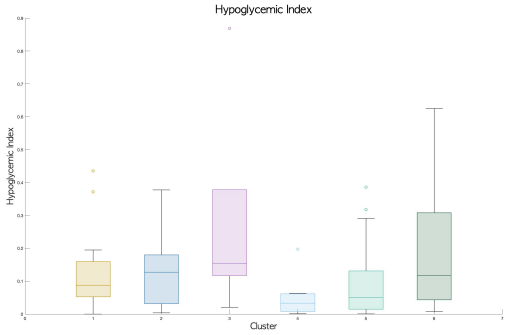

Glycemic Risk Assessment Diabetes Equation score (GRADE)
Glycemic Risk Assessment Diabetes Equation score (GRADE) (%)
Cluster

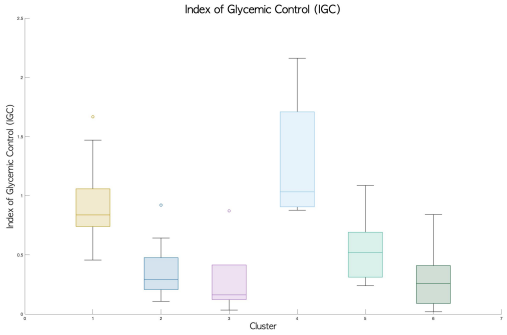

Glycemic Risk Assessment Diabetes Equation score nel range euglicemico (GRADEeu)
Cluster

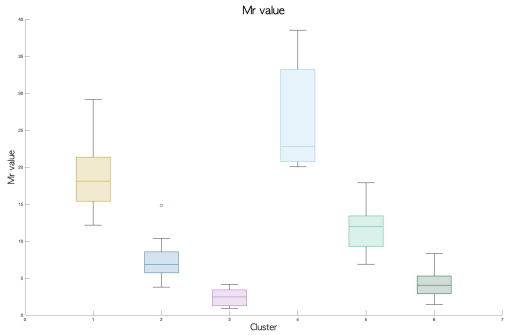

Glycemic Risk Assessment Diabetes Equation score nel range iperglicemico (GRADEhyper)
Cluster

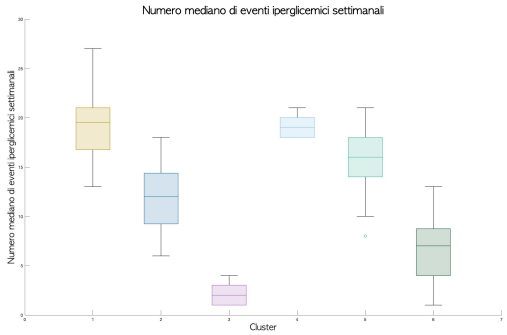

Glycemic Risk Assessment Diabetes Equation score nel range ipoglicemico (GRADEhypo)
Cluster

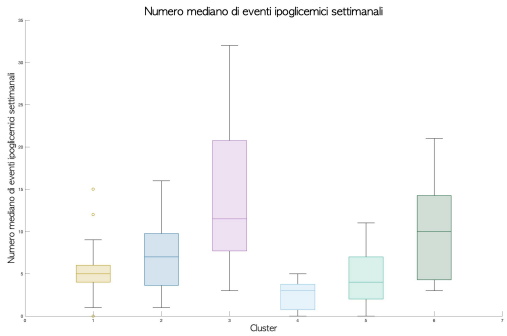

Hyperglycemic Index
Hyperglycemic Index
Cluster

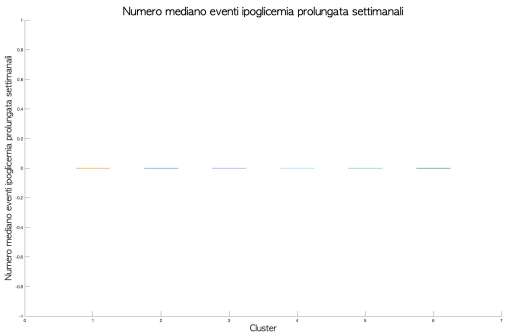

Hypoglycemic Index
Hypoglycemic Index
Cluster

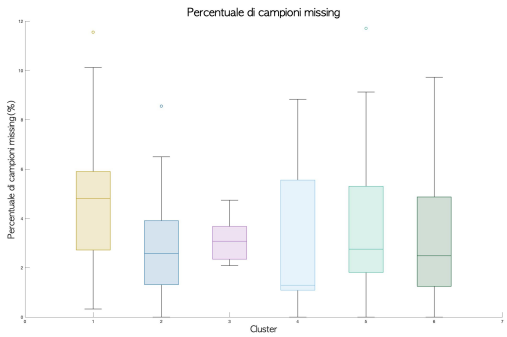

Index of Glycemic Control (IGC)
Index of Glycemic Control (IGC)
Cluster

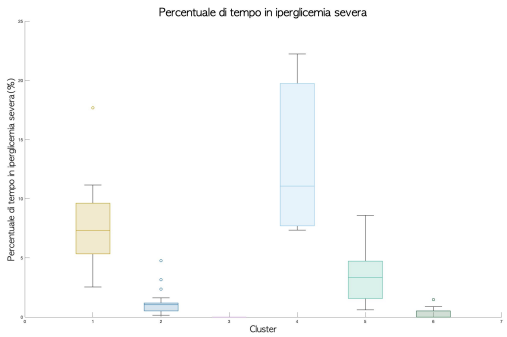

Mr value
Mr value
Cluster

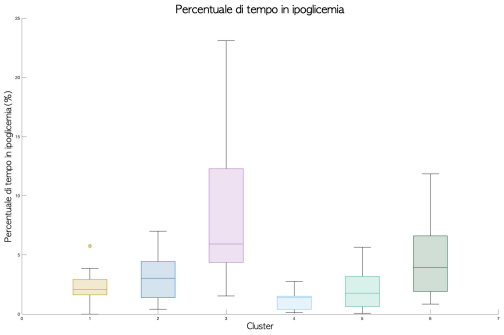

Numero mediano di eventi iperglicemici settimanali
Numero mediano di eventi iperglicemici settimanali
Cluster

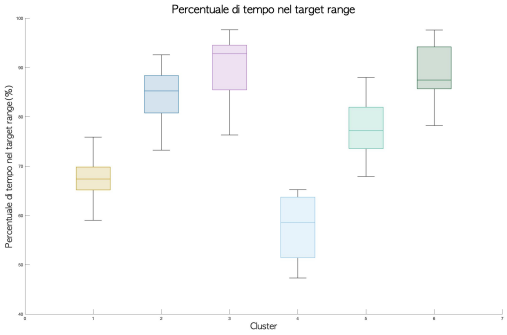

Numero mediano di eventi ipoglicemici settimanali
Numero mediano di eventi ipoglicemici settimanali
Cluster

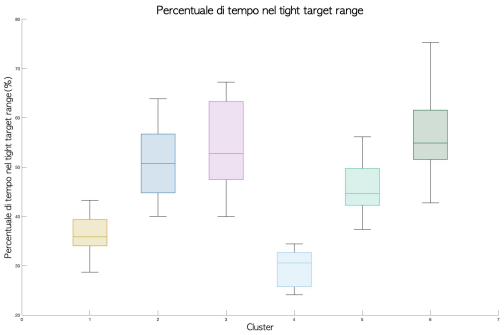

Numero mediano eventi ipoglicemia prolungata settimanali
Numero mediano eventi ipoglicemia prolungata settimanali
Cluster

