

Dati Sanitari Sintetici: Compromesso tra Privacy e Utilità nella Ricerca Medica

Autori:

Sara Casadio
Noemi Ferrara
Giorgia Pirelli

Anno accademico 2025/2026

10 Dicembre, 2025

1. Introduzione

Negli ultimi anni, l'Intelligenza Artificiale e il Machine Learning hanno aperto nuove prospettive nella medicina, consentendo diagnosi più precise e terapie personalizzate grazie all'analisi di grandi dataset clinici. Tuttavia, l'utilizzo di dati reali dei pazienti comporta rischi significativi legati alla privacy e alla sicurezza, oltre alla necessità di rispettare normative rigorose come il GDPR europeo e l'HIPAA statunitense. La sensibilità dei dati medici è confermata anche dal loro valore sul mercato illecito: una cartella clinica può arrivare a costare fino a 250 dollari, rendendoli un obiettivo attraente per i cybercriminali.

I dati sintetici si propongono come un'alternativa innovativa. Generati tramite tecniche come le Generative Adversarial Networks (GAN) o i modelli di diffusione, questi dati replicano le caratteristiche statistiche dei dataset reali senza contenere informazioni identificabili sui pazienti. Questo approccio offre la possibilità di sviluppare modelli predittivi efficaci e di condurre analisi approfondite riducendo i rischi etici e legali.

Tuttavia, rimane una questione centrale: è possibile produrre dati sintetici che siano al contempo sicuri e sufficientemente utili? Una protezione della privacy troppo restrittiva può compromettere l'utilità dei dati, mentre dati sintetici molto realistici possono essere vulnerabili a attacchi come il *membership inference*, che rivela se un individuo è presente nel dataset originale, o il *re-identification*, che collega dati sintetici a persone reali.

Diversi lavori precedenti hanno esplorato metodi di generazione e metriche di valutazione dei dati sintetici. Ad esempio, Figueira et al. descrivono vari approcci di sintesi, mentre Hernandez et al. confrontano strumenti di valutazione per determinarne l'efficacia. Il progetto si concentra sul bilanciamento tra privacy e utilità, analizzando quantitativamente come questi due aspetti influenzino le prestazioni dei dati sintetici in contesti sanitari.

1.1. Domande di ricerca

Questo studio si propone di rispondere alle seguenti domande:

- **RQ1:** Quali modelli mantengono le prestazioni più elevate quando addestrati su dati sintetici rispetto alla baseline con dati reali?
- **RQ2:** È possibile conciliare privacy e utilità in un dataset sintetico, o devono essere accettati compromessi significativi?

Per rispondere a queste domande, viene utilizzato l'UCI Diabetes Dataset, un dataset pubblico contenente 768 pazienti con variabili mediche predittive e diagnosi di diabete.

L'analisi si sviluppa attraverso quattro fasi. Innanzitutto, vengono generati dataset sintetici con diversi livelli di protezione della privacy: nessuna privacy, privacy moderata e privacy forte. Successivamente, viene valutata la somiglianza statistica confrontando distribuzioni e matrici di correlazione tra dati reali e sintetici mediante test statistici standard. Nella terza fase, viene verificata l'utilità dei dati sintetici addestrando modelli predittivi che imparano a diagnosticare il diabete valutandoli su dati reali attraverso le metriche Accuracy, Precision, Recall, F1-Score e ROC-AUC. Viene poi implementato un attacco di *membership inference* per testare la resistenza dei dati sintetici e individuare eventuali fughe di informazioni. Infine, il compromesso tra privacy e utilità viene visualizzato attraverso grafici che mostrano come le diverse configurazioni influenzino le prestazioni, permettendo di identificare il punto di equilibrio ottimale.

2. Generezione dati sintetici con privacy differenziale

La presente sezione illustra il processo di sintesi di tre dataset caratterizzati da differenti livelli di garanzie di privacy, a partire dal dataset reale di addestramento.

2.1. Caratteristiche del dataset

Il presente studio utilizza l'UCI Pima Indians Diabetes Dataset, un dataset pubblico composto da 768 osservazioni relative a pazienti di sesso femminile di età superiore ai 21 anni e di origine etnica.

