Generazione di dati sanitari sintetici: compromesso tra privacy e utilità nella ricerca medica

Progetto di Cybersecurity

Autori:
Sara Casadio - 0001186923
Noemi Ferrara - 0001170201
Giorgia Pirelli - 0001176116

Anno accademico 2025/2026

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 oltre 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.

1.1 Domande di ricerca

Questo studio si propone di rispondere alle seguenti domande di ricerca:

- **RQ1:** Quali modelli di machine learning mantengono le migliori prestazioni quando addestrati su dati sintetici, rispetto alla baseline ottenuta con dati reali?
- **RQ2:** È possibile conciliare tutela della privacy e utilità dei dati sintetici al punto da renderli un'alternativa affidabile ai dati clinici reali nella ricerca medica e nelle applicazioni di machine learning?

1.2 Riproducibilità e struttura del progetto

Il progetto è stato sviluppato per garantire la massima trasparenza e per essere completamente replicabile tramite l'esecuzione del notebook `progettoCybersecurity.ipynb`, disponibile al seguente repository GitHub. Tale notebook offre all'utente due distinti percorsi operativi in base alle proprie esigenze computazionali:

1. Riproduzione completa dell'esperimento (da zero): il notebook scarica il dataset originale da Kaggle, esegue la pre-elaborazione, addestra i modelli CTGAN e DP-CTGAN e genera nuovi dataset sintetici utilizzabili per tutte le fasi successive.
2. Analisi e consultazione rapida dei risultati dell'esperimento (caricando dati preesistenti): per facilitare la revisione del progetto senza la necessità di elevate risorse computazionali o lunghi tempi di attesa, è possibile procedere in due modi:
    - Consultazione statica: il notebook viene consegnato con gli output delle celle più importanti già renderizzati. È quindi possibile analizzare grafici, metriche di qualità e risultati degli attacchi MIA direttamente dalla versione pre-eseguita.

- Esecuzione parziale (caricamento dati): è possibile saltare le fasi di training caricando i dataset sintetici già generati. In questo modo l'utente può rieseguire istantaneamente solo le celle di analisi, verificando in tempo reale la validità dei risultati.

L'analisi è strutturata in cinque fasi principali ed utilizza il dataset pubblico *Diabetes Health Indicators*, scaricato al seguente link Kaggle, composto da 253.680 osservazioni e da un insieme di variabili cliniche e demografiche impiegate per la predizione della diagnosi di diabete. A partire da tale dataset reale, sono stati generati dataset sintetici caratterizzati da differenti livelli di protezione della privacy: nessuna privacy, privacy moderata e privacy forte. Nella fase successiva dell'analisi è stata valutata la fedeltà statistica dei dati sintetici, confrontando caratteristiche statistiche e strutture di dipendenza tramite matrici di correlazione, utilizzando test statistici standard per quantificare la similarità rispetto ai dati reali. Nella terza fase è stata analizzata l'utilità dei dati sintetici in ambito predittivo: diversi modelli di machine learning sono stati addestrati esclusivamente su dati sintetici e valutati su dati reali, al fine di misurarne la capacità di generalizzazione. Le prestazioni sono state confrontate mediante varie metriche. Successivamente, è stato implementato un attacco di *Membership Inference Attack* per valutare la robustezza dei dataset sintetici rispetto a potenziali fughe di informazioni. Infine, il compromesso tra privacy e utilità è stato analizzato in modo comparativo attraverso visualizzazioni grafiche che mostrano l'impatto delle diverse configurazioni di privacy sulle prestazioni dei modelli, consentendo l'individuazione di un possibile punto di equilibrio ottimale tra protezione dei dati e valore informativo.

2 Analisi del dataset

Il presente studio utilizza il dataset *Diabetes Health Indicators* derivante dalla *Behavioral Risk Factor Surveillance System (BRFSS)* del 2015. Il dataset è composto da 253.680 osservazioni e 22 variabili, rendendolo particolarmente adatto per l'addestramento di modelli generativi basati su Differential Privacy grazie all'elevata numerosità campionaria.

Le variabili principali includono indicatori dello stile di vita e parametri clinici:

- **Outcome** (Diabetes_binary): variabile target. Indica se il paziente ha o non ha il diabete (0 = no diabete, 1 = diabete).
- **HighBP**: indica se ha una pressione sanguigna elevata (0 = no, 1 = sì).
- **HighChol**: indica se ha il colesterolo totale elevato (0 = no, 1 = sì).
- **CholCheck**: indica se il paziente ha effettuato un controllo del colesterolo negli ultimi 5 anni (0 = no, 1 = sì).
- **BMI** (Body Mass Index): indica l'indice di massa corporea del paziente.
- **Smoker**: indica se il paziente ha fumato almeno 100 sigarette (circa 5 pacchetti) nell'arco della sua vita (0 = no, 1 = sì).
- **Stroke**: indica se al paziente è mai stato diagnosticato un ictus (0 = no, 1 = sì).
- **HeartDiseaseorAttack**: presenza di malattie coronariche (CHD) o storia di infarto miocardico (MI) (0 = no, 1 = sì).

- **PhysActivity**: svolgimento di attività fisica o esercizio nelle ultime 30 ore, escludendo l'attività lavorativa (0 = no, 1 = sì).
- **Fruits**: consumo di frutta una o più volte al giorno (0 = no, 1 = sì).
- **Veggies**: consumo di verdura una o più volte al giorno (0 = no, 1 = sì).
- **HvyAlcoholConsump**: consumo eccessivo di alcol. Definito come più di 14 drink a settimana per gli uomini e più di 7 per le donne (0 = no, 1 = sì).
- **AnyHealthcare**: indica che il paziente possiede una copertura sanitaria di qualsiasi tipo (0 = no, 1 = sì).
- **NoDocbcCost**: indica se il paziente ha dovuto rinunciare a una visita medica nell'ultimo anno a causa dei costi eccessivi (0 = no, 1 = sì).
- **GenHlth**: autovalutazione dello stato di salute generale su una scala da 1 a 5 (1 = eccellente, 2 = molto buono, 3 = buono, 4 = discreto, 5 = scarso).
- **MentHlth**: numero di giorni, negli ultimi 30, in cui la salute mentale non è stata buona (include stress, depressione e problemi emotivi). Range: 0-30.
- **PhysHlth**: numero di giorni, negli ultimi 30, in cui la salute fisica o infortuni non sono stati buoni. Range: 0-30.
- **DiffWalk**: presenza di serie difficoltà nel camminare o nel salire le scale (0 = no, 1 = sì).
- **Sex**: sesso del paziente (0 = femmina, 1 = maschio).
- **Age**: categoria di età del paziente suddivisa in 13 classi (1 = 18-24 anni, fino a 13 = 80 anni o più). Ogni incremento di classe corrisponde a un intervallo di 5 anni.
- **Education**: livello di istruzione raggiunto su una scala da 1 a 6 (1 = mai frequentato la scuola, 6 = laurea o istruzione superiore).
- **Income**: fascia di reddito familiare annuo su una scala da 1 a 8 (1 = meno di \$10.000, 8 = \$75.000 o più).