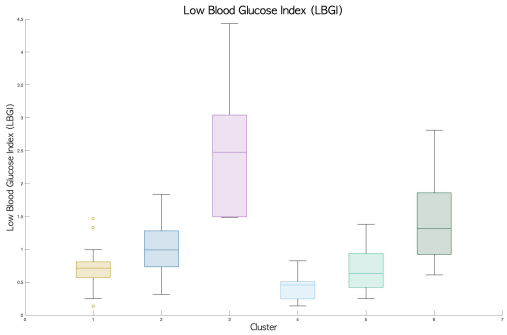

Percentuale di campioni missing
Percentuale di campioni missing (%)
Cluster

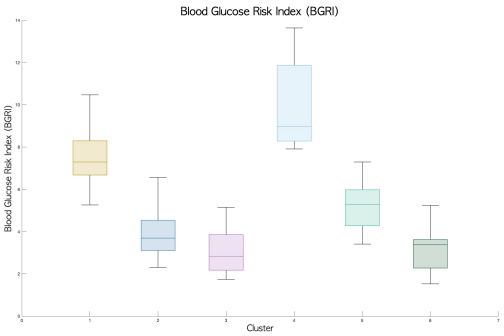

Percentuale di tempo in iperglicemia severa
Percentuale di tempo in iperglicemia severa (%)
Cluster

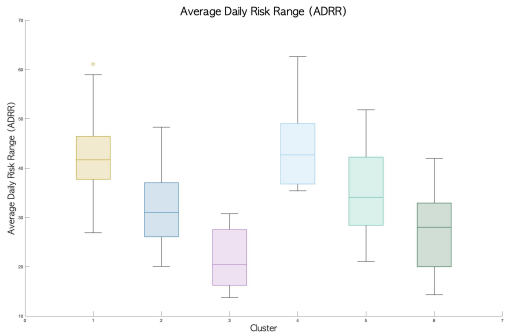

Percentuale di tempo in ipoglicemia
Percentuale di tempo in ipoglicemia (%)
Cluster

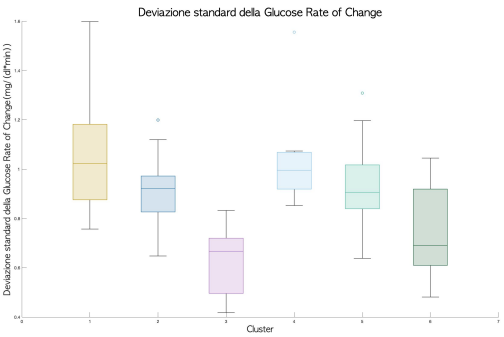

Percentuale di tempo nel target range
Percentuale di tempo nel target range (%)
Cluster

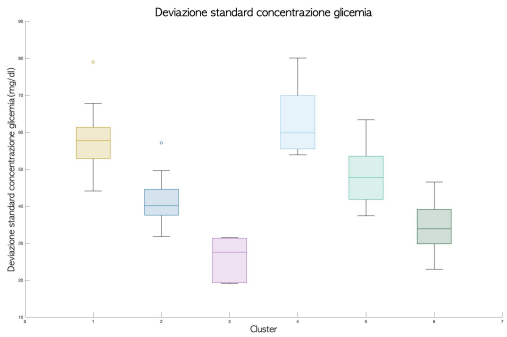

Percentuale di tempo nel tight target range
Percentuale di tempo nel tight target range (%)
Cluster

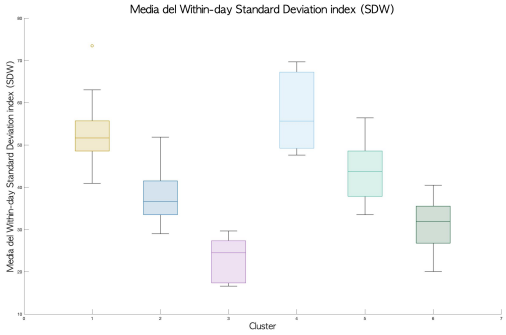

Low Blood Glucose Index (LBGI)
Low Blood Glucose Index (LBGI)
Cluster

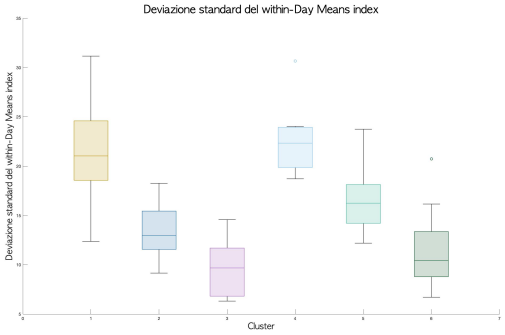

Blood Glucose Risk Index (BGRI)
Blood Glucose Risk Index (BGRI)
Cluster

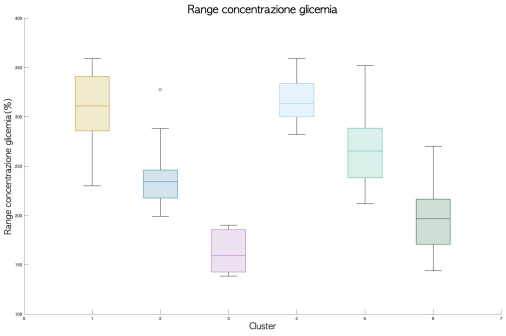

Average Daily Risk Range (ADRR)
Average Daily Risk Range (ADRR)
Cluster

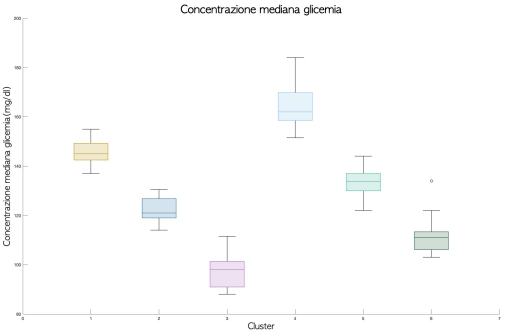

Deviazione standard della Glucose Rate of Change
Deviazione standard della Glucose Rate of Change (mg/(dl*min))
Cluster

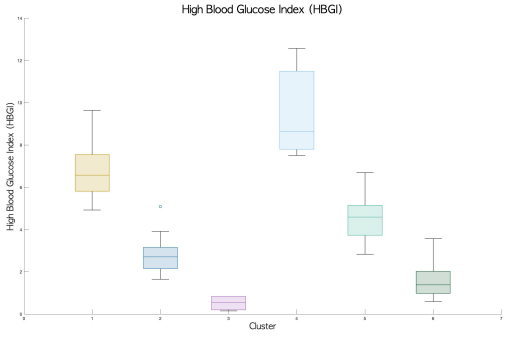

Deviazione standard concentrazione glicemia
Deviazione standard concentrazione glicemia (mg/dl)
Cluster

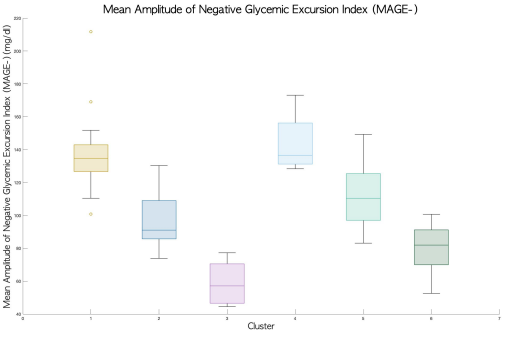

Media del Within-day Standard Deviation index (SDW)
Media del Within-day Standard Deviation index (SDW)
Cluster