Il dataset comprende 8 variabili predittive di natura clinica e una variabile target binaria:

- **Pregnancies:** numero di gravidanze (variabile discreta, range: 0-17)
- **Glucose:** concentrazione plasmatica di glucosio a 2 ore da un test orale di tolleranza al glucosio, espressa in mg/dL (variabile continua, range: 0-199)
- **BloodPressure:** pressione sanguigna diastolica, espressa in mm Hg (variabile continua, range: 0-122)
- **SkinThickness:** spessore della plica cutanea tricipitale, espresso in mm (variabile continua, range: 0-99)
- **Insulin:** livello di insulina sierica a 2 ore, espresso in $\mu\text{U/mL}$ (variabile continua, range: 0-846)
- **BMI:** indice di massa corporea, calcolato come $\text{peso in kg}/(\text{altezza in m})^2$ (variabile continua, range: 0-67.1)
- **DiabetesPedigreeFunction:** funzione che quantifica la storia familiare di diabete, calcolata mediante una funzione che considera la relazione e l'età di insorgenza del diabete nei familiari (variabile continua, range: 0.078-2.42)
- **Age:** età del paziente espressa in anni (variabile discreta, range: 21-81)
- **Outcome:** variabile target binaria che indica la presenza (1) o assenza (0) di diabete mellito

Il dataset presenta una distribuzione sbilanciata della variabile target, con approssimativamente il 65% di osservazioni negative (assenza di diabete) e il 35% di osservazioni positive (presenza di diabete).

2.2. Preparazione e preprocessing del dataset

La fase di preprocessing è fondamentale per garantire la qualità dei dati prima della generazione sintetica. Il dataset originale presenta una problematica rilevante: alcune variabili cliniche contengono valori pari a zero che risultano fisiologicamente implausibili. Ad esempio, valori nulli per la concentrazione di glucosio, la pressione sanguigna, lo spessore della plica cutanea, il livello di insulina o l'indice di massa corporea non sono clinicamente possibili in pazienti viventi.

Tali valori zero rappresentano in realtà dati mancanti che sono stati codificati impropriamente nel dataset originale. La strategia di trattamento adottata prevede:

1. **Identificazione dei valori anomali:** i valori pari a zero nelle colonne `glucose`, `blood_pressure`, `skin_thickness`, `insulin` e `bmi` vengono identificati come dati mancanti e sostituiti con `NaN`.
2. **Imputazione mediante mediana:** per ciascuna delle colonne sopra indicate, i valori mancanti vengono imputati utilizzando la mediana della distribuzione della rispettiva variabile calcolata sui valori non nulli. La scelta della mediana rispetto alla media è motivata dalla maggiore robustezza di questa statistica in presenza di outlier e distribuzioni asimmetriche, caratteristiche comuni nei dati clinici.
3. **Partizionamento del dataset:** il dataset preprocessato viene suddiviso in training set (70%, 537 osservazioni) e holdout set (30%, 231 osservazioni). La stratificazione garantisce che la proporzione tra classi positive e negative rimanga costante nei due sottoinsiemi, preservando la distribuzione originale della variabile da predire.

Il training set viene utilizzato esclusivamente per l'addestramento del modello generativo CTGAN, mentre l'holdout set viene riservato per la valutazione finale delle prestazioni dei modelli predittivi addestrati sui dati sintetici.

2.3. Architettura CTGAN

Per la sintesi dei dati è stato impiegato il framework SDV (Synthetic Data Vault) versione 1.0, nello specifico il modello `CTGANSynthesizer`. CTGAN (Conditional Tabular GAN) rappresenta una variante specializzata delle Generative Adversarial Networks, specificatamente progettata per la generazione di dati tabulari eterogenei, che supera le limitazioni delle GAN tradizionali nella modellazione di distribuzioni multimodali e nella gestione simultanea di variabili continue e categoriche.