2.1 Partizionamento del dataset

Data la natura del dataset, già consolidato e privo di valori mancanti critici, la fase di preparazione si è concentrata solo sulla suddivisione del dataset in training set (80%, 202.944 osservazioni) e holdout set (20%, 50.736 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 è stato utilizzato esclusivamente per l'addestramento dei modelli CTGAN, mentre l'holdout set è stato riservato per la valutazione finale delle prestazioni dei modelli predittivi addestrati sui dati sintetici.

3 Generezione dati sintetici con e senza 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 training.

3.1 Modelli Generativi: CTGAN e DP-CTGAN

La generazione dei dati sintetici è stata affidata a due architetture basate sul framework delle *Generative Adversarial Networks*, integrate attraverso le librerie `Synthetic Data Vault (SDV)` e `SmartNoise-Synth`.

3.1.1 CTGAN (Conditional Tabular GAN)

Per il dataset privo di garanzie di privacy, è stato impiegato il framework `SDV` versione 1.29.1, 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 sui dati tabulari grazie a tre innovazioni fondamentali:

- **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.

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 $G : \mathcal{Z} \to \mathcal{X}$ 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 $D : \mathcal{X} \to [0, 1]$ 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:

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

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

3.1.2 DP-CTGAN (SmartNoise-Synth)

Per la generazione dei dataset con protezione della privacy ($\varepsilon = 4.0$ e $\varepsilon = 2.0$), è stata utilizzata l'implementazione `DP-CTGAN` fornita dalla libreria `SmartNoise-Synth`. Questa versione estende l'architettura CTGAN integrando i principi della **Differential Privacy** durante la fase di ottimizzazione. Il processo di addestramento viene modificato attraverso:

- **Differentially Private SGD (DP-SGD)**: l'aggiornamento dei pesi del generatore e del discriminatore avviene tramite gradienti "perturbati". Ogni gradiente viene ritagliato (*clipping*) per limitare l'influenza del singolo record e successivamente addizionato con rumore gaussiano calibrato in base al budget ε scelto.
- **Privacy Accounting**: la libreria gestisce il calcolo cumulativo del consumo di privacy (*privacy budget tracking*) attraverso meccanismi avanzati come il *Rényi Differential Privacy*, garantendo che il livello di protezione dichiarato sia matematicamente verificabile.

3.2 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` viene addestrato senza vincoli di Differential Privacy. In questo scenario, l'algoritmo non subisce alcuna perturbazione dei gradienti, producendo campioni che replicano fedelmente le distribuzioni congiunte del training set originale. Rappresenta il limite superiore dell'utilità, ma anche il massimo rischio di successo per attacchi di *Membership Inference*.
2. **Privacy Moderata** (`synthetic_privacy_moderata.csv`, $\varepsilon = 4.0$): viene configurato il modello `DP-CTGAN` con un budget di privacy intermedio. Questo valore di ε introduce una quantità di rumore sufficiente ad aumentare la privacy, pur preservando le correlazioni necessarie per l'addestramento di modelli predittivi.
3. **Privacy Forte** (`synthetic_privacy_strong.csv`, $\varepsilon = 2.0$): viene applicato un rigore estremo nella protezione del dato. Un ε così ridotto, mira a ridurre drasticamente l'efficacia di un attacco *Membership Inference*, accettando tuttavia una potenziale degradazione della fedeltà statistica e delle performance dei classificatori.

La scelta del parametro ε, noto come *privacy budget*, rappresenta un compromesso diretto tra tutela della privacy e qualità dei dati generati. Valori più elevati di ε corrispondono a una protezione della privacy più debole, in quanto richiedono l'aggiunta di una quantità minore di rumore durante l'addestramento, consentendo al modello di apprendere in modo più accurato le distribuzioni sottostanti dei dati reali. Al contrario, valori più bassi di ε impongono una perturbazione più intensa dei gradienti, riducendo il rischio di memorizzazione di informazioni sensibili ma degradando potenzialmente la fedeltà statistica e l'utilità dei dati sintetici.

Nel contesto di questo studio, sono stati selezionati due livelli di privacy rappresentativi: $\varepsilon = 4.0$, considerato indicativo di una *privacy moderata*, e $\varepsilon = 2.0$, associato a una

privacy forte. Questa scelta consente di analizzare empiricamente l'impatto del budget di privacy sia sulla somiglianza statistica tra dati reali e sintetici, sia sulle prestazioni dei modelli di machine learning addestrati sui dati generati, permettendo di valutare in modo sistematico il compromesso tra privacy e utilità.

3.3 Archiviazione dei dataset

I tre dataset sintetici generati (ciascuno con lo stesso numero di osservazioni del dataset di training reale) sono stati salvati in formato CSV e resi disponibili per le successive fasi di analisi. Tale scelta consente non solo di garantire la riproducibilità degli esperimenti, ma anche di importare direttamente i dataset senza la necessità di riaddestrare i modelli generativi, permettendo così di valutare e confrontare le prestazioni dei modelli.

- `synthetic_no_privacy.csv`: 202.944 campioni sintetici senza protezione della privacy
- `synthetic_privacy_moderata.csv`: 202.944 campioni sintetici con $\varepsilon = 4.0$
- `synthetic_privacy_strong.csv`: 202.944 campioni sintetici con $\varepsilon = 2.0$

Inoltre, sono stati salvati anche i dataset reali utilizzati:

- `diabetes_train.csv`: training set reale (202.944 osservazioni)
- `diabetes_holdout.csv`: holdout set reale (50.736 osservazioni)

4 Risultati dell'analisi di qualità (SDV Metrics)

La qualità dei dataset sintetici generati è stata valutata utilizzando le metriche standard fornite dalla libreria *Synthetic Data Vault (SDV)*, ampiamente utilizzata per l'analisi della validità e della fedeltà statistica dei dati sintetici. In particolare, SDV fornisce due strumenti complementari: il *Diagnostic Report*, volto a verificare la correttezza strutturale dei dati generati, e il *Quality Score*, progettato per misurare la somiglianza statistica tra dati reali e sintetici.

4.1 Diagnostic Report

Il *Diagnostic Report* valuta se il dataset sintetico è strutturalmente valido, ossia se rispetta i vincoli formali e semantici definiti dallo schema dei dati originali. Tale analisi non misura la somiglianza statistica, ma verifica che i dati generati siano coerenti e utilizzabili da un punto di vista strutturale.