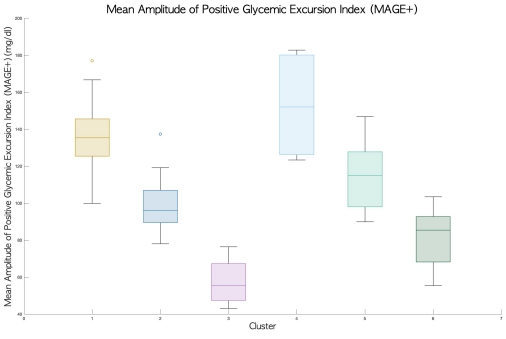

Deviazione standard del within-Day Means index
Deviazione standard del within-Day Means index
Cluster

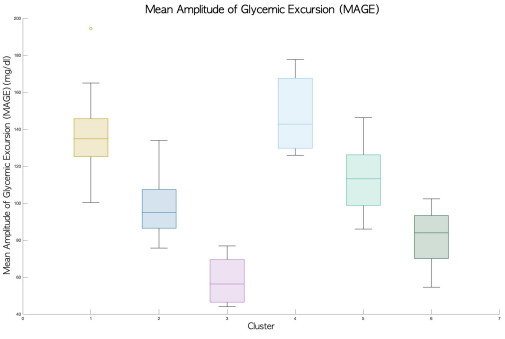

Range concentrazione glicemia
Range concentrazione glicemia (%)
Cluster

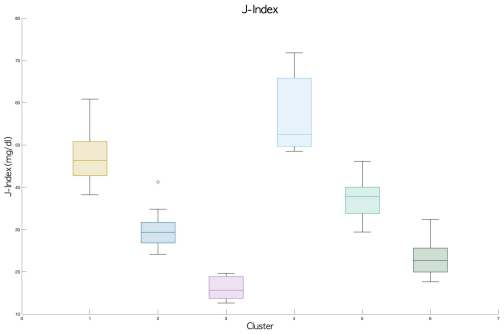

Concentrazione mediana glicemia
Concentrazione mediana glicemia (mg/dl)
Cluster

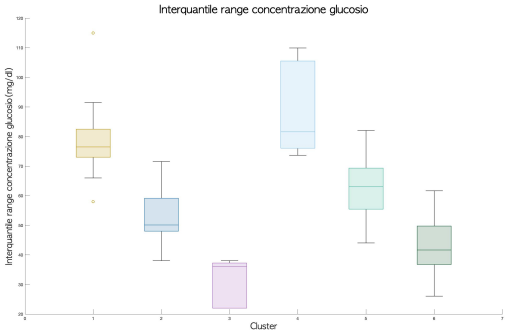

High Blood Glucose Index (HBGI)
High Blood Glucose Index (HBGI)
Cluster

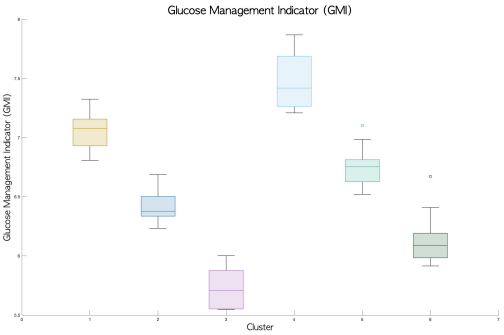

Mean Amplitude of Negative Glycemic Excursion Index (MAGE-)
Mean Amplitude of Negative Glycemic Excursion Index (MAGE-) (mg/dl)
Cluster

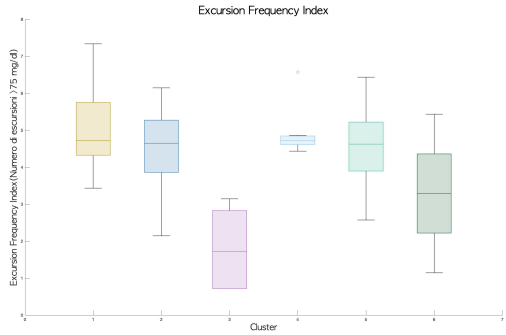

Mean Amplitude of Positive Glycemic Excursion Index (MAGE+)
Mean Amplitude of Positive Glycemic Excursion Index (MAGE+) (mg/dl)
Cluster

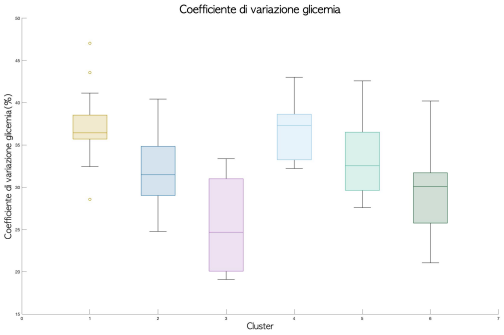

Mean Amplitude of Glycemic Excursion (MAGE)
Mean Amplitude of Glycemic Excursion (MAGE) (mg/dl)
Cluster

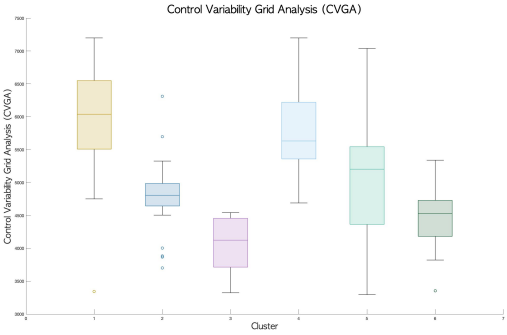

J-Index
J-Index (mg/dl)
Cluster

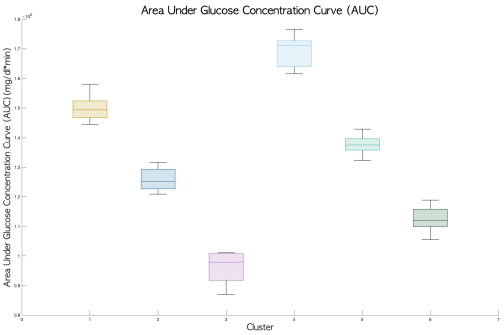

Interquantile range concentrazione glucosio
Interquantile range concentrazione glucosio (mg/dl)
Cluster

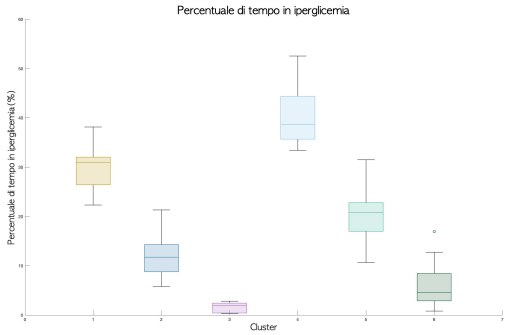

Glucose Management Indicator (GMI)
Glucose Management Indicator (GMI)
Cluster


Excursion Frequency Index
Excursion Frequency Index (Numero di escursioni ↓75 mg/dl)
Cluster


Coefficiente di variazione glicemia
Coefficiente di variazione glicemia (%)
Cluster

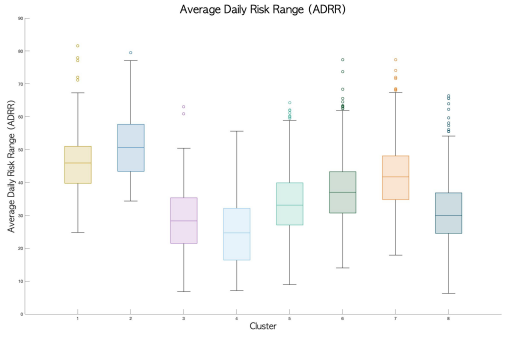

Control Variability Grid Analysis (CVGA)
Control Variability Grid Analysis (CVGA)
Cluster