2.3.1. Configurazione del training

Il processo di addestramento è stato configurato con i seguenti iperparametri:

- **Numero di epoche:** 3000 iterazioni complete sul training set
- **Batch size:** 500 campioni per batch
- **Learning rate:** 2×10^{-4} sia per il generatore che per il discriminatore
- **Ottimizzatore:** Adam con parametri di default ($\beta_1 = 0.9$, $\beta_2 = 0.999$)

Tali iperparametri sono stati selezionati sulla base delle raccomandazioni della letteratura e attraverso sperimentazione preliminare, al fine di garantire una convergenza stabile del modello preservando la qualità dei campioni sintetici generati.

2.3.2. Meccanismo di funzionamento

Il processo di apprendimento di CTGAN si articola attraverso il seguente algoritmo adversarial:

1. Il generatore G apprende una funzione di mappatura $G : \mathcal{Z} \rightarrow \mathcal{X}$ che trasforma vettori di rumore $z \sim \mathcal{N}(0, I)$ in campioni sintetici $\tilde{x} = G(z)$ che approssimano la distribuzione dei dati reali $x \sim p_{data}$.
2. Il discriminatore D apprende una funzione $D : \mathcal{X} \rightarrow [0, 1]$ che stima la probabilità che un campione provenga dalla distribuzione reale piuttosto che da quella generata.
3. I due modelli vengono addestrati alternando aggiornamenti del discriminatore e del generatore, ottimizzando rispettivamente le seguenti funzioni obiettivo:

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{data}} [\log D(x)] - \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))] \quad (1)$$

$$\mathcal{L}_G = -\mathbb{E}_{z \sim p_z} [\log D(G(z))] \quad (2)$$

Una caratteristica distintiva di CTGAN rispetto alle GAN tradizionali è l'utilizzo di tecniche specializzate per dati tabulari:

- **Mode-specific normalization:** le variabili continue vengono normalizzate utilizzando una Gaussian Mixture Model che identifica automaticamente le diverse modalità della distribuzione, permettendo di catturare distribuzioni multimodali complesse.
- **Conditional generation:** durante il training, il generatore viene condizionato su specifiche categorie della variabile target, garantendo che il dataset sintetico mantenga la distribuzione delle classi del dataset reale.
- **Training-by-sampling:** per gestire il bilanciamento delle classi, i campioni vengono selezionati durante il training in modo da garantire una rappresentazione uniforme di tutte le modalità delle variabili categoriche.

2.4. Livelli di privacy differenziale

Al fine di analizzare quantitativamente il trade-off tra protezione della privacy e preservazione dell'utilità, sono stati generati tre dataset sintetici con differenti garanzie formali di privacy:

1. **Nessuna privacy** (`synthetic_no_privacy.csv`): Il modello CTGAN genera dati sintetici senza alcuna protezione aggiuntiva della privacy. I campioni prodotti replicano fedelmente le caratteristiche statistiche del training set.
2. **Privacy moderata** (`synthetic_privacy_moderate.csv`, $\epsilon = 5.0$): Dopo la generazione tramite CTGAN, viene applicato rumore differenzialmente privato mediante il meccanismo di Laplace con dominio limitato, mediante l'utilizzo del parametro $\epsilon = 5.0$.
3. **Privacy forte** (`synthetic_privacy_strong.csv`, $\epsilon = 1.0$): Viene applicato un livello più rigoroso di protezione della privacy attraverso il medesimo meccanismo di Laplace ma con $\epsilon = 1.0$, garantendo maggior privacy.

Il parametro ε (epsilon) quantifica formalmente il livello di privacy garantito: valori più bassi corrispondono a maggiori garanzie di privacy ma introducono maggiore distorsione nei dati. Un valore di $\varepsilon = 1.0$ è generalmente considerato uno standard rigoroso per applicazioni ad alta sensibilità, mentre $\varepsilon = 5.0$ rappresenta un compromesso più permissivo ma ancora significativo.

2.5. Implementazione della privacy differenziale

L'applicazione della privacy differenziale avviene mediante la libreria `diffprivlib`, un'implementazione open-source di algoritmi differenzialmente privati conforme agli standard crittografici. Per ciascuna variabile numerica del dataset sintetico (escludendo la variabile target binaria), viene applicato il meccanismo `LaplaceBoundedDomain`, che garantisce la privacy differenziale mantenendo i valori entro un intervallo predefinito.