Tutti i dataset sintetici considerati hanno ottenuto un punteggio pari al 100% sia in termini di *Data Validity* sia di *Data Structure*, come riportato in Tabella 1.

Il punteggio massimo in *Data Validity* indica che tutte le variabili sintetiche rispettano i vincoli di dominio e di tipo delle corrispondenti variabili reali: non sono presenti valori fuori range, categorie invalide o violazioni dei tipi di dato. Analogamente, il punteggio massimo in *Data Structure* conferma che la struttura del dataset, inclusa la cardinalità delle relazioni tra variabili, è stata preservata correttamente durante il processo di generazione.

Tabella 1: Risultati Diagnostic Report SDV

Dataset	Data Validity	Data Structure	Overall Score
No Privacy	100.0%	100.0%	100.0%
Privacy Moderata ($\varepsilon = 4.0$)	100.0%	100.0%	100.0%
Privacy Strong ($\varepsilon = 2.0$)	100.0%	100.0%	100.0%

È importante sottolineare che un punteggio perfetto nel *Diagnostic Report* non implica che i dati sintetici siano statisticamente simili ai dati reali, ma unicamente che essi risultano formalmente corretti e coerenti con lo schema originale.

4.2 Quality Score

Il *Quality Score* di SDV fornisce invece una misura quantitativa della fedeltà statistica dei dati sintetici rispetto al dataset reale. Tale punteggio è ottenuto combinando metriche di similarità delle distribuzioni univariate e metriche basate sulle correlazioni tra coppie di variabili, restituendo un valore normalizzato compreso tra 0 e 1, dove valori più elevati indicano una maggiore somiglianza statistica. I risultati riportati in Tabella 2 mostrano differenze contenute tra i diversi livelli di protezione della privacy. In particolare, il dataset generato con privacy moderata ($\varepsilon = 4.0$) ottiene il punteggio più elevato, seguito dal dataset con privacy forte ($\varepsilon = 2.0$), mentre il dataset generato senza meccanismi di privacy presenta un *Quality Score* leggermente inferiore. Tuttavia, le differenze osservate tra i punteggi sono relativamente ridotte, suggerendo che il *Quality Score* da solo non è sufficiente per valutare l'effettiva utilità dei dati sintetici in contesti applicativi. Per questo motivo, l'analisi della qualità statistica è stata integrata da ulteriori analisi mirate, volte a confrontare più in dettaglio le proprietà dei dataset sintetici rispetto a quelle del dataset reale.

Tabella 2: Quality Score SDV per livello di privacy

Dataset	Quality Score
Privacy Moderata ($\varepsilon = 4.0$)	**0.8939**
Privacy Strong ($\varepsilon = 2.0$)	0.8908
No Privacy (CTGAN)	0.8757

5 Valutazione della somiglianza statistica

La valutazione della somiglianza statistica si concentra su diversi livelli, ciascuno dei quali fornisce informazioni complementari sulla coerenza e la fedeltà dei dati generati:

- **Statistiche univariate**: per ciascuna variabile vengono confrontate importanti statistiche descrittive, come media e deviazione standard, al fine di verificare la coerenza delle distribuzioni tra dataset reale e sintetici.
- **Distribuzioni empiriche**: vengono generati istogrammi sovrapposti per visualizzare le distribuzioni delle singole variabili e identificare eventuali discrepanze significative.

- **Matrice di correlazione**: per ciascun dataset viene calcolata la matrice di correlazione di Pearson tra tutte le variabili numeriche. La somiglianza tra matrici reali e sintetiche è quantificata mediante la distanza di Frobenius:

$$d_{Frob}(C_{real}, C_{synth}) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(C_{real}^{ij} - C_{synth}^{ij})^2}, \tag{3}$$

  Valori di distanza vicini a zero indicano che il dataset sintetico riproduce fedelmente le relazioni lineari tra le variabili. Al contrario, valori elevati segnalano che le correlazioni tra alcune variabili differiscono significativamente tra reale e sintetico, suggerendo una perdita di fedeltà statistica nella generazione dei dati.

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

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

5.1 Dataset senza protezione della privacy

Il dataset sintetico senza alcuna protezione della privacy (`synthetic_no_privacy.csv`) mostra una buona preservazione delle proprietà statistiche del dataset originale. L'analisi delle statistiche descrittive (Tabella 3) evidenzia differenze contenute per tutte le variabili.

L'analisi delle correlazioni tra le variabili nei dati sintetici mostra che CTGAN è in grado di preservare le principali relazioni presenti nei dati reali. Le matrici di correlazione evidenziano come la struttura di dipendenza tra le feature sia riprodotta in maniera fedele, confermando la capacità del modello di mantenere le caratteristiche multivariate dei dati originali. Questo risultato è supportato dalla distanza di Frobenius tra le matrici di correlazione, che risulta pari a **1.22**, indicando una differenza complessiva molto contenuta tra le correlazioni reali e sintetiche.

Tabella 3: Confronto tra statistiche descrittive dei dati reali e sintetici

Feature	**Mean**$_{Real}$	**Mean**$_{Synth}$	**ΔMean**	**Std**$_{Real}$	**Std**$_{Synth}$	**ΔStd**
outcome	0.1393	0.2320	0.0926	0.3463	0.4221	0.0758
HighBP	0.4290	0.5762	0.1472	0.4949	0.4942	-0.0008
HighChol	0.4243	0.5773	0.1530	0.4942	0.4940	-0.0002
CholCheck	0.9624	0.9600	-0.0024	0.1902	0.1960	0.0058
BMI	28.3780	29.6326	1.2546	6.5983	6.0720	-0.5263
Smoker	0.4429	0.5933	0.1505	0.4967	0.4912	-0.0055
Stroke	0.0404	0.0511	0.0107	0.1969	0.2202	0.0233
HeartDiseaseorAttack	0.0942	0.1401	0.0459	0.2922	0.3471	0.0550
PhysActivity	0.7570	0.7954	0.0384	0.4289	0.4034	-0.0255
Fruits	0.6344	0.7035	0.0691	0.4816	0.4567	-0.0249
Veggies	0.8114	0.8526	0.0412	0.3912	0.3545	-0.0367
HvyAlcoholConsump	0.0557	0.1115	0.0558	0.2294	0.3148	0.0854
AnyHealthcare	0.9511	0.9354	-0.0158	0.2156	0.2459	0.0303
NoDocbcCost	0.0846	0.0539	-0.0307	0.2783	0.2258	-0.0525
GenHlth	2.5118	2.5234	0.0116	1.0684	1.2587	0.1903
MentHlth	3.1898	4.5953	1.4055	7.4176	8.5479	1.1302
PhysHlth	4.2508	7.6605	3.4097	8.7256	11.6871	2.9615
DiffWalk	0.1682	0.1458	-0.0223	0.3740	0.3529	-0.0211
Sex	0.4410	0.3421	-0.0989	0.4965	0.4744	-0.0221
Age	8.0328	8.0341	0.0012	3.0514	2.7493	-0.3021
Education	5.0499	5.1339	0.0839	0.9864	0.9905	0.0041
Income	6.0508	5.3074	-0.7434	2.0727	2.1627	0.0900