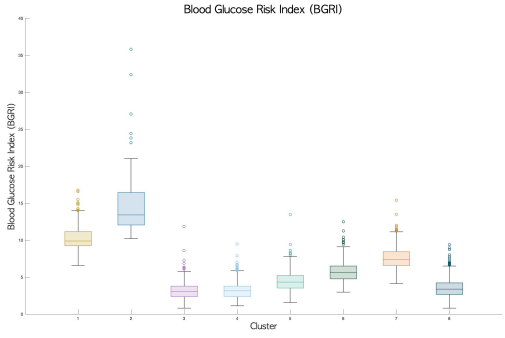

Area Under Glucose Concentration Curve (AUC)
Area Under Glucose Concentration Curve (AUC) (mg/dl*min)
Cluster

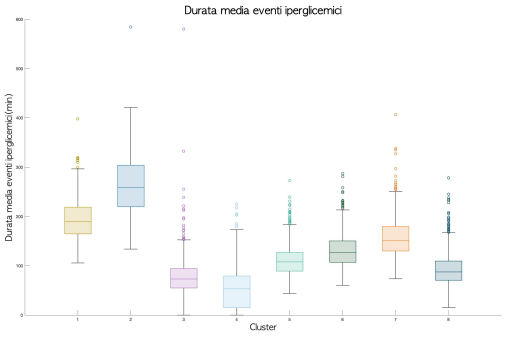

Percentuale di tempo in iperglicemia
Percentuale di tempo in iperglicemia (%)
Cluster

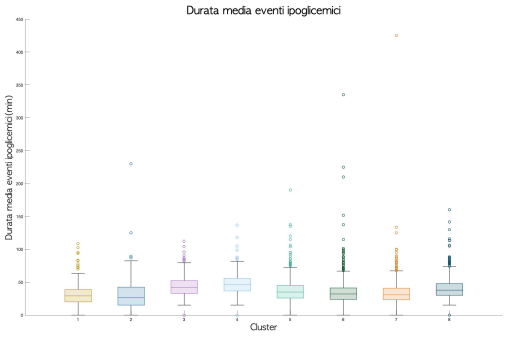
Appendice B

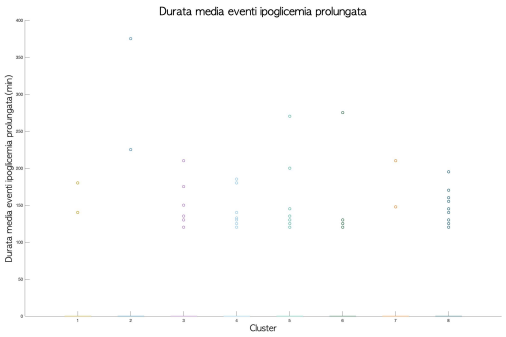
Boxplot risultati clustering dei profili settimanali

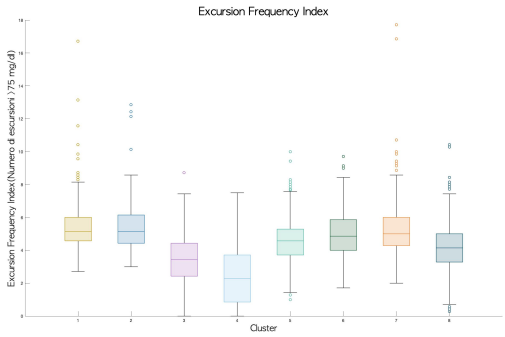

Average Daily Risk Range (ADRR)
Average Daily Risk Range (ADRR)
Cluster

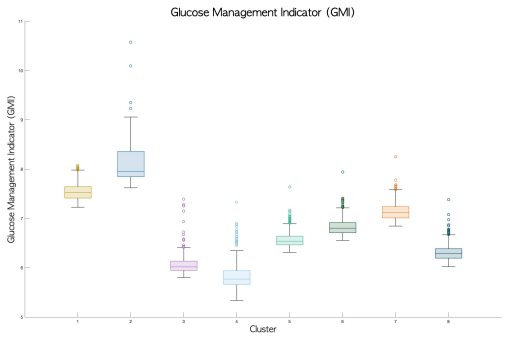

Blood Glucose Risk Index (BGRI)
Blood Glucose Risk Index (BGRI)
Cluster

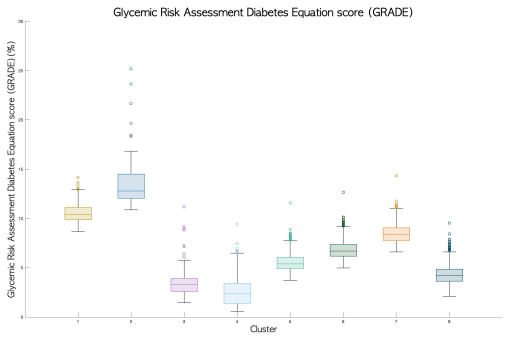

Durata media eventi iperglicemici
Durata media eventi iperglicemici(min)
Cluster

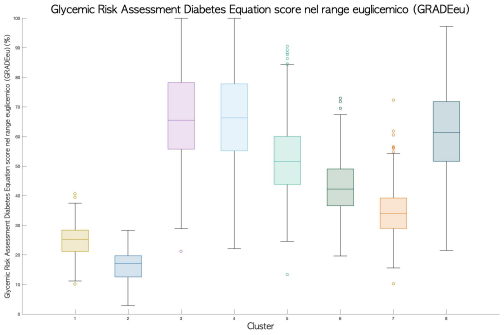

Durata media eventi ipoglicemici
Durata media eventi ipoglicemici(min)
Cluster

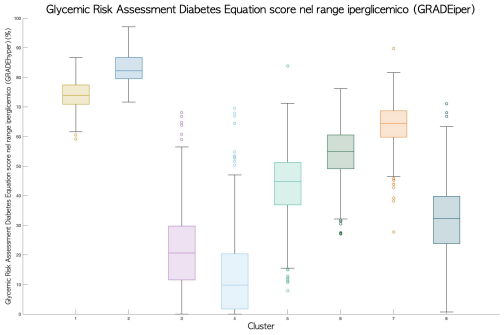

Durata media eventi ipoglicemia prolungata
Durata media eventi ipoglicemia prolungata (min)
Cluster

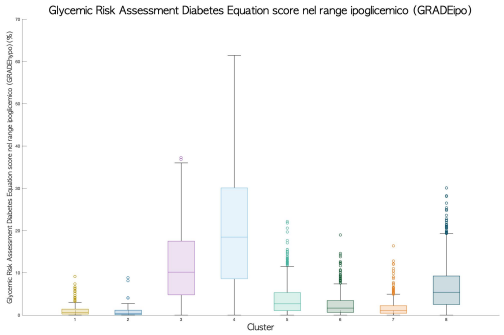

Excursion Frequency Index
Excursion Frequency Index(Numero di escursioni >75 mg/dl)
Cluster

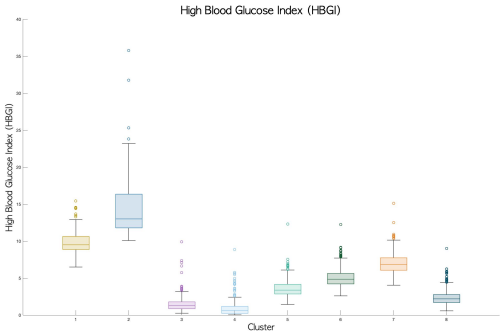

Glucose Management Indicator (GMI)
Glucose Management Indicator (GMI)
Cluster

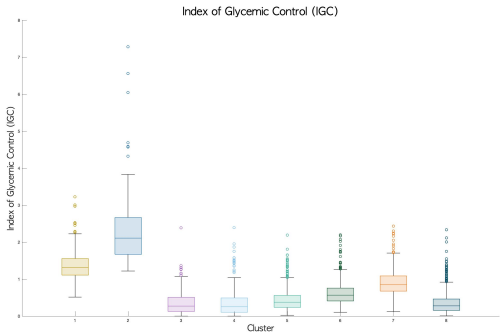

Glycemic Risk Assessment Diabetes Equation score (GRADE)
Glycemic Risk Assessment Diabetes Equation score (GRADE) (%)
Cluster