2.5.1. Meccanismo di Laplace con dominio limitato

Per ogni feature f e ogni valore x_i^f nel dataset sintetico, il valore perturbato è calcolato come:

$$\tilde{x}_i^f = \text{clip}\left(x_i^f + \text{Lap}\left(\frac{s_f}{\varepsilon}\right), l_f, u_f\right) \quad (3)$$

dove:

- $\text{Lap}(\lambda)$ denota una variabile aleatoria estratta dalla distribuzione di Laplace con parametro di scala $\lambda = \frac{s_f}{\varepsilon}$
- $s_f = u_f - l_f$ rappresenta la sensibilità globale della query, definita come la differenza tra il limite superiore e inferiore del dominio
- l_f e u_f sono rispettivamente il limite inferiore e superiore del dominio della feature
- ε è il parametro di privacy
- La funzione $\text{clip}(x, l, u) = \max(l, \min(x, u))$ assicura il rispetto rigoroso dei vincoli di dominio

La distribuzione di Laplace è definita dalla funzione di densità di probabilità:

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) \quad (4)$$

dove μ è il parametro di posizione (tipicamente 0) e $b = \frac{s_f}{\varepsilon}$ è il parametro di scala.

2.5.2. Determinazione dei limiti di dominio

I limiti inferiore l_f e superiore u_f per ciascuna variabile sono determinati sulla base della distribuzione empirica nel dataset reale, secondo le seguenti regole:

- **Variabili discrete non negative** (pregnancies, age):

$$l_f = \max(0, \min(x_i^f)), \quad u_f = \max(x_i^f) + 2 \quad (5)$$

Il margine di +2 unità sul limite superiore permette una limitata extrapolazione per età e numero di gravidanze.

- **Variabili continue** (glucose, blood_pressure, skin_thickness, insulin, bmi, diabetes_pedigree):

$$l_f = Q_{0.01}^f - 0.05 \cdot (Q_{0.99}^f - Q_{0.01}^f), \quad u_f = Q_{0.99}^f + 0.05 \cdot (Q_{0.99}^f - Q_{0.01}^f) \quad (6)$$

dove Q_p^f denota il p-esimo percentile della distribuzione della feature f nel dataset reale. L'utilizzo dei percentili 1° e 99° anziché dei valori minimo e massimo assoluti rende il metodo robusto rispetto a outlier estremi. Il margine aggiuntivo del 5% dell'intervallo interquantile consente una moderata variabilità oltre i valori osservati.

2.5.3. Garanzie teoriche di privacy

Il meccanismo implementato garantisce ϵ -privacy differenziale secondo la seguente definizione formale:

Definizione (ϵ -Privacy Differenziale). Un meccanismo randomizzato \mathcal{M} soddisfa ϵ -privacy differenziale se per ogni coppia di dataset adiacenti D e D' (che differiscono per un singolo record) e per ogni sottoinsieme misurabile S dello spazio degli output:

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D') \in S] \quad (7)$$

Questa proprietà garantisce che la presenza o assenza di un singolo individuo nel dataset abbia un impatto limitato (quantificato da ϵ) sulla distribuzione degli output, rendendo computazionalmente difficile determinare se un particolare individuo ha contribuito al dataset.

2.6. Post-processing e normalizzazione

Successivamente all'applicazione del rumore differenzialmente privato, viene eseguita una fase di post-processing denominata *clipping conservativo*, che garantisce la coerenza dei dati sintetici con i vincoli del dominio applicativo senza violare le garanzie di privacy differenziale (il post-processing preserva la privacy differenziale secondo il teorema di post-processing).