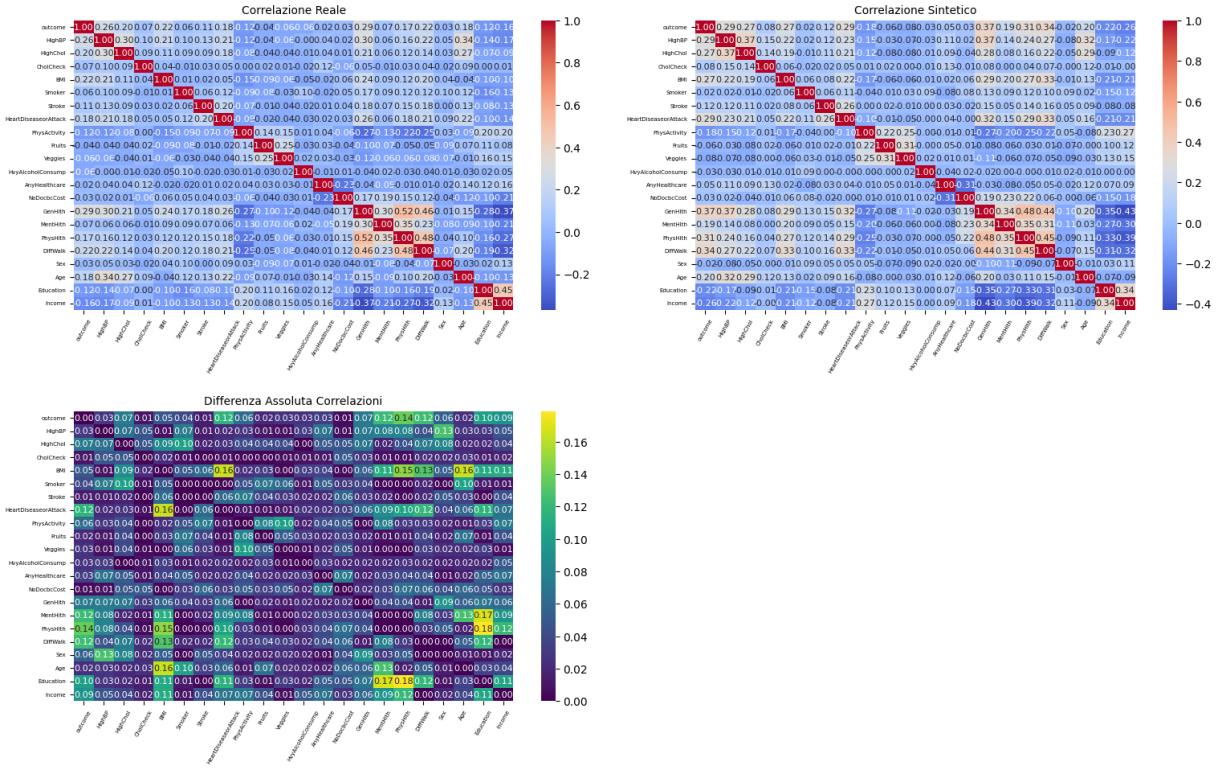

Figura 1: Matrici di correlazione dati reali (prima), sintetici NO PRIVACY (seconda) e la loro differenza (terza).

I grafici seguenti mostrano la sovrapposizione tra le distribuzioni delle feature dei dati reali (in blu) e di quelli sintetici (in giallo), evidenziando come CTGAN sia in grado di riprodurre in modo accurato la forma e le caratteristiche statistiche delle distribuzioni originali.

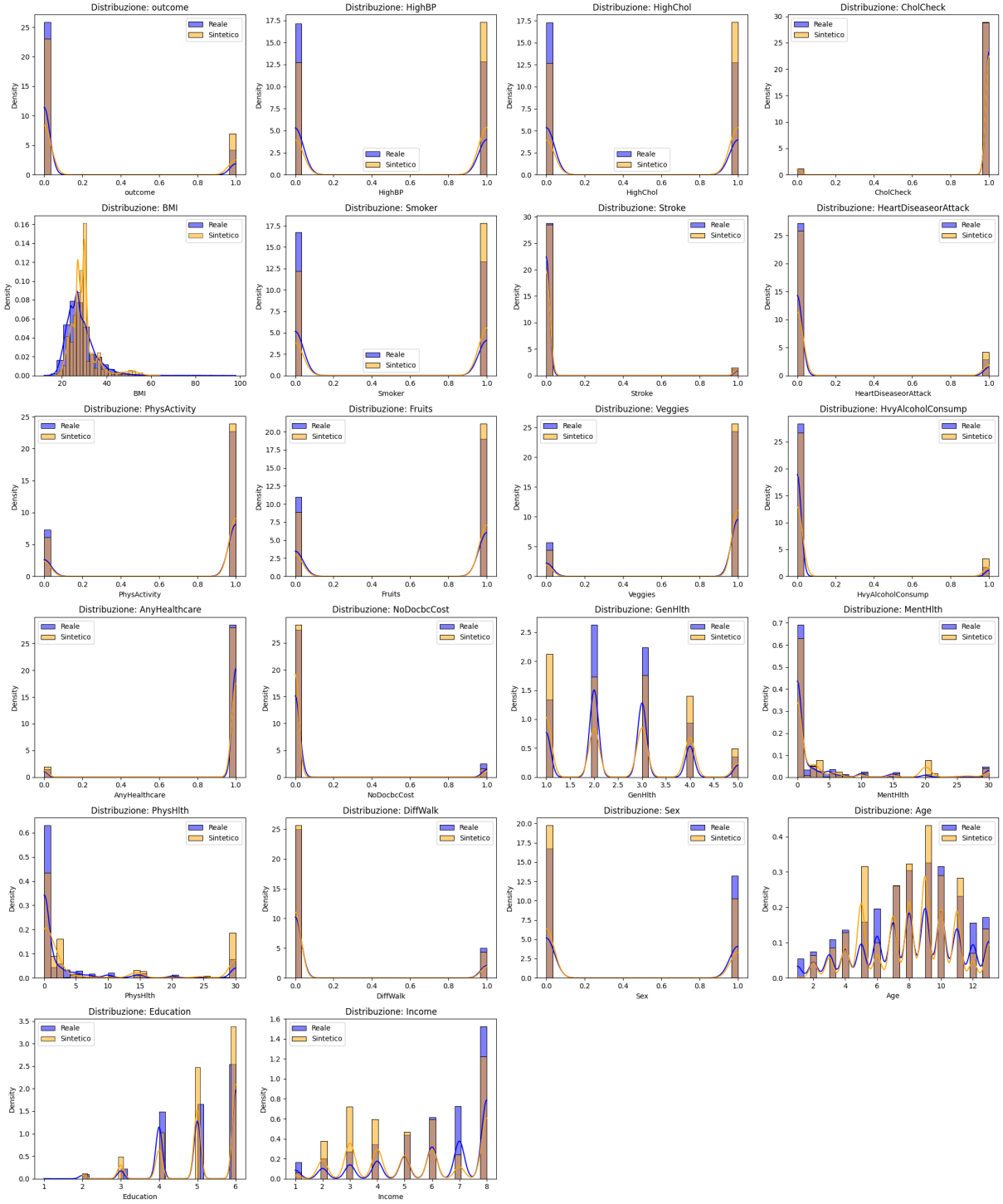

Figura 2: Sovrapposizione tra le distribuzioni delle feature dei dati reali (blu) e sintetici NO PRIVACY (giallo).