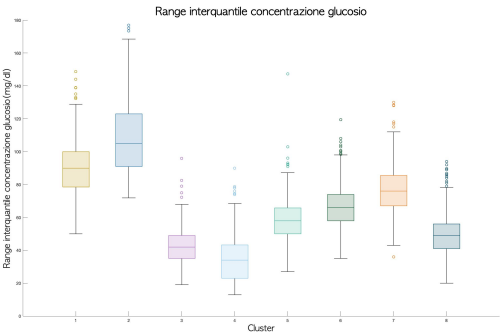

Glycemic Risk Assessment Diabetes Equation score nel range euglicemico (GRADEeu)
Glycemic Risk Assessment Diabetes Equation score nel range euglicemico (GRADEeu) (%)
Cluster

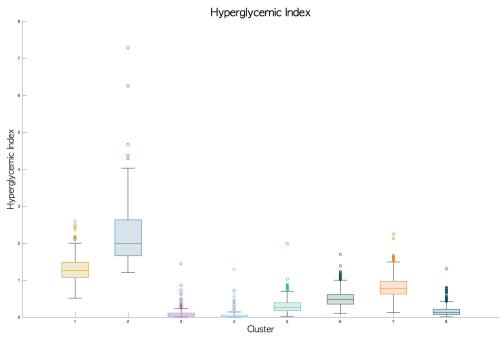

Glycemic Risk Assessment Diabetes Equation score nel range iperglicemico (GRADEiper)
Glycemic Risk Assessment Diabetes Equation score nel range iperglicemico (GRADEiper) (%)
Cluster

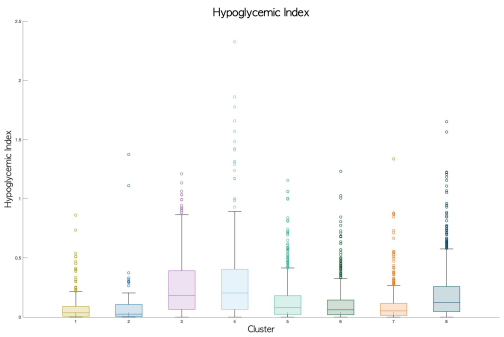

Glycemic Risk Assessment Diabetes Equation score nel range ipoglicemico (GRADEipo)
Glycemic Risk Assessment Diabetes Equation score nel range ipoglicemico (GRADEipo) (%)
Cluster

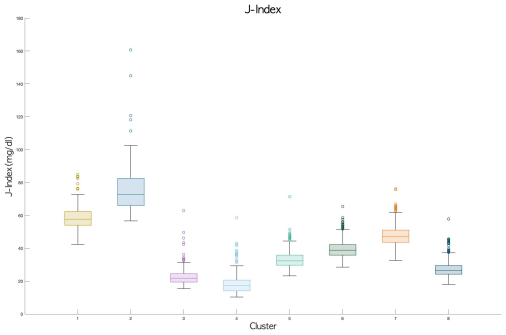

High Blood Glucose Index (HBGI)
High Blood Glucose Index (HBGI)
Cluster

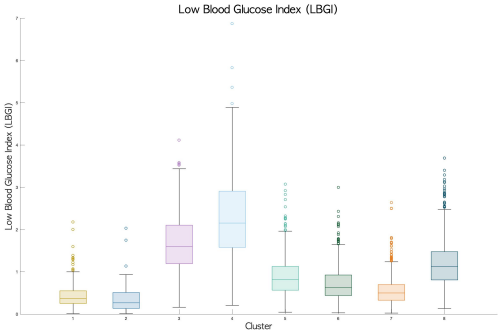

Index of Glycemic Control (IGC)
Index of Glycemic Control (IGC)
Cluster

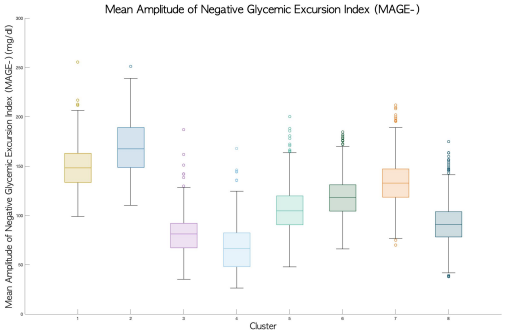

Range interquantile concentrazione glucosio
Range interquantile concentrazione glucosio (mg/dl)
Cluster

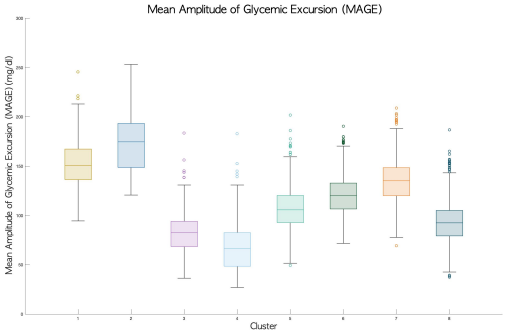

Hyperglycemic Index
Hyperglycemic Index
Cluster

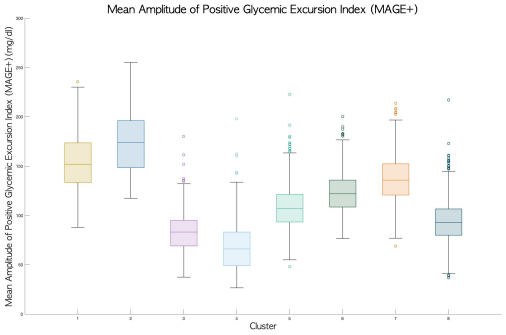

Hypoglycemic Index
Hypoglycemic Index
Cluster

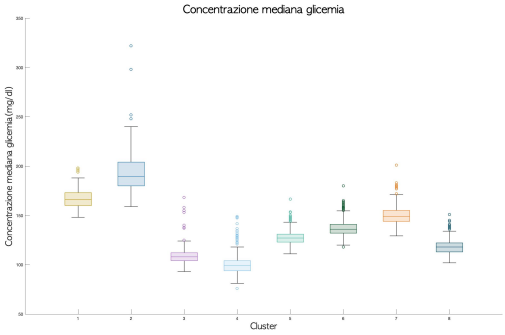

J-Index
J-Index (mg/dl)
Cluster

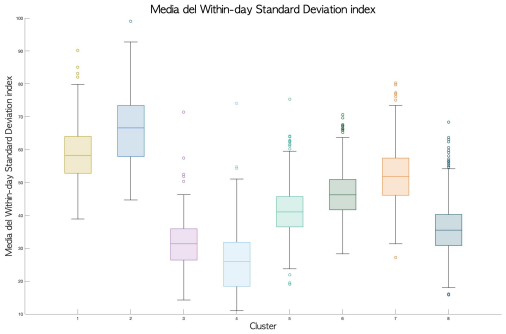

Low Blood Glucose Index (LBGI)
Low Blood Glucose Index (LBGI)
Cluster

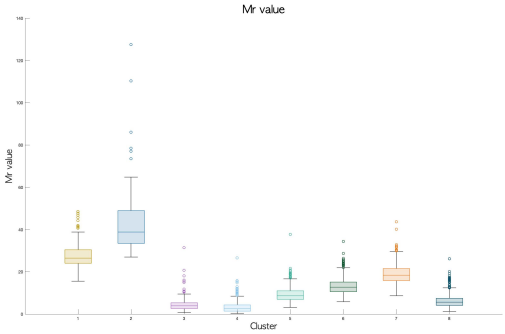

Mean Amplitude of Negative Glycemic Excursion Index (MAGE-)
Mean Amplitude of Negative Glycemic Excursion Index (MAGE-) (mg/dl)
Cluster