Le operazioni di post-processing comprendono:

1. **Arrotondamento delle variabili discrete:** le variabili *pregnancies*, *age*, *glucose*, *blood_pressure*, *skin_thickness* e *insulin* vengono arrotondate al più vicino intero mediante la funzione $\lfloor x + 0.5 \rfloor$.
2. **Clipping ai percentili:** per ciascuna variabile continua, i valori vengono limitati all'intervallo $[Q_{0.01}, Q_{0.99}]$ calcolato sul dataset reale:

$$x_i^f \leftarrow \text{clip}(x_i^f, Q_{0.01}^f, Q_{0.99}^f) \quad (8)$$

3. **Normalizzazione della variabile target:** la variabile *outcome* viene convertita in valori binari $\{0, 1\}$ mediante arrotondamento e successivo clipping:

$$\text{outcome}_i \leftarrow \text{clip}(\lfloor \text{outcome}_i + 0.5 \rfloor, 0, 1) \quad (9)$$

4. **Conversione dei tipi:** le variabili discrete vengono convertite al tipo intero (`int64`), mentre le variabili continue mantengono la precisione a virgola mobile (`float64`).

Questa procedura assicura che i dati sintetici generati rispettino rigorosamente i vincoli di tipo e di dominio delle variabili originali, preservando al contempo le garanzie formali di privacy differenziale introdotte nella fase precedente.

2.7. Validazione statistica preliminare

Al termine del processo di generazione, viene condotta un'analisi comparativa sistematica tra i dataset reali e sintetici per verificare la preservazione delle proprietà statistiche fondamentali. Tale validazione fornisce una prima indicazione qualitativa del trade-off tra privacy e utilità prima di procedere con le valutazioni quantitative approfondite.

Le metriche di confronto includono:

- **Statistiche univariate:** per ciascuna variabile vengono confrontati minimo, massimo, media, mediana, deviazione standard e quartili tra dataset reale e sintetici.
- **Distribuzione empirica:** vengono generati istogrammi sovrapposti per visualizzare le distribuzioni empiriche e identificare eventuali discrepanze significative.
- **Matrice di correlazione:** viene calcolata la matrice di correlazione di Pearson per entrambi i dataset e ne viene valutata la somiglianza attraverso la distanza di Frobenius:

$$d_{Frob}(C_{real}, C_{synth}) = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (C_{real}^{ij} - C_{synth}^{ij})^2} \quad (10)$$

- **Distribuzione della variabile target:** viene verificata la preservazione della proporzione tra classi positive e negative.

I risultati di questa analisi preliminare sono riportati in forma tabellare per ciascuno dei tre dataset sintetici generati, evidenziando come l'incremento delle garanzie di privacy (ϵ decrescente) influenzi progressivamente la fedeltà statistica rispetto al dataset originale.

2.8. Archiviazione dei dataset

I tre dataset sintetici generati vengono persistiti in formato CSV per le successive fasi di analisi:

- `synthetic_no_privacy.csv`: 537 campioni sintetici senza protezione della privacy
- `synthetic_privacy_moderate.csv`: 537 campioni sintetici con $\epsilon = 5.0$
- `synthetic_privacy_strong.csv`: 537 campioni sintetici con $\epsilon = 1.0$

Parallelamente vengono archiviati anche i dataset reali utilizzati:

- `diabetes_train.csv`: training set reale (537 osservazioni)
- `diabetes_holdout.csv`: holdout set reale (231 osservazioni)

2.9. Validazione statistica preliminare

Al termine del processo di generazione, viene condotta un'analisi comparativa sistematica tra i dataset reali e sintetici per verificare la preservazione delle proprietà statistiche fondamentali. Tale validazione fornisce una prima indicazione qualitativa del trade-off tra privacy e utilità prima di procedere con le valutazioni quantitative approfondite.

Le metriche di confronto includono:

- **Statistiche univariate:** per ciascuna variabile vengono confrontati minimo, massimo, media, mediana, deviazione standard e quartili tra dataset reale e sintetici.
- **Distribuzione empirica:** vengono generati istogrammi sovrapposti per visualizzare le distribuzioni empiriche e identificare eventuali discrepanze significative.
- **Matrice di correlazione:** viene calcolata la matrice di correlazione di Pearson per entrambi i dataset e ne viene valutata la somiglianza attraverso la distanza di Frobenius:

$$d_{Frob}(C_{real}, C_{synth}) = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (C_{real}^{ij} - C_{synth}^{ij})^2} \quad (11)$$

- **Distribuzione della variabile target:** viene verificata la preservazione della proporzione tra classi positive e negative.