5.2 Dataset con privacy moderata ($\varepsilon = 4.0$)

L'introduzione di meccanismi di protezione della privacy nella generazione del dataset (`synthetic_privacy_moderata.csv`) comporta una leggera degradazione della precisione statistica a favore della sicurezza dei dati. Come mostrato nella Tabella 4, il modello riesce a mantenere una buona approssimazione per la maggior parte delle variabili, sebbene si osservino scostamenti più marcati rispetto al caso senza protezione.

Notabili sono le variazioni nelle variabili continue e di salute mentale/fisica (`BMI`, `MentHlth`, `PhysHlth`), dove la deviazione standard del dataset sintetico risulta sensibilmente inferiore rispetto all'originale.

Tabella 4: Confronto tra le statistiche descrittive dei dati reali e sintetici

Feature	Mean_Real	Mean_Sint	ΔMean	Std_Real	Std_Sint	ΔStd
outcome	0.1393	0.1040	-0.0353	0.3463	0.3053	-0.0410
HighBP	0.4290	0.3238	-0.1052	0.4949	0.4679	-0.0270
HighChol	0.4243	0.4251	0.0007	0.4942	0.4944	0.0001
CholCheck	0.9624	0.9814	0.0190	0.1902	0.1351	-0.0551
BMI	28.3780	26.2855	-2.0925	6.5983	3.4085	-3.1897
Smoker	0.4429	0.3699	-0.0730	0.4967	0.4828	-0.0140
Stroke	0.0404	0.0044	-0.0360	0.1969	0.0660	-0.1309
HeartDiseaseorAttack	0.0942	0.1071	0.0129	0.2922	0.3093	0.0171
PhysActivity	0.7570	0.7969	0.0399	0.4289	0.4023	-0.0266
Fruits	0.6344	0.6744	0.0400	0.4816	0.4686	-0.0130
Veggies	0.8114	0.8374	0.0259	0.3912	0.3690	-0.0221
HvyAlcoholConsump	0.0557	0.0546	-0.0011	0.2294	0.2272	-0.0022
AnyHealthcare	0.9511	0.9753	0.0241	0.2156	0.1553	-0.0603
NoDocbcCost	0.0846	0.1527	0.0681	0.2783	0.3597	0.0814
GenHlth	2.5118	2.2305	-0.2814	1.0684	1.0426	-0.0258
MentHlth	3.1898	1.4448	-1.7450	7.4176	5.1926	-2.2250
PhysHlth	4.2508	2.1259	-2.1250	8.7256	6.8011	-1.9245
DiffWalk	0.1682	0.1657	-0.0024	0.3740	0.3718	-0.0022
Sex	0.4410	0.4504	0.0094	0.4965	0.4975	0.0010
Age	8.0328	8.5076	0.4748	3.0514	2.8713	-0.1801
Education	5.0499	5.1312	0.0812	0.9864	0.8430	-0.1434
Income	6.0508	6.5158	0.4649	2.0727	1.8585	-0.2142

L'analisi delle correlazioni evidenzia come la struttura multivariata rimanga identificabile, sebbene con un'intensità dei legami leggermente attenuata. Questo suggerisce che, nonostante la protezione della privacy, il modello DP-CTGAN riesca a preservare le relazioni logiche fondamentali, pur introducendo un "rumore" protettivo che ne riduce la precisione puntuale. Infatti, in questo caso si registra un incremento del valore della distanza di Frobenius, pari a **1.64**.

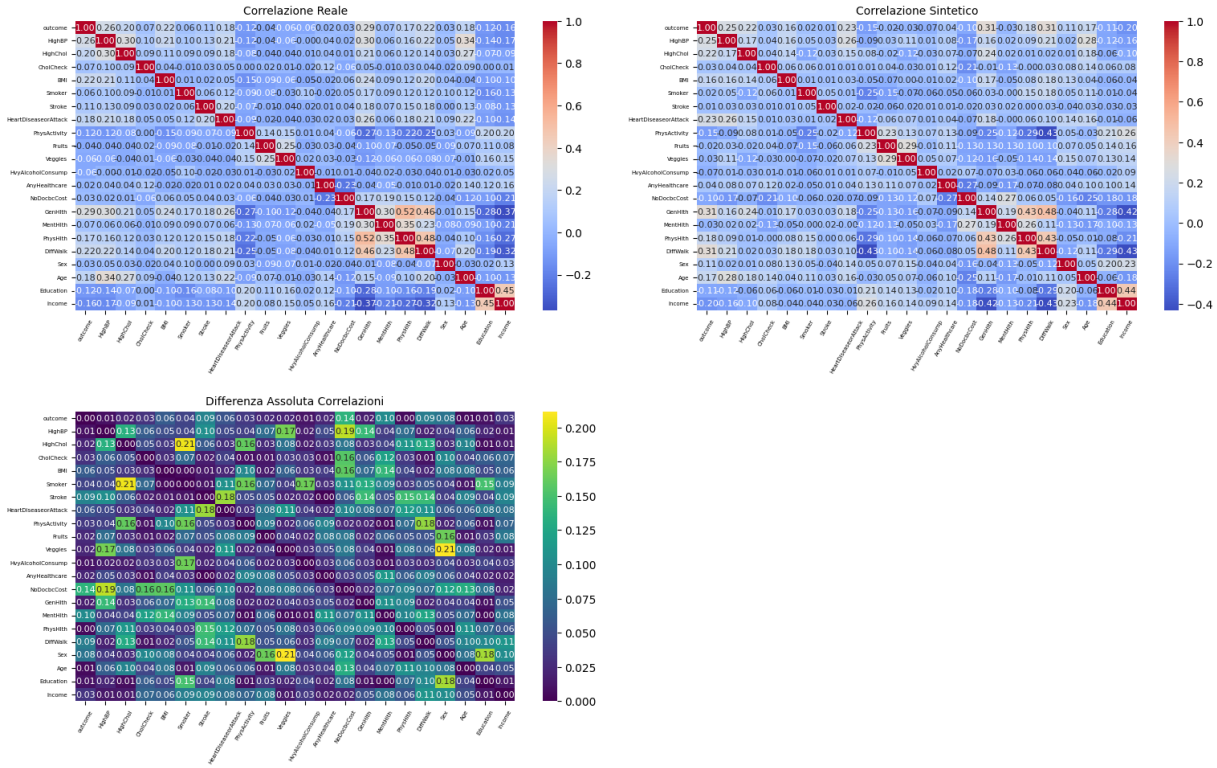

Figura 3: Matrici di correlazione dati reali (prima), sintetici PRIVACY MODERATA (seconda) e la loro differenza (terza).