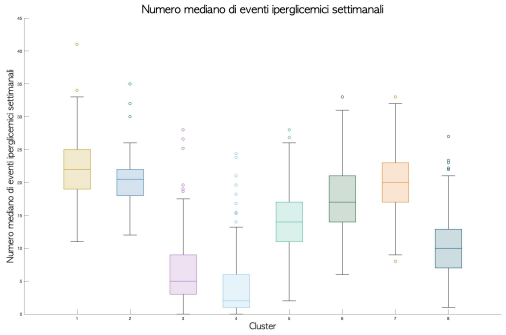

Mean Amplitude of Glycemic Excursion (MAGE)
Mean Amplitude of Glycemic Excursion (MAGE) (mg/dl)
Cluster

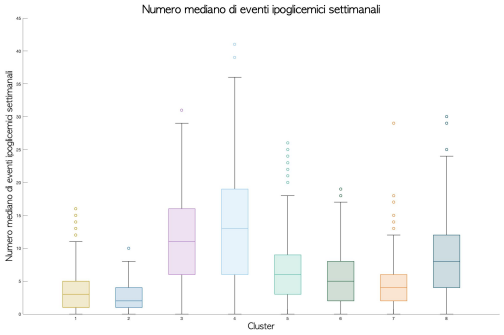

Mean Amplitude of Positive Glycemic Excursion Index (MAGE+)
Mean Amplitude of Positive Glycemic Excursion Index (MAGE+) (mg/dl)
Cluster

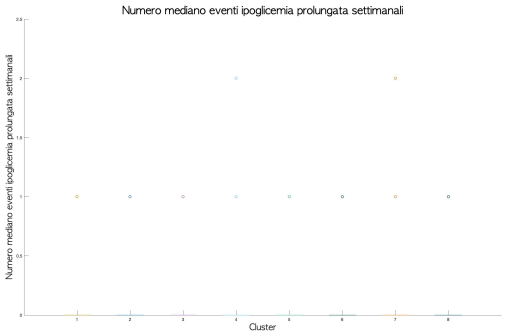

Concentrazione mediana glicemia
Concentrazione mediana glicemia (mg/dl)
Cluster

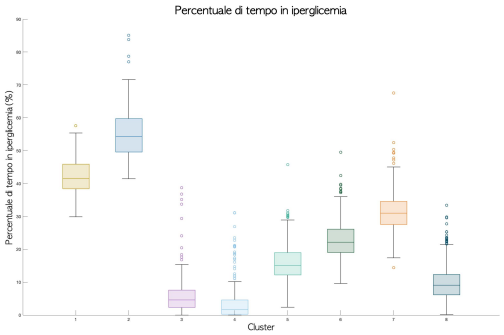

Media del Within-day Standard Deviation index
Media del Within-day Standard Deviation index
Cluster

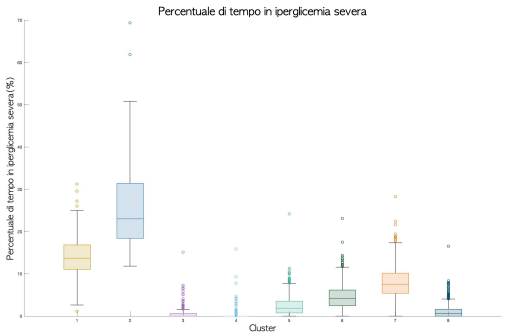

Mr value
Mr value
Cluster

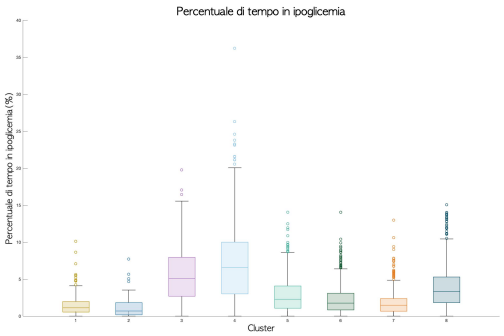

Numero mediano di eventi iperglicemici settimanali
Numero mediano di eventi iperglicemici settimanali
Cluster

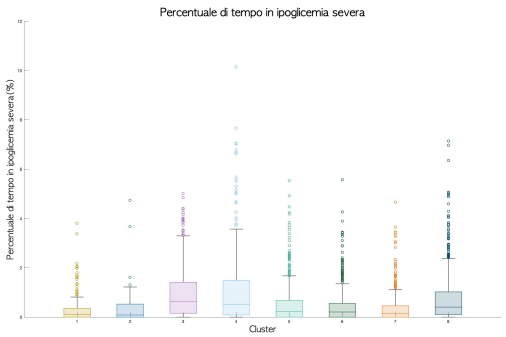

Numero mediano di eventi ipoglicemici settimanali
Numero mediano di eventi ipoglicemici settimanali
Cluster

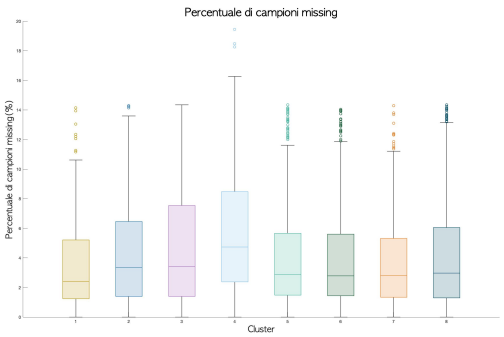

Numero mediano eventi ipoglicemia prolungata settimanali
Numero mediano eventi ipoglicemia prolungata settimanali
Cluster

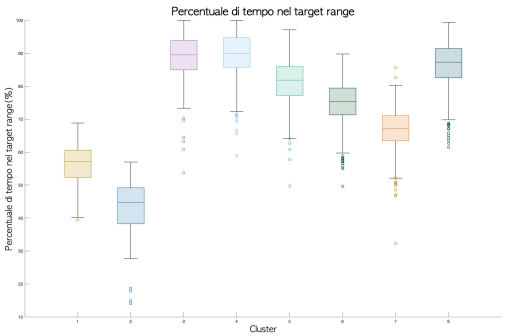

Percentuale di tempo in iperglicemia
Percentuale di tempo in iperglicemia (%)
Cluster

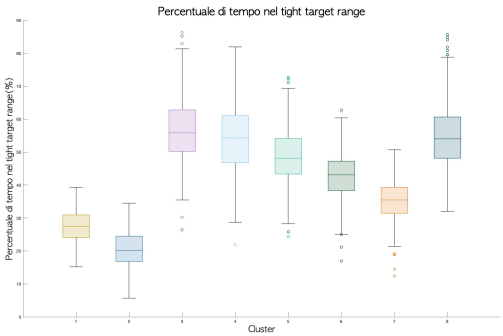

Percentuale di tempo in iperglicemia severa
Percentuale di tempo in iperglicemia severa (%)
Cluster

