# Esplorazione Statistica di Dati Sanitari: il Caso Shanghai T2DM

Progetto di
*Statistica e Analisi dei Dati*

**Carotenuto Rosa**

**Guida Luigi**

**Data:** July 17, 2025

`Link alla repository`

# Contents

# 1 Introduzione

Il Diabete di Tipo 2 (T2D) rappresenta una delle principali sfide sanitarie globali, con una prevalenza in crescita e un impatto significativo sulla qualità della vita dei pazienti e sui sistemi sanitari nazionali. Caratterizzato da un'alterazione cronica del metabolismo del glucosio, il T2D viene monitorato clinicamente mediante parametri biochimici come l'emoglobina glicata (HbA1c) e, più recentemente, attraverso tecnologie di *Continuous Glucose Monitoring* (CGM), che forniscono un tracciamento ad alta frequenza dei livelli glicemici nel tempo.

## 1.1 Struttura dei dati

Nel contesto di questo progetto, si è disposto di due fonti principali di dati:

- **Summary**: un dataset aggregato (`Shanghai_T2DM_Summary.xlsx`) che raccoglie, per ciascun paziente, informazioni anagrafiche (età, sesso), cliniche (HbA1c, creatinina, BMI, comorbidità), comportamentali (fumo, consumo di alcol), e terapeutiche (uso di insulina e farmaci orali).
- **Dataset individuali**: file separati per ciascun paziente (es. `2000_0_20201230.xlsx`), contenenti dati CGM ad alta risoluzione temporale, insieme ad annotazioni relative a pasti, somministrazioni di insulina e altri eventi rilevanti. Questi dati sono organizzati in serie temporali e consentono analisi dinamiche del controllo glicemico.

## 1.2 Obiettivi del progetto

L'obiettivo di questo studio è analizzare e modellizzare i dati a disposizione mediante strumenti di **statistica descrittiva**, **inferenziale**, **modellazione multivariata**, **serie temporali** e **clustering**, con il supporto dell'ambiente statistico `R`. In particolare, le analisi statistiche condotte mirano a rispondere a una serie di domande di ricerca rilevanti per la comprensione del controllo glicemico nei pazienti con T2D. Gli obiettivi principali includono domande come:

- Esiste una correlazione tra le metriche derivate dal CGM (glicemia media, deviazione standard, TIR, TAR) e il valore di HbA1c nei pazienti T2D?
- Le metriche CGM (glicemia media, deviazione standard, TIR e TAR) permettono, nel loro insieme, di spiegare in modo significativo la variabilità osservata nei valori di HbA1c? Quali di queste metriche contribuiscono in modo indipendente alla previsione di HbA1c?
- Quali profili clinici caratterizzano i diversi gruppi di pazienti individuati tramite clustering dei segnali glicemici continui?
- Esiste un'associazione significativa tra il consumo di alcol e i valori di HbA1c nei pazienti con T2D?
- L'effetto del consumo di alcol rimane significativo anche controllando per età, sesso, BMI, fumo e tipo di terapia?
- L'eventuale associazione osservata è confusa da altre variabili?

L'analisi statistica mira dunque a estrarre insight clinicamente rilevanti, validare ipotesi e descrivere con rigore le dinamiche glicemiche osservate nei pazienti, con una particolare attenzione all'integrazione tra dati statici (clinico-anagrafici) e dinamici (CGM).

# 2 Analisi della relazione tra metriche CGM e HbA1c

Nel monitoraggio e nella gestione del T2D, la misura dell'emoglobina glicata (HbA1c) è una delle principali metriche cliniche di riferimento. Essa riflette l'andamento medio della glicemia negli ultimi 2–3 mesi e rappresenta un importante marcatore per la valutazione dell'efficacia della terapia. Con l'introduzione dei dispositivi di monitoraggio continuo del glucosio (CGM, Continuous Glucose Monitoring), è possibile ottenere informazioni molto più dettagliate e frequenti sull'andamento della glicemia, sia in termini di valore medio, sia in termini di variabilità intra-giornaliera.
Le metriche che sono state utilizzate per riassumere le informazioni derivate dal CGM sono:

- **Glicemia media**
- **Deviazione standard della glicemia**
- **TIR — Time In Range**
- **TAR — Time Above Range**

Il nostro obiettivo iniziale è stato quello di indagare se queste metriche fossero associate al valore di HbA1c. In particolare, abbiamo voluto testare se esistesse una relazione significativa tra ciascun indicatore derivato dal CGM e i valori di HbA1c osservati nei pazienti.

**🔍 ($RQ_1$). Esiste una correlazione tra le metriche derivate dal CGM (glicemia media, deviazione standard, TIR, TAR) e il valore di HbA1c nei pazienti con T2D?**

## 2.1 Preparazione dei dati

Come descritto nei capitoli precedenti, il dataset su cui si basa la nostra analisi include sia informazioni cliniche riepilogative sia misurazioni CGM ad alta frequenza per ciascun paziente. In questa sezione ci concentriamo sulle operazioni di pulizia e aggregazione necessarie per rendere i dati analizzabili nel contesto delle nostre research question.

**Importazione del file Summary.** La prima operazione ha riguardato il caricamento del file `Shanghai_T2DM_Summary.xlsx`. Sono stati selezionati solo i campi rilevanti all'analisi, ovvero:

- **Patient Number** rinominato in `PatientID`.
- **HbA1c (mmol/mol)**, rinominato in `HbA1c`.

Poiché in alcuni casi i valori di HbA1c risultavano codificati come testo, si è proceduto a una conversione esplicita in formato numerico per garantire la correttezza dell'analisi statistica successiva.

**Lettura dei file contenenti le misurazioni CGM.** Per ciascun file contenente i dati CGM dei singoli pazienti:

- È stato estratto l'identificativo del paziente, basato sul nome del file.
- La colonna che conteneva i valori glicemici, inizialmente identificata da nomi come "CGM (mg / dl)" è stata rinominata in `glicemia`, per migliorare la semplicità di scrittura.
- È stato effettuato il parsing della colonna temporale, assumendo il formato `%d/%m/%Y %H:%M`. Il timestamp è stato convertito in oggetto di tipo `POSIXct`, utile per le analisi successive basate sul tempo.

**Estrazione delle metriche CGM e unione con i dati clinici.** Dopo aver raccolto tutte le misurazioni glicemiche individuali per ciascun paziente, si è proceduto con la costruzione delle metriche riassuntive di interesse, utili per analizzare la relazione con il valore di HbA1c.

**Calcolo delle metriche di variabilità glicemica** Per ogni paziente è stato calcolato un insieme di indicatori derivati dalla serie temporale di valori glicemici. Le metriche estratte sono:

- **Glicemia media** (`mean_gl`): rappresenta l'andamento medio della glicemia sul periodo monitorato.
- **Varianza** (`var_gl`) e **Deviazione standard** (`sd_g`): misurano la variabilità intra-paziente della glicemia, indicando quanto oscillano i valori rispetto alla media.
- **TIR – Time In Range** (`TIR_70_1800`): proporzione di tempo in cui la glicemia si mantiene all'interno del range considerato ottimale (70–180 mg/dL).
- **TAR – Time Above Range** (`TAR_over180`): proporzione di tempo in cui la glicemia supera i 180 mg/dL, indice di iperglicemia.