I grafici di sovrapposizione delle distribuzioni (Figura 4) confermano visivamente quanto osservato: le distribuzioni univariata mostrano ancora un'ottima copertura del supporto dei dati reali, ma con curve di densità più smussate in corrispondenza dei valori meno frequenti.

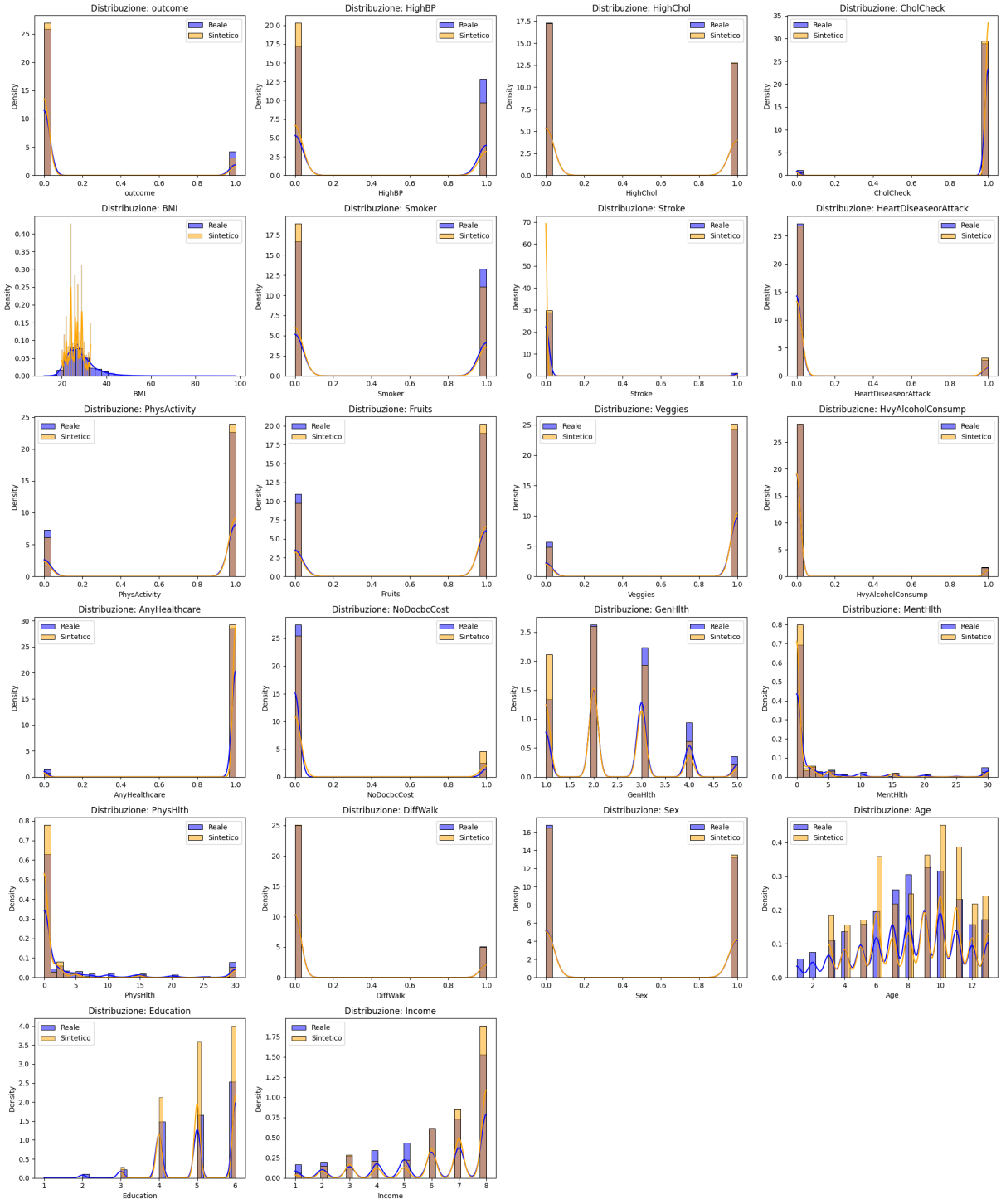

Figura 4: Sovrapposizione tra le distribuzioni delle feature dei dati reali (blu) e sintetici PRIVACY MODERATA (giallo).

5.3 Dataset con privacy strong ($\varepsilon = 2.0$)

L'adozione di un regime di privacy stringente ($\varepsilon = 2.0$) introduce un livello significativo di rumore durante l'addestramento del modello DP-CTGAN. Come evidenziato nella Tabella 5, le statistiche descrittive dei dati sintetici mostrano scostamenti medi relativamente contenuti rispetto ai valori reali per molte variabili. In particolare, le medie e le deviazioni standard appaiono più vicine al centro della distribuzione, suggerendo che i valori sintetici

tendono a convergere verso i valori medi, riducendo la presenza di estremi rispetto ai dati originali. Questa tendenza è confermata anche dai grafici delle distribuzioni sovrapposte (vedi Figura 5), dove si osserva come le distribuzioni dei dati sintetici siano più compatte e meno estese rispetto a quelle reali.