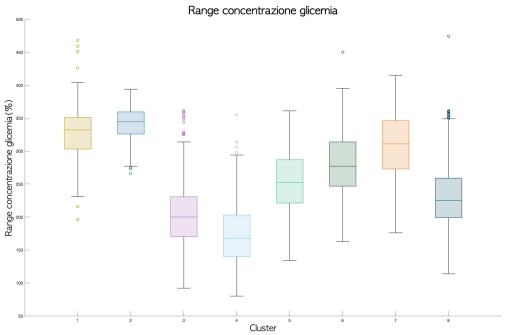

Percentuale di tempo in ipoglicemia
Percentuale di tempo in ipoglicemia (%)
Cluster

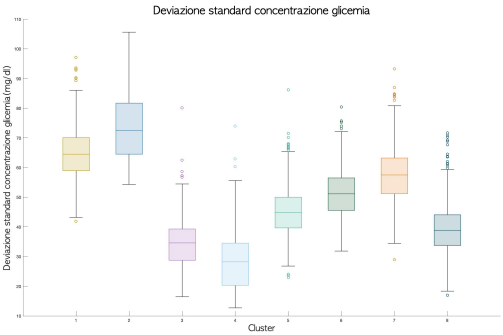

Percentuale di tempo in ipoglicemia severa
Percentuale di tempo in ipoglicemia severa (%)
Cluster

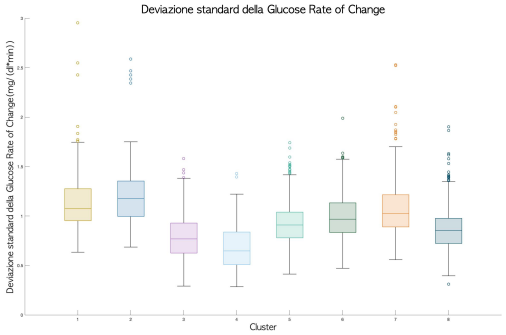

Percentuale di campioni missing
Percentuale di campioni missing (%)
Cluster

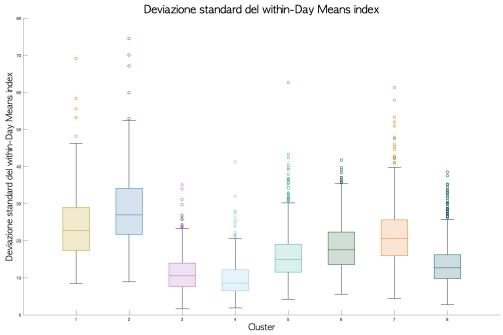

Percentuale di tempo nel target range
Percentuale di tempo nel target range(%)
Cluster


Percentuale di tempo nel tight target range
Percentuale di tempo nel tight target range(%)
Cluster


Range concentrazione glicemia
Range concentrazione glicemia(%)
Cluster


Deviazione standard concentrazione glicemia
Deviazione standard concentrazione glicemia(mg/dl)
Cluster


Deviazione standard della Glucose Rate of Change
Deviazione standard della Glucose Rate of Change(mg/(dl*min))
Cluster


Deviazione standard del within-Day Means index
Deviazione standard del within-Day Means index
Cluster

Bibliografia

[1] World Health Organization, “Global report on diabetes” , 2016, DOI: non disponibile;

[2] Centers for Disease Control and Prevention, “National Diabetes Statistics Report, 2020”, 2020, Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services, DOI: non disponibile;

[3] Istituto Nazionale di Statistica (ISTAT), “Annuario Statistico Italiano 2016”, Sanità e Salute, vol.4, 2016, DOI: non disponibile;

[4] Peric S. et al., “Diabetes and COVID-19”, Wien Klin Wochenschr, vol.132, pp. 356–361, 2020, DOI:10.1007/s00508-020-01672-3;

[5] American Diabetes Association, “Diagnosis and Classification of Diabetes Mellitus”, Diabetes Care, vol.37, 2014, DOI:10.2337/dc14-S081;

[6] David C. Klonoff, “Continuous Glucose Monitoring: Roadmap for 21st century diabetes therapy”, Diabetes Care, vol.28, pp. 1231-1239, 2004, DOI:10.2337/diacare.28.5.1231;

[7] Michael R. Kosorok et al., “Precision Medicine”, Annual Review of Statistics and Its Application, vol. 6, pp. 263–86, 2019, DOI:10.1146/annurev-statistics-030718-105251;

[8] Enzo Bonora e Giorgio Sesti, “Il Diabete in Italia”, Società Italiana di Diabetologia, 2016, DOI: non disponibile;

[9] C. Cobelli, et al., “A compartmental model to quantitate in vivo glucose transport in the human forearm”, American Journal of Physiology-Endocrinology and Metabolism, vol. 257, pp. E943-E958, 1989, DOI:10.1152/ajpendo.1989.257.6.E943;

[10] Thomas Danne et al., "International Consensus on Use of Continuous Glucose Monitoring", Diabetes Care, vol.40, pp. 1631-1640, 2017, DOI:10.2337/dc17-1600;

[11] American Diabetes Association, "Microvascular Complications and Foot Care: Standards of Medical Care in Diabetes 2021", Diabetes Care, vol. 44, pp. S151–S167, 2021, DOI:10.2337/dc21-S011;

[12] Katherine R. Tuttle et al., "Diabetic Kidney Disease: A Report From an ADA Consensus Conference", Diabetes Care, vol.37, pp.2864–2883, 2014, DOI:10.2337/dc14-1296;

[13] American Diabetes Association, "Comprehensive Medical Evaluation and Assessment of Comorbidities: Standards of Medical Care in Diabetes 2021", Diabetes Care, vol.44, pp. S40–S52, 2021, DOI:10.2337/dc21-S004;

[14] Salvatore Carbone et al., "Obesity, risk of diabetes and role of physical activity, exercise training and cardiorespiratory fitness", Progress in Cardiovascular Diseases, vol.62, pp.327–333, 2019, DOI:10.1016/j.pcad.2019.08.004;

[15] S.Walford et al., "Self-Monitoring of Blood Glucose: Improvement of Diabetic Control", The Lancet, vol.311, pp. 732-735, 1978, DOI: 10.1016/S0140-6736(78)90855-3;

[16] Martina Montagnana et al., "Overview on self monitoring of blood glucose", Clinica Chimica Acta, vol.402, pp. 7-13, 2009, DOI: 10.1016/j.cca.2009.01.002;

[17] David Rodbard, "Continuous Glucose Monitoring: A Review of Successes, Challenges, and Opportunities", Diabetes Technology and Therapeutics , vol.18, 2016, DOI: 10.1089/dia.2015.0417;

[18] John C. Pickup et al., "Insulin-Pump Therapy for Type 1 Diabetes Mellitus", Clinical Therapeutics, The New English Journal, vol.366, pp.1616-24, 2012, DOI: non disponibile;

[19] Andreas Melmer et al., "Glycaemic control in individuals with type 1 diabetes using an open source artificial pancreas system (OpenAPS)", Diabetes Obese Metabolism, vol. 21, pp. 2333–2337, 2019, DOI: 10.1111/dom.13810;

[20] Christine Knoll et al., "Real-world evidence on clinical outcomes of people with type 1 diabetes using open-source and commercial automated insulin dosing systems: A systematic review", Diabetic Medicine, 2021, DOI: 10.1111/dme.14741;

[21] Rui Tao et al., "Multilevel clustering approach driven by continuous glucose monitoring data for further classification of type 2 diabetes", BMJ Open Diabetes Research Care, vol.9, 2021, DOI: 10.1136/bmjdrc-2020-001869;