Il risultato è stato un nuovo dataset contenente una riga per ciascun paziente e le cinque metriche CGM appena descritte.

**Integrazione con i dati clinic** Successivamente, le metriche CGM sono state unite al dataset contenente i valori di HbA1c, tramite un merge per `PatientID`. In questo modo, per ogni paziente è stato possibile accoppiare le informazioni dinamiche ricavate dal monitoraggio continuo con l'indicatore clinico statico di riferimento.

**Rimozione dei pazienti con dati incompleti** Durante il processo di fusione, è emerso che alcuni pazienti non presentavano un valore valido di HbA1c nel file summary. Poiché questo valore è essenziale per rispondere alla research question, tali pazienti sono stati esclusi dall'analisi.

Dopo il filtraggio dei valori `NA`, il dataset finale è risultato composto da un sottoinsieme dei pazienti originari, tutti caratterizzati da:

- Misurazioni glicemiche sufficienti per calcolare le metriche CGM
- Un valore valido di HbA1c

Questo dataset finale (`data_clean`) è stato utilizzato per tutte le analisi statistiche successive.

## 2.2 Analisi di correlazione tra metriche CGM e HbA1c

Una volta costruito il dataset finale, per indagare la relazione tra il controllo glicemico (ottenuto da CGM) e il valore di emoglobina glicata (HbA1c), è stata condotta un'analisi della correlazione lineare.

**Correlazione.** È stato inizialmente calcolato il coefficiente di correlazione di Pearson, che quantifica la forza e la direzione del legame lineare tra ciascuna metrica glicemica e HbA1c. I risultati ottenuti sono illustrati nella Tabella 1.

Table 1: Coefficiente di correlazione

| **Metrica** | **Coefficiente di Pearson (r)** |
|---|---|
| Glicemia media | 0.449 |
| Deviazione standard | 0.499 |
| Time In Range TIR | −0.333 |
| Time Above Range TAR | 0.406 |

Le correlazioni osservate mostrano tutte un'intensità moderata e si presentano con un segno coerente rispetto a quanto ci si aspetterebbe dal punto di vista clinico. In particolare, emerge che i pazienti con valori più elevati di glicemia media o con una maggiore variabilità glicemica — espressa dalla deviazione standard — tendono ad avere anche valori più alti di HbA1c. Invece, la correlazione tra HbA1c e TIR (Time In Range) è di segno opposto: si osserva infatti che all'aumentare della percentuale di tempo in cui la glicemia resta nel range considerato ottimale (70–180 mg/dL), i valori di HbA1c diminuiscono. Infine, la correlazione positiva tra HbA1c e TAR (Time Above Range) indica che i pazienti che trascorrono una quota maggiore del tempo in iperglicemia tendono ad avere HbA1c più elevata. Tuttavia, il coefficiente di correlazione da solo non è sufficiente per stabilire se l'associazione osservata sia significativa a livello statistico: è necessaria una verifica formale tramite test di ipotesi.

**Test di significatività.** Per ogni correlazione calcolata è stato eseguito un test di significatività tramite la funzione `cor.test()`. Oltre al p-value, `cor.test()` restituisce anche un intervallo di confidenza al 95% per il coefficiente di correlazione, che rappresenta l'intervallo plausibile in cui si può trovare il valore reale della correlazione nella popolazione. I risultati ottenuti sono riassunti nella Tabella 2.

Table 2: Significatività della correlazione

| **Metrica** | **r** | **p-value** | **IC95%** |
|---|---|---|---|
| Glicemia media | 0.449 | $2.48 \times 10^{-6}$ | [0.278, 0.592] |
| Deviazione standard | 0.499 | $1.07 \times 10^{-7}$ | [0.337, 0.633] |
| Time In Range TIR | $-0.333$ | $6.63 \times 10^{-4}$ | [-0.496, -0.147] |
| Time Above Range TAR | 0.406 | $2.52 \times 10^{-5}$ | [0.229, 0.557] |

Tutti i coefficienti di correlazione risultano statisticamente significativi, con p-value ampiamente inferiori alla soglia convenzionale di 0.05. Inoltre, tutti gli intervalli di confidenza escludono lo zero, rafforzando l'evidenza di un legame reale e non casuale tra le metriche CGM e i valori di HbA1c.

## 2.3 Modellazione multivariata: regressione lineare multipla

Le analisi di correlazione effettuate hanno messo in evidenza l'esistenza di una relazione statisticamente significativa tra ciascuna delle metriche CGM e i valori di HbA1c, considerati però singolarmente. Tuttavia, dal punto di vista statistico e clinico, è lecito ipotizzare che queste metriche interagiscano tra loro nel determinare il valore finale di HbA1c. Per questo motivo, si è deciso di adottare un approccio multivariato, attraverso l'utilizzo di un modello di regressione lineare multipla. In questo contesto, il valore di HbA1c viene modellato come funzione lineare delle metriche CGM precedentemente calcolate: glicemia media, deviazione standard, TIR e TAR.

**🔍 ($RQ_2$). Le metriche CGM (glicemia media, deviazione standard, TIR e TAR) permettono, nel loro insieme, di spiegare in modo significativo la variabilità osservata nei valori di HbA1c? Quali di queste metriche contribuiscono in modo indipendente alla previsione di HbA1c?**

**Visualizzazione delle relazioni tra HbA1c e le metriche CGM** Prima di procedere alla modellazione multivariata, è stata condotta un'analisi grafica preliminare per visualizzare la relazione tra HbA1c e ciascuna metrica derivata dal CGM. Sono stati realizzati scatterplot, visibili in Figura 1, Figura 2, Figura 3 e Figura 4, con sovrapposta la retta di regressione lineare semplice, al fine di valutare soprattutto la coerenza visiva con i risultati ottenuti dalla correlazione di Pearson.