Tabella 5: Confronto tra le statistiche dei dati reali e sintetici

Feature	Mean_Real	Mean_Sint	ΔMean	Std_Real	Std_Sint	ΔStd
outcome	0.1393	0.0371	-0.1022	0.3463	0.1891	-0.1572
HighBP	0.4290	0.4913	0.0623	0.4949	0.4999	0.0050
HighChol	0.4243	0.3441	-0.0802	0.4942	0.4751	-0.0192
CholCheck	0.9624	0.9690	0.0066	0.1902	0.1733	-0.0169
BMI	28.3780	26.4376	-1.9404	6.5983	3.1468	-3.4515
Smoker	0.4429	0.4032	-0.0396	0.4967	0.4905	-0.0062
Stroke	0.0404	0.0610	0.0206	0.1969	0.2394	0.0425
HeartDiseaseorAttack	0.0942	0.0736	-0.0207	0.2922	0.2611	-0.0311
PhysActivity	0.7570	0.8009	0.0439	0.4289	0.3993	-0.0296
Fruits	0.6344	0.7001	0.0657	0.4816	0.4582	-0.0234
Veggies	0.8114	0.7812	-0.0302	0.3912	0.4134	0.0223
HvyAlcoholConsump	0.0557	0.0213	-0.0344	0.2294	0.1444	-0.0850
AnyHealthcare	0.9511	0.9495	-0.0016	0.2156	0.2189	0.0034
NoDocbcCost	0.0846	0.0554	-0.0292	0.2783	0.2288	-0.0495
GenHlth	2.5118	2.2902	-0.2216	1.0684	1.1605	0.0921
MentHlth	3.1898	2.0610	-1.1288	7.4176	6.2766	-1.1410
PhysHlth	4.2508	3.5752	-0.6756	8.7256	8.6035	-0.1222
DiffWalk	0.1682	0.1574	-0.0108	0.3740	0.3641	-0.0099
Sex	0.4410	0.5381	0.0971	0.4965	0.4985	0.0020
Age	8.0328	8.4181	0.3853	3.0514	2.5539	-0.4975
Education	5.0499	5.1853	0.1354	0.9864	0.9194	-0.0670
Income	6.0508	6.1802	0.1293	2.0727	2.0893	0.0167

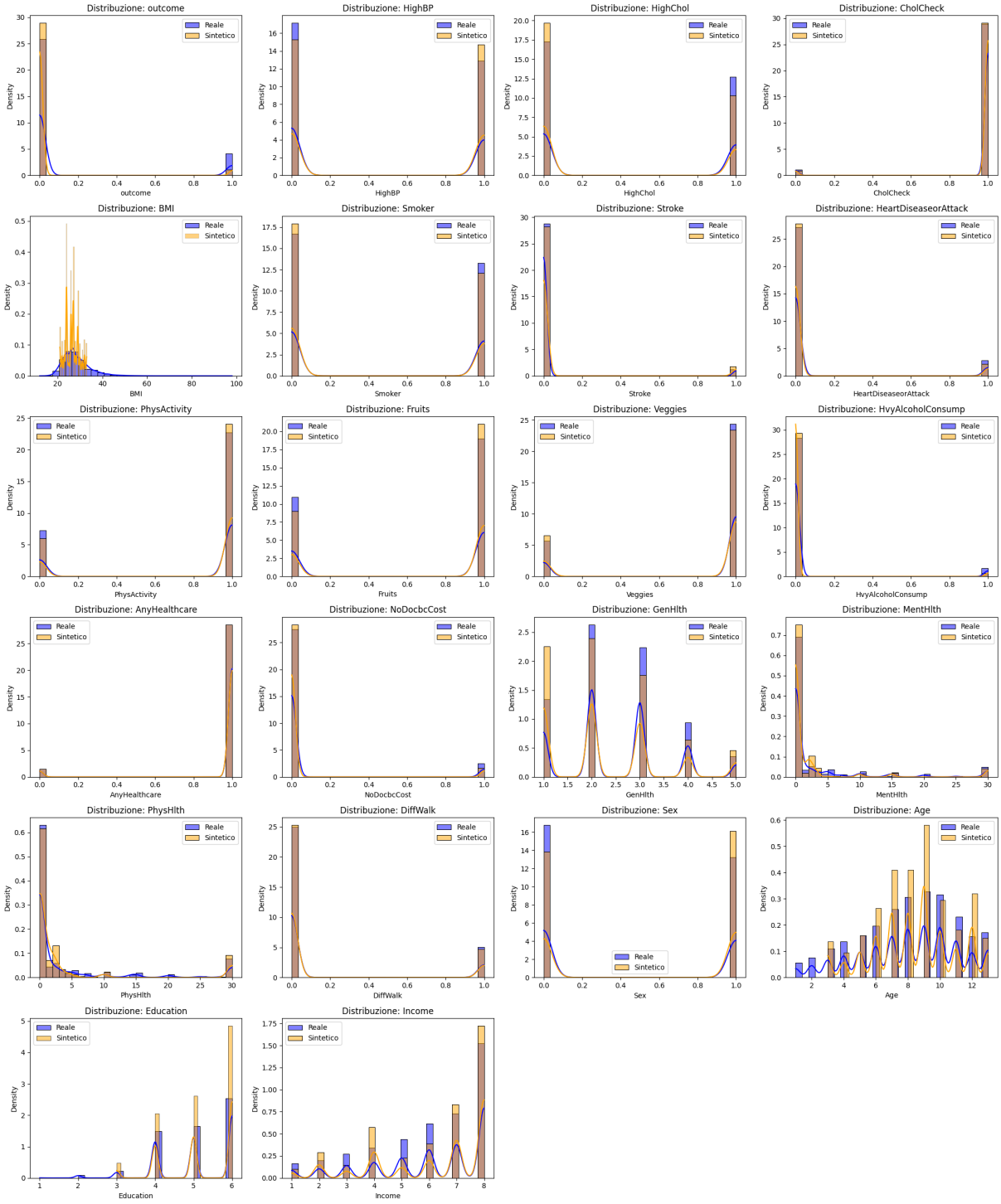

Figura 5: Sovrapposizione tra le distribuzioni delle feature dei dati reali (blu) e sintetici PRIVACY STRONG (giallo).

Tuttavia, nonostante gli scostamenti puntuali delle statistiche descrittive siano moderati, la distanza di Frobenius tra le matrici di correlazione aumenta a **2.03**, indicando una modifica significativa nelle relazioni lineari tra variabili. Questo risultato riflette l'effetto del rumore introdotto per garantire la privacy: mentre le singole statistiche di margine restano simili, le interazioni multivariate tra le variabili vengono parzialmente smussate, con correlazioni leggermente ridotte o alterate.

In sintesi, sebbene le statistiche univariate mostrino una buona resilienza al rumore,

l'elevata distanza di Frobenius evidenzia come la privacy strong agisca da regolarizzatore aggressivo, portando a una semplificazione del dataset sintetico che tende a gravitare attorno ai valori medi, trasformando la struttura delle correlazioni originali.