[22] Lyvia Biagi et al., "Individual categorisation of glucose profiles using compositional data analysis", Statistical Methods in Medical Research, pp. 1-18, 2018, DOI: 10.1177/0962280218808819;

[23] Corinna Schröder et al., "Classification of postprandial glycemic patterns in type 1 diabetes subjects under closed-loop control: an in silicon study", Annual International Conference IEEE Eng. Med. Biol. Soc. , pp. 5443-5446, DOI: 10.1109/EMBC.2019.8857246.;

[24] Ivan Contreras et al., "Profiling intra-patients type I diabetes behaviour", Computational Methods Programs Biomed., vol. 136, pp. 131-41, DOI: 10.1016/j.cmpb.2016.08.022;

[25] Zekai Wu et al., "Use of a do-it-yourself artificial pancreas system is associated with better glucose management and higher quality of life among adults with type 1 diabetes", Therapeutic Advances in Endocrinology and Metabolism, vol. 11, pp.1-11, 2020, DOI: 10.1177/2042018820950146;

[26] Martin C. Nwadiugwu et al., "Identifying Glycemic Variability in Diabetes Patient Cohorts and Evaluating Disease Outcomes", Journal of Clinical Medicine, vol.10, 2021, DOI: 10.3390/ jcm10071477;

[27] Kazuhiko Sakaguchi, "Glucose area under the curve during oral glucose tolerance test as an index of glucose intolerance", Diabetolog Int., vol. 7, pp: 53-58, 2016, DOI: 10.1007/s13340-015-0212-4;

[28] David B. Allison et al., "The Use of Areas Under Curves in Diabetes Research", Diabetes Care, vol. 18, 1995, DOI: 10.2337/diacare.18.2.245;

[29] Lalo Magni et al., "Evaluating the Efficacy of Closed-Loop Glucose Regulation via Control-Variability Grid Analysis", Journal of Diabetes Science and Technology, vol.2, pp. 630-635, 2008, DOI: 10.1177/193229680800200414;

[30] Fabris Chiara et al., “Glucose variability indices in type 1 diabetes: Parsimonious set of indices revealed by sparse principal component analysis”, Diabetes Technologies and Therapeutics, vol.16, pp. 644–652, 2014, DOI: 10.1089/dia.2013.0252;

[31] Louis Monnier et al., “Toward Defining the Threshold Between Low and High Glucose Variability in Diabetes”, Diabetes Care, vol. 40, pp. 832–838, 2017, DOI:10.2337/dc16-1769;

[32] Richard M. Bergenstal et al., “Glucose Management Indicator (GMI): A New Term for Estimating A1C From Continuous Glucose Monitoring”, Diabetes Care, vol. 41, pp. 2275–2280, 2018, DOI:10.2337/dc18-1581;

[33] Wójcicki et al., ”J-index. A new proposition of the assessment of current glucose control in diabetic patients”, Hormone and Metabolic Research, vol. 27, pp. 41-42, 1995, DOI: 10.1055/s-2007-979906;

[34] Clarke et al., ”Statistical Tools to Analyze Continuous Glucose Monitor Data”, Diabetes Technol Ther, vol. 11, pp. S45-S54, 2009, DOI: 10.1089=dia.2008.0138;

[35] Service et al., ”Mean amplitude of glycemic excursions, a measure of diabetic instability”, Diabetes, vol. 19, pp. 644-655, 1970, DOI: 10.2337/diab.19.9.644;

[36] Kovatchev et al., ”Symmetrization of the blood glucose measurement scale and its applications”, Diabetes Care, vol. 20, pp. 1655-1658, 1997, DOI: 10.2337/diacare.20.11.1655;

[37] Kovatchev et al., ”Evaluation of a new measure of blood glucose variability in diabetes”, Diabetes Care, vol. 29, pp. 2433-2438, 2006, DOI: 10.2337/dc06-1085;

[38] N. R. Hill et al., ”A method for assessing quality of control from glucose profiles”, Diabetic Medicine, vol. 24, pp. 753-758, 2007, DOI: 10.1111/j.1464-5491.2007.02119.x;

[39] Schlichtkrull et al., ”The M-value, an index of blood-sugar control in diabetics”, Acta Medica Scandinavica, vol. 177, pp. 95-102, 1965, DOI: 10.1111/j.0954-6820.1965.tb01810.x;

[40] Rich Caruana et al., "An Empirical Comparison of Supervised Learning Algorithms", Proceedings of the 23rd International Conference on Machine Learning, pp. 161–168, 2006, DOI: 10.1145/1143844.1143865;

[41] Jain, A. K. and Dubes, R. C., "Algorithms for Clustering Data", Prentice Hall Inc., Advanced Reference Series : Computer Science, 1988, DOI: non disponibile;

[42] Saeed Aghabozorgi et al., "Time-series clustering – A decade review", Information Systems, vol. 53, pp. 16–38, 2015, DOI: 10.1016/j.is.2015.04.007;

[43] Amit Saxena et al., "A review of clustering techniques and developments", Neurocomputing, vol. 267, pp. 664–681, 2017, DOI:10.1016/j.neucom.2017.06.053;

[44] S. P. Loyd, "Least squares quantization in PCM", IEEE Trans. on Infor. Theory, vol. 28, pp. 129-137, 1982, DOI:10.1109/TIT.1982.1056489;

[45] Rui Xu et al., "Clustering Algorithms in Biomedical Research: A Review", IEEE Reviews in Biomedical Engineering, vol. 3, 2010, DOI: 10.1109/RBME.2010.2083647;

[46] Boris Mirkin, "Choosing the number of clusters", WIREs Data Mining Knowl Discov, vol.1, pp. 252–260, 2011, DOI: 10.1002/widm.15;

[47] Jonathan Baarsch et al., "Investigation of Internal Validity Measures for K-Means Clustering", Proceedings of the International MultiConference Of Engineers and Computer Scientists, vol.1, pp.14-16, 2012, DOI: non disponibile;

[48] Kaufman Leonard and Peter J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis", 1990, DOI: non disponibile;

[49] G. Lance and W. Williams, "A general theory of classification sorting strategies 1. Hierarchical systems," Computer J., vol. 9, pp. 373–380, 1967, DOI: non disponibile;

[50] Yunjae Jung et al., "A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering", Journal of Global Optimization, vol. 25, pp. 91–111, 2003, DOI: 10.1023/A:1021394316112;

[51] David Rodbard, “Interpretation of Continuous Glucose Monitoring Data: Glycemic Variability and Quality of Glycemic Control”, Diabetes Technology and Therapeutics, vol. 11, 2009, DOI: 10.1089=dia.2008.0132;

Sitografia

- [1.1] International Diabetes Federation website. https://diabetesatlas.org/data/en/indicators/1/.
  Accesso in data: 3-12-2021;
- [1.2] Glucagone: Cos'è? Funzioni, Glicemia, Salute. https://m.prezzisalute.com/Salute/Glucagone.html.
  Accesso in data: 4-12-2021 ;
- [1.3] Troppi diabetici: tra pochi anni il 50% di loro rimarrà senza insulina. https://medicinaonline.co/2018/11/21/troppi-diabetici-tra-pochi-anni-il-50-di-loro-rimarra-senza-insulina/.
  Accesso in data: 4-12-2021;
- [1.4] Diabete, basta punture sul dito. La glicemia si misura con FreeStyle Libre di Abbott. https://www.wakeupnews.eu/diabete-basta-punture-sul-dito-la-glicemia-si-misura-freestyle-libre-abbott/.
  Accesso in data: 5-12-2021;
- [1.5] openAPS documentation. https://openaps.readthedocs.io/en/latest/.
  Accesso in data: 10-07-2021;