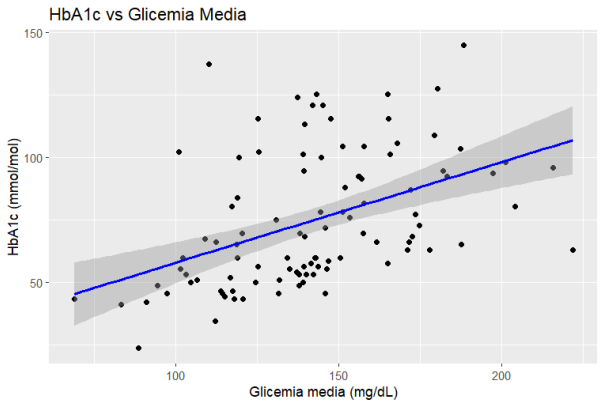


Figure 1: Glicemia media vs Hb1Ac

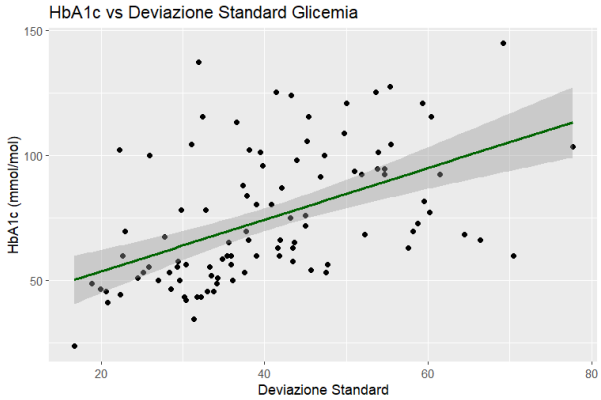


Figure 2: TIR vs Hb1Ac

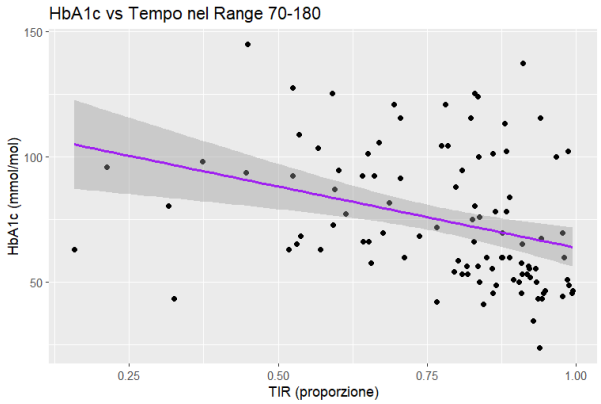


Figure 3: TIR vs Hb1Ac

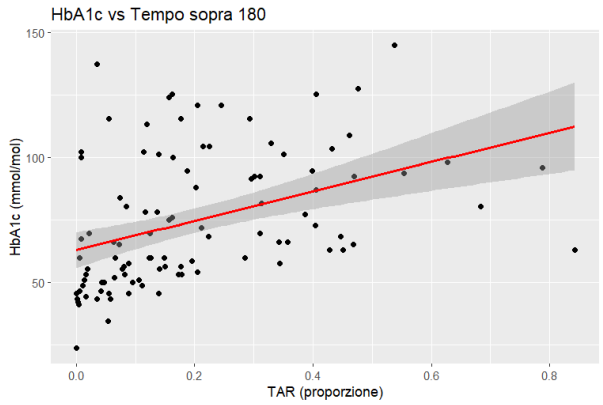


Figure 4: TAR vs Hb1Ac

L'andamento crescente o decrescente della retta nei diversi grafici riflette la direzione dell'associazione già evidenziata nei test precedenti.

**Verifica dei presupposti per l'applicazione della regressione lineare multipla** Dopo aver osservato associazioni significative tra le metriche CGM e HbA1c, sia a livello numerico (correlazioni) sia grafico (scatterplot), si è deciso di procedere con la stima di un modello di regressione lineare multipla, con l'obiettivo di valutare il contributo simultaneo di tutte le metriche CGM alla spiegazione della variabilità di HbA1c.

Prima di stimare il modello, è stato necessario valutare la possibilità teorica e statistica di applicare la regressione lineare multipla. Le ipotesi fondamentali da verificare sono:

- **Linearità**
- **Indipendenza degli errori**
- **Omoschedasticità**
- **Normalità degli errori**
- **Collinearità**

**Modello completo con tutte le metriche CGM** A partire da queste considerazioni, è stato stimato un primo modello includendo tutte le metriche CGM calcolate in precedenza. Il `summary` del modello è visibile nella Tabella 3.

Table 3: Risultati del modello di regressione multipla completo

| Variabile | Coefficiente | Errore Std. | t-value | p-value | Significativo |
|---|---|---|---|---|---|
| Intercetta | -9.5807 | 32.5473 | -0.294 | 0.7691 | No |
| `mean_gl` | 0.3523 | 0.2975 | 1.184 | 0.2392 | No |
| `sd_gl` | 0.7895 | 0.2459 | 3.210 | 0.0018 | Si |
| `TIR_70_180` | 8.9390 | 41.5968 | 0.215 | 0.8303 | No |
| `TAR_over180` | -23.1046 | 74.2969 | -0.311 | 0.7565 | No |
| **Statistiche globali del modello** | | | | | |
| Multiple $R^2$ | 0.2816 | | | | |
| Adjusted $R^2$ | 0.2517 | | | | |
| Errore standard dei residui | 22.98 (96 gradi di libertà) | | | | |
| F-statistic | 9.41 su 4 e 96 DF | | | | |
| $p$-value globale | $1.846 \times 10^{-6}$ modello significativo | | | | |

Il modello stimato spiega circa il 28% della variabilità di HbA1c. Il test F globale risulta altamente significativo ($p < 0.001$), indicando che almeno una delle variabili predittive incluse

nel modello contribuisce in modo rilevante alla spiegazione della variabile risposta. Analizzando i coefficienti individuali, si osservano le seguenti evidenze:

- Solo la deviazione standard della glicemia (`sd_gl`) risulta significativamente associata a HbA1c ($p = 0.0018$).
- La glicemia media (`mean_gl`), pur mostrando un coefficiente positivo, non raggiunge la significatività statistica.
- Le metriche `TIR_70_180` e `TAR_over180`, non risultano significative.
- Il coefficiente relativo a `TAR_over180` presenta inoltre un errore standard molto elevato e una stima instabile, rafforzando l'ipotesi della sua inadeguatezza nel modello attuale.

Nel complesso, il modello fornisce un primo quadro descrittivo interessante, ma le evidenze suggeriscono l'opportunità di procedere con specificazioni più parsimoniose, rimuovendo le variabili instabili e ridondanti.

**Diagnosi della collinearità: esclusione di TAR** Un'analisi della collinearità tramite il *Variance Inflation Factor* (VIF) ha evidenziato la presenza di forti relazioni lineari tra alcune variabili predittive. I risultati sono presenti nella Tabella 4.

Table 4: Valori del VIF per i predittori del modello completo

| Variabile | VIF |
|---|---|
| `mean_gl` | 14.733309 |
| `sd_gl` | 1.883912 |
| `TIR_70_180` | 10.758685 |
| `TAR_over180` | 35.293997 |

In particolare, `TAR_over180` mostra un VIF elevatissimo (35,3), seguito da `TIR_70_180` (10,8), valori superiori alla soglia critica di 10 generalmente indicativa di collinearità dannosa. Questi risultati sono coerenti con la natura complementare delle due variabili: TIR e TAR rappresentano porzioni percentuali di una distribuzione condivisa e quindi tendono a essere fortemente inversamente correlate. Per questo motivo, si è deciso di procedere con una nuova specificazione del modello, escludendo la variabile `TAR_over180`, al fine di ridurre la collinearità e migliorare la stabilità del modello stimato.

**Modello senza `TAR_over180`** Per migliorare la stabilità del modello e ridurre la collinearità tra i predittori, si è stimata una nuova specificazione escludendo la variabile `TAR_over180`. I risultati sono riportati nella Tabella 5.

Table 5: Risultati del modello di regressione multipla senza `TAR_over180`

| Variabile | Coefficiente | Errore Std. | t-value | p-value | Significativo |
|---|---|---|---|---|---|
| Intercetta | -11.4285 | 31.8509 | -0.359 | 0.7205 | No |
| `mean_gl` | 0.2696 | 0.1326 | 2.033 | 0.0448 | Sì |
| `sd_gl` | 0.7951 | 0.2441 | 3.257 | 0.0016 | Sì |
| `TIR_70_180` | 20.1644 | 20.5744 | 0.980 | 0.3295 | No |
| **Statistiche globali del modello** | | | | | |
| Multiple $R^2$ | 0.2809 | | | | |
| Adjusted $R^2$ | 0.2587 | | | | |
| Errore standard dei residui | 22.87 (97 gradi di libertà) | | | | |
| F-statistic | 12.63 su 3 e 97 DF | | | | |
| $p$-value globale | $4.87 \times 10^{-7}$ modello significativo | | | | |

Il nuovo modello spiega circa il 28% della variabilità osservata nei valori di HbA1c, in linea con la specificazione precedente. Il test F globale è altamente significativo, a conferma dell'utilità complessiva del modello.

Dal punto di vista dei singoli coefficienti:

- La deviazione standard della glicemia (`sd_gl`) si conferma significativamente associata ad HbA1c, con un effetto positivo stabile e rilevante.
- Anche la glicemia media (`mean_gl`) risulta ora statisticamente significativa ($p = 0.0448$), suggerendo che entrambi gli aspetti della distribuzione glicemica (media e variabilità) contribuiscono in modo indipendente alla previsione di HbA1c.
- Al contrario, la metrica `TIR_70_180` rimane non significativa nel nuovo modello, rafforzando l'ipotesi che la sua informazione sia in parte assorbita dalle altre metriche.

**Collinearità nel modello semplificato** L'analisi dei *Variance Inflation Factor* (VIF) conferma il miglioramento ottenuto con l'esclusione della variabile `TAR_over180`. I valori sono riportati nella Tabella 6.

Table 6: Valori del VIF per i predittori del modello senza `TAR_over180`

| Variabile | VIF |
|---|---|
| `mean_gl` | 2.955166 |
| `sd_gl` | 1.873851 |
| `TIR_70_180` | 2.656800 |

Tutti i valori risultano ampiamente inferiori alla soglia (10), indicando un basso livello di collinearità tra i predittori e una buona stabilità delle stime.

**Valutazione del modello con esclusione di `TIR_70_180`** È stata valutata anche la possibilità di costruire un modello ulteriormente semplificato, includendo soltanto `mean_gl` e la deviazione standard `sd_gl`. Tale specificazione ha mostrato una capacità esplicativa quasi identica rispetto al modello con `TIR_70_180`, con un *Adjusted* $R^2$ pari a 0.259 contro 0.2587. Tuttavia, l'eliminazione di `TIR_70_180` ha comportato un peggioramento del livello di significatività statistica del coefficiente associato a `mean_gl`, il cui $p$-value è salito da 0.0448 a 0.0751, perdendo così la soglia convenzionale di significatività. Nonostante `TIR_70_180` non sia statisticamente significativo nel modello, la sua presenza sembra favorire una stima più precisa e più stabile dell'effetto della glicemia media. Inoltre, trattandosi di una metrica clinicamente rilevante e priva di collinearità dannosa, si è deciso di mantenerla nel modello finale.

**Verifica della normalità dei residui e proposte di miglioramento** Una delle ipotesi della regressione lineare è la normalità dei residui. Per verificarla, è stato applicato il test di Shapiro–Wilk sui residui del modello contenente `sd_gl`, `mean_gl` e `TIR_70_180`. Il valore di $p$ ottenuto è pari a $5.86 \times 10^{-5}$ e, poiché risulta inferiore alla soglia convenzionale di 0.05, si rifiuta l'ipotesi nulla di normalità. In altre parole, vi è evidenza statistica che i residui del modello non seguono una distribuzione normale. In prospettiva futura, per migliorare ulteriormente l'aderenza del modello ai presupposti teorici e rafforzare l'affidabilità statistica delle inferenze, si potrebbero implementare:

- trasformazioni delle variabili (es. logaritmica, Box–Cox) al fine di migliorare la distribuzione dei residui
- metodi di regressione robusta (es. Huber o MM-estimators), meno sensibili a valori anomali o violazioni delle assunzioni classiche

# 3 Identificazione e caratterizzazione di profili glicemici tramite clustering

Il monitoraggio continuo della glicemia (CGM) rappresenta una risorsa preziosa per cogliere le dinamiche glicemiche nei pazienti con T2D. Sebbene due pazienti possano avere valori medi di glicemia simili, le loro fluttuazioni quotidiane e settimanali possono essere molto diverse, riflettendo differenze nella risposta alla terapia, nello stile di vita o nella progressione della malattia. In questa parte del lavoro ci siamo posti l'obiettivo di indagare se sia possibile individuare, a partire dalle serie temporali di glicemia, dei gruppi omogenei di pazienti. Successivamente, abbiamo voluto capire se questi gruppi si differenziano anche per alcune caratteristiche cliniche misurate separatamente, come l'emoglobina glicata (HbA1c), l'albumina glicata (GA), la presenza di ipoglicemie o l'abitudine al consumo di alcol.

**(RQ$_3$). Quali profili clinici caratterizzano i diversi gruppi di pazienti individuati tramite clustering dei segnali glicemici continui?**

## 3.1 Lettura e preparazione dei dati

Il caricamento dei dati e la costruzione delle serie temporali glicemiche sono stati effettuati seguendo lo stesso procedimento descritto nel Paragrafo 2.1.

## 3.2 Costruzione delle serie temporali regolari

Una volta importate e pulite le misurazioni glicemiche, per ciascun paziente abbiamo costruito una serie temporale regolare con intervalli di 15 minuti. Questo passaggio è stato fondamentale per garantire l'allineamento temporale dei dati tra pazienti diversi e per rendere le serie confrontabili tra loro.

Inizialmente, sono state rimosse tutte le righe contenenti valori mancanti nella data o nella glicemia. Successivamente, la colonna contenente le date è stata convertita in oggetti temporali standard R (`POSIXct`), cercando di interpretare automaticamente i formati temporali. Il formato principale utilizzato è stato quello ISO (`"YYYY-MM-DD HH:MM:SS"`), e, in pochi casi, è stato adottato un formato alternativo europeo del tipo `"DD/MM/YYYY HH:MM"`. Dopo aver scartato le osservazioni con timestamp non interpretabili, i dati sono stati ordinati cronologicamente. A partire dal primo e dall'ultimo timestamp disponibili per ciascun paziente, è stata generata una griglia temporale regolare con passo di 15 minuti. Su questa griglia è stata quindi interpolata la serie glicemica, mediante interpolazione lineare, costruendo un oggetto `zoo` contenente i valori glicemici corrispondenti ai nuovi timestamp regolari. Gli oggetti `zoo` sono una struttura particolarmente utile in R per rappresentare serie temporali indicizzate da date o orari. Nel nostro caso, hanno permesso di trattare in modo flessibile le misurazioni CGM dei pazienti. Infine, i pazienti per cui non è stato possibile costruire una serie regolare (a causa di un numero troppo basso di dati validi) sono stati esclusi, mantenendo soltanto le serie temporali complete nella lista `ts_list`, che costituirà la base per le analisi successive.

## 3.3 Analisi wavelet e costruzione delle feature

Dopo aver ottenuto per ciascun paziente una serie temporale regolare della glicemia, abbiamo proceduto ad analizzarne la struttura interna utilizzando la *Discrete Wavelet Transform* (DWT), una tecnica molto utile per decomporre un segnale in componenti che catturano la variazione locale a diverse scale temporali. Questo approccio è particolarmente indicato per analizzare segnali come quelli glicemici, che presentano sia oscillazioni lente che variazioni improvvise.

**Trasformata wavelet discreta (DWT)** Per ciascuna serie temporale, è stata applicata una trasformata wavelet discreta utilizzando la funzione `dwt()` del pacchetto `wavelets`. Sono stati utilizzati i seguenti parametri:

- `filter = "la8"`: corrisponde al filtro wavelet Daubechies con 8 coefficienti (valore predefinito)
- `n.levels = J`: il numero massimo di livelli, con un limite massimo di 4 (valore predefinito)
- `boundary = "reflection"`: è stato scelto il metodo di estensione `reflection` per trattare i bordi del segnale, aal posto di `"periodic"`, che assume che il segnale si ripeta ciclicamente.

**Costruzione della matrice delle feature** Il risultato della DWT è una serie di vettori:

- i coefficienti di *dettaglio* $(D_1, D_2, \ldots, D_J)$, che rappresentano le variazioni del segnale più brusche
- un vettore di *approssimazione* finale $(A_J)$, che sintetizza l'andamento più lento e globale della serie

**Scelta della rappresentazione: media e varianza dei coefficienti wavelet** Nel contesto dell'analisi, l'obiettivo era costruire una matrice di caratteristiche (feature matrix) con una rappresentazione numerica comparabile tra pazienti. Tuttavia, una criticità emersa durante la fase di sviluppo è stata la lunghezza variabile delle serie temporali dei pazienti: poiché ogni serie può avere durata e granularità differenti, anche la decomposizione wavelet produce un numero diverso di livelli e quindi vettori di coefficienti di lunghezza non uniforme. Questo impedisce l'uso diretto dei coefficienti grezzi per il clustering, in quanto non è possibile trasformarli in una matrice regolare. Una prima soluzione prevedeva l'estrazione di tutti i coefficienti solo se ogni serie avesse restituito vettori della stessa lunghezza. Questa condizione non era verificata nei nostri dati. Abbiamo quindi adottato un approccio più stabile e raccomandato: la riduzione di ciascun vettore wavelet a statistiche di sintesi a lunghezza fissa. In particolare, per ogni livello di decomposizione $j = 1, \ldots, J$ sono state calcolate due statistiche riassuntive:

- la media dei coefficienti $\mu_{D_j} = \text{mean}(D_j)$
- la varianza dei coefficienti $\sigma^2_{D_j} = \text{var}(D_j)$

Lo stesso è stato fatto per il livello di approssimazione finale. In questo modo, ogni paziente è rappresentato da un vettore di lunghezza fissa pari a $2 \cdot (J + 1)$, garantendo compatibilità e confrontabilità tra osservazioni. Questa strategia ha permesso di mantenere la massima quantità di informazione disponibile, pur assicurando una struttura matriciale coerente per le analisi successive.

**Normalizzazione** Le feature sono state standardizzate tramite normalizzazione Z-score, così da avere media 0 e deviazione standard 1, per garantire che tutte le variabili contribuiscano equamente al calcolo delle distanze durante il clustering, evitando che scale diverse distorcano i risultati.

**Scelta della metrica di distanza e del numero di cluster** Per eseguire il clustering dei pazienti, è stato preliminarmente valutato quale metrica di distanza fosse più adatta a rappresentare la similarità tra i profili glicemici. Sono state considerate tre metriche:

- Distanza **Euclidea**;

- Distanza **Manhattan**;
- Distanza basata su **1 - correlazione di Pearson**.

Per ciascuna di queste distanze è stato calcolato il valore medio dell'indice di silhouette (*Sil*) per un numero di cluster variabile tra 2 e 10. L'indice di silhouette è una misura della qualità del clustering, e un valore più elevato indica una separazione più netta tra i gruppi. Il numero di cluster ottimale è stato scelto come quello che massimizzava la silhouette media. I risultati sono riportati nella Tabella 7.

Table 7: Confronto tra metriche di distanza: numero ottimale di cluster e silhouette media

| **Metrica di distanza** | **Cluster ottimale (k)** | **Silhouette media** |
|---|---|---|
| Euclidea | 2 | 0.1582 |
| Manhattan | 2 | 0.1930 |
| 1 - Correlazione | 3 | 0.3516 |

Come evidente, la distanza basata su correlazione ha ottenuto la silhouette media più elevata, suggerendo che è la più efficace nel catturare le somiglianze nei profili glicemici. Per questo motivo, è stata scelta la `1 - correlazione di Pearson` come metrica finale di distanza per la costruzione della matrice di dissimilarità.

### 3.4 Clustering

Dopo aver selezionato come distanza ottimale `1 - correlazione di Pearson` (Tabella 7), è stato applicato il clustering dei pazienti mediante due tecniche distinte:

- **PAM (Partitioning Around Medoids)**, un metodo robusto rispetto agli outlier e particolarmente adatto a lavorare con matrici di dissimilarità
- **Clustering gerarchico agglomerativo**, con metodo di linkage di *Ward.D2*, seguito da taglio del dendrogramma a k = 3 gruppi.

Il valore di k=3 è stato scelto in base alla silhouette media massima ottenuta con la distanza di correlazione.

**Cluster assegnati** I pazienti sono stati assegnati ai gruppi come riportato nella Tabella 8.

Table 8: Distribuzione dei pazienti nei cluster PAM e Gerarchico (con k=3)

| **Metodo** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|
| PAM | 26 | 47 | 36 |
| Gerarchico | 41 | 44 | 24 |

**Dendrogramma del clustering gerarchico** Per visualizzare la struttura dei gruppi identificati tramite il clustering gerarchico, è stato tracciato il seguente dendrogramma:

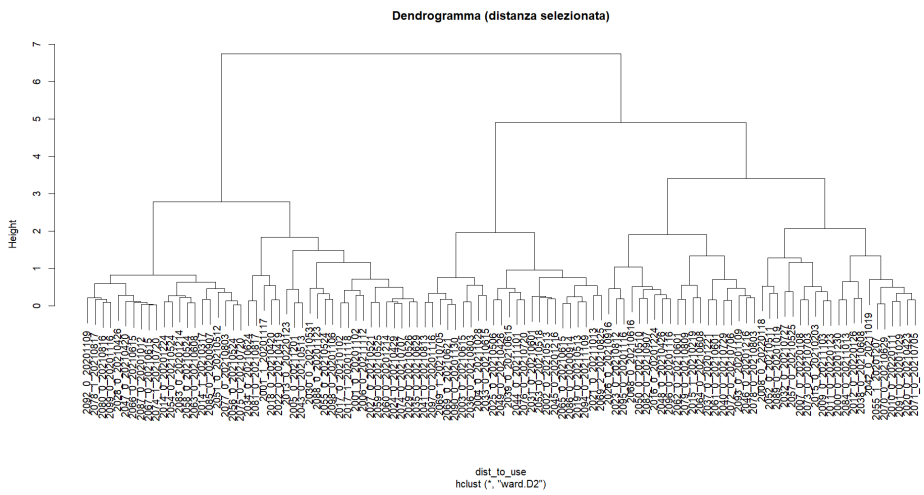


Figure 5: Dendrogramma.

**Scelta finale** Sebbene entrambi i metodi abbiano restituito suddivisioni coerenti, si è scelto di adottare la partizione risultante dal **clustering gerarchico**, in quanto più stabile e visivamente interpretabile attraverso il dendrogramma.

### 3.5 Analisi dei cluster: confronto con variabili cliniche

Dopo aver assegnato i pazienti a tre cluster utilizzando un algoritmo gerarchico, abbiamo iniziato a confrontare i gruppi rispetto ad alcune variabili cliniche disponibili nel file `Shanghai_T2DM_Summary.xlsx`.

**Variabili categoriche: ipoglicemia e consumo di alcol** Per confrontare la distribuzione di variabili categoriche tra i cluster (ipoglicemia e consumo di alcol), abbiamo utilizzato il test del $\chi^2$. Di seguito sono riportate le tabelle di contingenza:

Table 9: Distribuzione dell'ipoglicemia per cluster

| Cluster | No | Yes |
|---|---|---|
| 1 | 37 | 4 |
| 2 | 39 | 5 |
| 3 | 23 | 1 |

Table 10: Distribuzione del consumo di alcol per cluster

| Cluster | Drinker | Non-drinker |
|---|---|---|
| 1 | 6 | 35 |
| 2 | 5 | 39 |
| 3 | 2 | 22 |

In entrambi i casi, i test del $\chi^2$ non risultano significativi ($p = 0.609$ e $p = 0.743$), suggerendo che la distribuzione di queste variabili non differisce significativamente tra i gruppi.

**Albumina glicata (GA)** Per la variabile continua `Glycated Albumin (GA)`, abbiamo prima verificato l'assunzione di normalità all'interno di ciascun cluster tramite il test di Shapiro–Wilk.

Nessuno dei tre cluster ha mostrato distribuzione normale (tutti i p-value $< 0.05$), per cui si è optato per il test di Kruskal–Wallis. L'uso di questo test è giustificato dal fatto che, a differenza dell'ANOVA, il Kruskal–Wallis non richiede la normalità dei dati nei gruppi, risultando quindi più appropriato in presenza di violazioni di questa ipotesi. Il test ha restituito un risultato significativo ($p = 0.0021$), suggerendo che almeno un gruppo differisce dagli altri. Per identificare i confronti significativi, è stato applicato il test post-hoc di Dunn con correzione di Bonferroni:

Table 11: Confronti multipli su GA (test di Dunn con Bonferroni)

| **Confronto** | **Z** | **p (non corretto)** | **p (Bonferroni)** |
|---|---|---|---|
| 1 vs 2 | 1.41 | 0.159 | 0.479 |
| 1 vs 3 | -2.24 | 0.025 | 0.074 |
| 2 vs 3 | -3.50 | 0.00047 | **0.0014** |

Solo il confronto tra i cluster 2 e 3 risulta significativamente diverso anche dopo la correzione ($p < 0.05$). Questo indica che i pazienti del cluster 3 presentano valori di albumina glicata significativamente diversi rispetto a quelli del cluster 2.

**Emoglobina glicata (HbA1c)** Anche per `HbA1c (mmol/mol)` è stata condotta una verifica preliminare di normalità. Solo il cluster 3 ha mostrato distribuzione compatibile con la normalità ($p = 0.117$), mentre gli altri due no. Pertanto, è stato applicato nuovamente il test di Kruskal–Wallis. Il test ha restituito un risultato significativo ($p = 0.00896$), e anche in questo caso si è proceduto con il test post-hoc di Dunn:

Table 12: Confronti multipli su HbA1c (test di Dunn con Bonferroni)

| **Confronto** | **p (Bonferroni)** |
|---|---|
| 1 vs 2 | 1.0000 |
| 1 vs 3 | **0.0397** |
| 2 vs 3 | **0.0094** |

Il cluster 3 differisce in modo significativo da entrambi gli altri gruppi, confermando quanto già osservato per la GA. Questo rafforza l'evidenza che i pazienti nel cluster 3 presentano un profilo glicemico cronico diverso, sia in termini di GA che di HbA1c.

### 3.6 Conclusioni

Nonostante i cluster ottenuti non definiscano in modo netto un unico profilo clinico dominante, l'analisi condotta ha evidenziato comunque delle differenze rilevanti tra i gruppi. In particolare:

- Le variabili continue `HbA1c` e `GA`, entrambe indicatori fondamentali del controllo glicemico nel lungo periodo, hanno mostrato differenze statisticamente significative tra cluster.
- Il **cluster 3** si è distinto in modo marcato rispetto agli altri due, mostrando valori di glicemia cronica significativamente diversi.
- Al contrario, variabili categoriali come l'ipoglicemia e il consumo di alcol non hanno mostrato differenze significative tra i gruppi, suggerendo che potrebbero non essere fattori discriminanti nella segmentazione dei pazienti basata su pattern glicemici.

Dal punto di vista metodologico:

- L'utilizzo della trasformata wavelet e l'estrazione di statistiche sintetiche ha permesso di ridurre la complessità delle serie temporali mantenendone l'informazione essenziale.
- L'approccio al clustering su misura di distanza 1–correlazione ha restituito una struttura interpretabile e coerente con i pattern glicemici osservati.

In sintesi, pur non identificando un singolo "profilo tipo" per ciascun cluster, lo studio ha dimostrato la possibilità di distinguere sottogruppi clinicamente diversi sulla base di metriche di variabilità glicemica e valori di glicemia cronica.

# 4 Analisi dell'effetto del consumo di alcol sui valori di HbA1c

Nel contesto della gestione del diabete di tipo 2, la misura dell'emoglobina glicata (HbA1c) rappresenta una delle principali metriche cliniche per valutare il controllo glicemico nel lungo periodo. Riflettendo l'andamento medio della glicemia nelle 8-12 settimane precedenti, l'HbA1c consente di monitorare l'efficacia della terapia e l'aderenza del paziente al trattamento.

Oltre ai parametri fisiologici e clinici, anche **comportamenti e abitudini di vita** possono influenzare significativamente il valore dell'HbA1c. Tra questi, il **consumo di alcol** è oggetto di studio per il suo possibile ruolo nella modulazione della sensibilità insulinica e nel controllo glicemico.

In questa sezione ci proponiamo di valutare l'associazione tra il consumo di alcol e i valori di HbA1c nei pazienti con diabete di tipo 2, tramite approcci sia **univariati** che **multivariati**, tenendo conto di potenziali variabili confondenti.

Per approfondire il ruolo del consumo di alcol nei valori di HbA1c, ci poniamo le seguenti domande:

- **(RQ$_4$). Esiste un'associazione significativa tra il consumo di alcol e i valori di HbA1c nei pazienti con diabete di tipo 2?**
- **(RQ$_5$). L'effetto del consumo di alcol rimane significativo anche controllando per età, sesso, BMI, fumo e tipo di terapia?**
- **(RQ$_6$). L'eventuale associazione osservata è confusa da altre variabili?**

## 4.1 Preparazione dei dati

Abbiamo utilizzato il dataset `Shanghai_T2DM_Summary.xlsx`, che contiene informazioni anagrafiche, comportamentali, cliniche e terapeutiche di oltre 100 pazienti affetti da diabete di tipo 2. Le principali variabili utilizzate sono:

- `HbA1c`: emoglobina glicata (mmol/mol);
- `Alcohol Drinking History`: consumo di alcol (drinker/non-drinker);
- `Age`, `BMI`, `Smoking History`, `Gender`, `Hypoglycemic Agents`.

**Pulizia e trasformazione**

- Sono stati standardizzati i valori mancanti (es. "none", "-", "/", "NA") convertendoli in veri NA.
- Le variabili numeriche (HbA1c, Age, BMI, Smoking) sono state convertite in `numeric`, gestendo le virgole come separatori decimali.
- Le variabili categoriali (Alcohol, Gender, Terapia) sono state convertite in `factor`.
- È stata costruita la variabile `Terapia_cat` con quattro categorie:
  - Solo Metformin
  - Solo Insulina
  - Insulina + Orali
  - Orali (senza insulina né metformina)

## 4.2 Analisi univariata: effetto del solo consumo di alcol

Abbiamo inizialmente valutato se i valori di HbA1c differissero tra pazienti che consumano alcol e quelli che non lo fanno.

- Il test di normalità (Shapiro-Wilk) ha mostrato distribuzione non normale nel gruppo *non-drinker* ($p = 0.0001$), ma normale nei *drinker* ($p = 0.8661$);
- È stato applicato il test di Wilcoxon ($p = 0.0107$), che ha mostrato una differenza significativa;
- Mediane:
  - Drinker: 94.5 mmol/mol (media 91.1)
  - Non-drinker: 63.9 mmol/mol (media 72.2)

Il boxplot evidenzia valori più alti di HbA1c nel gruppo "drinker". Tuttavia, questo effetto potrebbe essere confuso da altre variabili cliniche e comportamentali.

## 4.3 Analisi multivariata: regressione lineare

Abbiamo stimato il modello:

```
lm(HbA1c ~ Age + BMI + Smoking + Gender + Alcohol + Terapia_cat)
```

- L'effetto dell'alcol non è più significativo ($p = 0.252$);
- Solo la categoria "Solo Insulina" risulta significativa ($p = 0.01175$), con aumento di circa 24 mmol/mol;
- $R^2 = 0.2175$, quindi il modello spiega il 21.7% della variabilità.

## 4.4 Diagnostica del modello

- **Normalità dei residui**: non verificata ($p < 0.001$);
- **Omocedasticità**: verificata ($p = 0.985$);
- **Autocorrelazione**: non significativa ($p = 0.073$);
- Dopo la rimozione delle osservazioni influenti (Cook's Distance), i risultati rimangono stabili.

## 4.5 Confronto tra modelli: evidenza di confounding

- **Modello semplice** (HbA1c $\sim$ Alcohol): coeff. $= -18.6$, $p = 0.025$
- **Modello completo** (HbA1c $\sim$ Age + BMI + Smoking + Gender + Alcohol + Terapia_cat): coeff. $= -12.2$, $p = 0.252$

Il calo della significatività conferma la presenza di confounding da parte di altre variabili, in particolare la terapia.

## 4.6 Conclusioni

- Il consumo di alcol appare associato a valori più alti di HbA1c solo in analisi univariata;
- In presenza di altre variabili, l'effetto dell'alcol non è più significativo;
- La categoria "Solo Insulina" è fortemente associata a valori più alti di HbA1c;

# 5 Validazione di un Dataset Sintetico Generato con LLM

## 5.1 Obiettivo e Contesto

L'obiettivo di questo studio è utilizzare un **Large Language Model** (LLM), nello specifico GPT-4, per generare un dataset sintetico che riproduca le caratteristiche di un dataset reale composto da pazienti affetti da diabete di tipo 2 (T2DM). L'intento è validare statisticamente tale dataset sintetico, verificando che:

- la struttura e le distribuzioni delle variabili siano rispettate;
- le relazioni cliniche tra le variabili siano mantenute;
- la frequenza delle categorie e la percentuale dei dati mancanti siano riprodotte correttamente;
- il formato e i tipi di dato siano coerenti.

Il dataset reale di riferimento è `Shanghai_T2DM_Summary.xlsx`, da cui sono stati estratti i parametri statistici e le regole cliniche usate per la generazione.

## 5.2 Metodologia

**Selezione e Preprocessing del Dataset** Il dataset reale contiene informazioni su 109 pazienti, ciascuno descritto tramite variabili numeriche e categoriali. Sono state calcolate medie, deviazioni standard, minimi, massimi e frequenze modali per ciascuna variabile, e sono state definite relazioni cliniche interne da rispettare (es. $eGFR \sim 1/\text{Creatinina}$, $BMI = \text{Peso}/\text{Altezza}^2$, ecc.).

**Prompt Engineering e Generazione** È stato progettato un prompt dettagliato da fornire al modello GPT-4, specificando:

- range e medie attese per le variabili numeriche;
- etichette e frequenze per le variabili categoriali;
- regole di coerenza interna (es. età $\geq$ durata diabete);
- formato dell'output richiesto (CSV o JSON).

Il prompt completo è riportato in **Appendice A**. Il dataset sintetico prodotto è stato poi salvato e confrontato con l'originale per la validazione.

## 5.3 Validazione Statistica

**Variabili Numeriche** Sono stati confrontati media, deviazione standard, minimo e massimo per ciascuna variabile. Di seguito la tabella completa dei risultati ottenuti:

Table 13: Statistiche descrittive (Reale vs Sintetico)

| Variabile | Media (R) | SD (R) | Media (S) | SD (S) | Min (R) | Max (R) | Min (S) | Max (S) |
|---|---|---|---|---|---|---|---|---|
| Weight (kg) | 66.29 | 12.01 | 68.39 | 14.67 | 40.00 | 100.00 | 40.00 | 100.00 |
| Smoking History (pack year) | 3.56 | 11.66 | 8.69 | 12.60 | 0.00 | 80.00 | 0.00 | 51.00 |
| Fasting C-peptide (nmol/L) | 0.47 | 0.27 | 0.48 | 0.24 | 0.04 | 1.24 | 0.04 | 0.97 |
| Fasting Insulin (pmol/L) | 111.99 | 257.09 | 278.48 | 337.60 | 10.73 | 2089.80 | 10.70 | 1639.00 |
| Gender (Female=1, Male=2) | 1.54 | 0.50 | 1.50 | 0.50 | 1.00 | 2.00 | 1.00 | 2.00 |
| 2-hour Postprandial C-peptide (nmol/L) | 0.96 | 0.72 | 1.14 | 0.90 | 0.15 | 4.42 | 0.15 | 4.13 |
| Age (years) | 60.30 | 14.01 | 61.81 | 16.67 | 22.00 | 97.00 | 22.00 | 97.00 |
| Creatinine ($\mu$mol/L) | 63.83 | 21.23 | 65.25 | 26.94 | 27.70 | 136.10 | 28.90 | 116.40 |
| Duration of diabetes (years) | 8.69 | 8.26 | 10.28 | 8.59 | 0.01 | 40.00 | 0.00 | 35.13 |
| Triglyceride (mmol/L) | 1.79 | 1.08 | 1.82 | 1.37 | 0.61 | 7.65 | 0.61 | 5.78 |
| HbA1c (mmol/mol) | 74.70 | 26.57 | 74.74 | 28.88 | 23.50 | 144.80 | 23.46 | 144.55 |
| Height (m) | 1.66 | 0.10 | 1.66 | 0.11 | 1.42 | 1.88 | 1.42 | 1.87 |
| eGFR (mL/min/1.73 $m^2$) | 116.66 | 42.18 | 118.59 | 43.20 | 34.00 | 257.00 | 43.99 | 212.53 |
| BMI (kg/$m^2$) | 24.06 | 3.31 | 24.04 | 6.60 | 17.09 | 36.73 | 15.21 | 30.52 |
| LDL Cholesterol (mmol/L) | 3.12 | 1.00 | 3.14 | 1.02 | 0.98 | 5.27 | 1.10 | 4.92 |
| Total Cholesterol (mmol/L) | 4.84 | 1.16 | 4.79 | 1.28 | 2.51 | 7.79 | 2.73 | 6.93 |
| 2-hour Postprandial Glucose (mg/dL) | 264.76 | 96.04 | 271.35 | 118.91 | 97.00 | 610.40 | 97.00 | 544.50 |
| Glycated Albumin (%) | 24.00 | 8.66 | 23.98 | 9.76 | 7.10 | 50.90 | 8.00 | 43.80 |

**Osservazioni**:

- Le medie e le dispersioni risultano coerenti per la quasi totalità delle variabili.
- Alcune variabili mostrano lievi scostamenti (es. *Fasting Insulin*, *Smoking History*), ma restano nel range clinicamente plausibile.
- Le distribuzioni complessive risultano compatibili.

**Variabili Categoriali** Sono state confrontate moda, frequenza della moda e numero di modalità. Di seguito i principali risultati:

Table 14: Statistiche categoriali (Moda e Frequenza)

| Variabile | Moda (R) | Moda (S) | Freq (R) | Freq (S) |
|---|---|---|---|---|
| Alcohol | non-drinker | non-drinker | 96 | 96 |
| Hypoglycemia | no | no | 99 | 99 |
| Hypoglycemic Agents | metformin | insulin asp. | 9 | 19 |
| Comorbidities | none | hypertension | 39 | 10 |

**Osservazioni**:

- Le modalità principali sono coerenti nella maggior parte dei casi.
- Alcune discrepanze (es. moda nei farmaci) sono state rilevate, ma non invalidano il dataset in quanto la distribuzione complessiva resta fedele.

## 5.4 Discussione

Il confronto tra dati reali e sintetici ha evidenziato una forte somiglianza in termini di media, dispersione e valori estremi. Le relazioni cliniche (età $\geq$ durata diabete, BMI corretto, eGFR coerente con creatinina) sono state rispettate. Le variabili categoriali conservano una distribuzione plausibile, con valori mancanti simulati correttamente. Uniche criticità rilevate:

- Scostamento della media in *Fasting Insulin* e *Smoking History*.
- Differenze di moda in alcune variabili categoriali.

Tali differenze non compromettono la qualità complessiva del dataset sintetico.

### 5.5 Conclusioni

Il dataset sintetico generato tramite LLM (GPT-4) è statisticamente valido e rappresenta fedelmente la struttura e le relazioni del dataset reale, tranne in alcuni casi che non risultano essere critici ai fini della validità statistica della generazione. Il processo richiede molto effort preliminare per la preparazione di tutti i range di valori in modo da fornire all'LLM ogni minimo dettaglio, sia in termini statistici che di consistenza dei dati. Nonostante non sia emersa semplicità nel processo di generazione dei dati sintetici, il risultato finale risulta comunque essere ampiamente positivo, dimostrando come, con il giusto utilizzo, gli LLM possano essere utilizzati in maniera efficace per la creazione di dataset sintetici, utili per lavori di tipo statistico.

## A Prompt Utilizzato

Genera un dataset sintetico di 109 pazienti affetti da diabete di tipo 2, che riproduca fedelmente la struttura e le distribuzioni osservate nel dataset reale. Ogni paziente deve essere rappresentato da una riga tabellare con le seguenti variabili:

**1. Variabili Numeriche**

- Gender (1 = Female, 2 = Male): [1, 2], media $\approx$ 1.54
- Age (years): [22, 97], media $\approx$ 60.3
- Height (m): [1.415, 1.88], media $\approx$ 1.66
- Weight (kg): [40.0, 100.0], media $\approx$ 66.3
- BMI (kg/m$^2$): [17.09, 36.73], media $\approx$ 24.1
- Smoking History (pack-year): [0.0, 80.0], media $\approx$ 3.56
- Duration of diabetes (years): [0.005, 40.0], media $\approx$ 8.7
- Fasting Plasma Glucose (mg/dL): [55.8, 432.0], media $\approx$ 164.9
- 2-hour Postprandial Glucose (mg/dL): [97.0, 610.4], media $\approx$ 264.8
- Fasting C-peptide (nmol/L): [0.04, 1.24], media $\approx$ 0.47
- 2-hour Postprandial C-peptide (nmol/L): [0.15, 4.42], media $\approx$ 0.96
- Fasting Insulin (pmol/L): [10.7, 2089.8], media $\approx$ 112.0
- HbA1c (mmol/mol): [23.5, 144.8], media $\approx$ 74.7
- Glycated Albumin (%): [7.1, 50.9], media $\approx$ 24.0
- Total Cholesterol (mmol/L): [2.51, 7.79], media $\approx$ 4.84
- Triglyceride (mmol/L): [0.61, 7.65], media $\approx$ 1.79
- HDL Cholesterol (mmol/L): [0.69, 2.56], media $\approx$ 1.13
- LDL Cholesterol (mmol/L): [0.98, 5.27], media $\approx$ 3.12
- Creatinine ($\mu$mol/L): [27.7, 136.1], media $\approx$ 63.8
- eGFR (mL/min/1.73 m$^2$): [34.0, 257.0], media $\approx$ 116.7

2. **Variabili Categoriali**
   - Alcohol Drinking History: *non-drinker* (96), *drinker* (13)
   - Comorbidities: multipli valori possibili, ∼36% mancanti
   - Hypoglycemic Agents: es. *metformin*, ∼5% mancanti
   - Other Agents: es. *aspirin*, ∼39% mancanti
   - Hypoglycemia: *yes* (10), *no* (99)
3. **Regole di Coerenza**
   - Età $\geq$ Durata del diabete
   - BMI = Peso / Altezza$^2$
   - eGFR inversamente correlato con creatinina
   - Valori glicemici coerenti con HbA1c e C-peptide
   - Chi ha solo *metformin* tende ad avere HbA1c più bassa
   - Formato dei missing values: `null` o `NaN`
4. **Output** Restituisci il dataset come `CSV` o `JSON`, con intestazioni identiche al dataset reale.