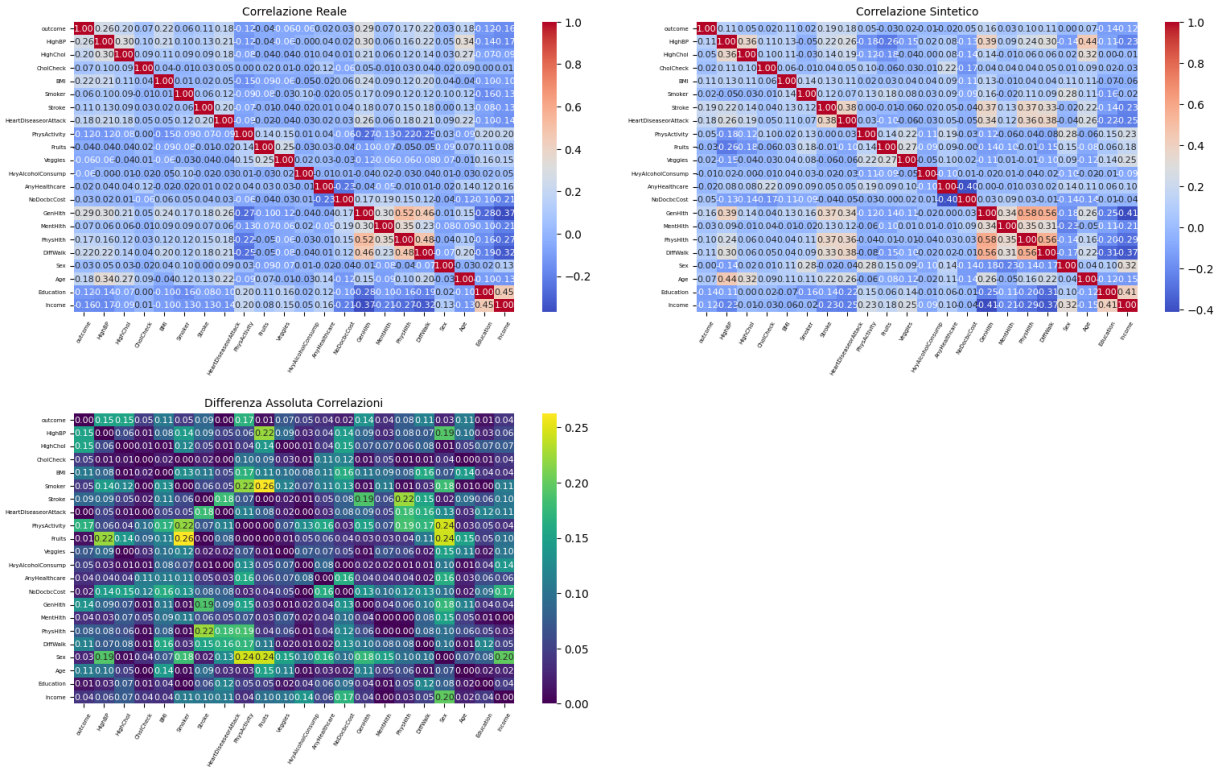

Figura 6: Matrici di correlazione dati reali (prima), sintetici PRIVACY STRONG(seconda) e la loro differenza (terza).

5.4 Sintesi comparativa

La Tabella 6 riassume quantitativamente l'impatto progressivo delle garanzie di privacy sulla qualità statistica dei dataset sintetici.

Tabella 6: Confronto quantitativo della fedeltà statistica tra i tre livelli di privacy

Metrica	**No Privacy**	**Privacy Moderata**	**Privacy Strong**
Errore medio medie (Δ Mean)	0.358	0.353	0.239
Errore medio std. dev. (Δ Std)	0.253	0.381	0.268
Distanza di Frobenius	1.22	1.64	2.03

L'osservazione dei risultati evidenzia una dinamica complessa tra l'accuratezza dei singoli valori e la conservazione della struttura relazionale del dataset. Il fenomeno più interessante emerge dall'analisi dell'errore medio sulle medie (Δ Mean), dove si riscontra un andamento non lineare. Paradossalmente, il dataset No Privacy mostra un errore superiore rispetto alla configurazione Privacy Strong. Questo accade perché, in assenza di vincoli, il modello generativo tende a essere instabile, tentando di rincorrere ogni singola fluttuazione e valore anomalo del dataset originale, finendo per spostare il baricentro delle variabili. Al contrario, l'applicazione di una protezione forte introduce un rumore che funge da regolarizzatore: impedendo al modello di apprendere i dettagli più rari e

specifici, lo costringe a ripiegare sui trend centrali e più frequenti, producendo medie che, seppur meno ricche di dettagli, risultano numericamente più vicine a quelle reali. Tuttavia, questa stabilità delle medie viene pagata con un peggioramento dell'errore sulla deviazione standard (Δ Std). Questo fenomeno è dovuto alla natura stessa della Differential Privacy, che tende a "schiacciare" le distribuzioni verso il centro per proteggere l'anonimato degli individui che presentano valori estremi.

Un'ulteriore conferma della degradazione qualitativa arriva dalla Distanza di Frobenius: in questo caso, osserviamo una crescita costante dell'errore all'aumentare della privacy. Le relazioni tra le variabili sono le informazioni più fragili e difficili da sintetizzare; il rumore introdotto per la protezione dei dati agisce come un interferente che "rompe" i legami tra le colonne. Se nel caso No Privacy il modello riesce a ricostruire molto bene le correlazioni, nella configurazione Strong queste connessioni tendono a perdersi. In conclusione, si osserva che sebbene la Privacy Strong offra un'ottima tenuta sui valori medi centrali, essa compromette la struttura profonda e la variabilità del dataset. La scelta del livello di privacy non è dunque solo una questione di sicurezza, ma determina quanto fedelmente il dataset sintetico potrà sostituire quello reale nelle analisi predittive avanzate. Questi risultati sollevano la questione critica affrontata nelle sezioni successive: in che misura queste alterazioni statistiche si traducono in una perdita di utilità pratica per l'addestramento di modelli predittivi? E quale configurazione rappresenta il punto di equilibrio ottimale tra protezione della privacy e preservazione dell'utilità?

6 Valutazione dell'utilità mediante modelli predittivi

La valutazione dell'utilità dei dataset sintetici costituisce un aspetto cruciale per determinare la loro applicabilità pratica in contesti di ricerca medica. A differenza delle metriche statistiche descrittive, che misurano la somiglianza strutturale tra dati reali e sintetici, l'utilità pratica si manifesta nella capacità dei dati sintetici di supportare l'addestramento di modelli predittivi performanti. Questa sezione presenta un'analisi sistematica delle prestazioni di diversi algoritmi di machine learning addestrati esclusivamente sui dataset sintetici e valutati su dati reali mai visti.

6.1 Metodologia sperimentale

6.1.1 Protocollo di valutazione

Il protocollo sperimentale adottato riflette un caso d'uso realistico: un ricercatore dispone esclusivamente di dati sintetici per l'addestramento e deve sviluppare un modello capace di generalizzare su pazienti reali. Questo scenario è particolarmente rilevante in contesti dove i dati originali non possono essere condivisi per vincoli di privacy o regolamentari.
La procedura di valutazione si articola nelle seguenti fasi:

1. **Addestramento sui dati sintetici**: ciascun modello di classificazione viene addestrato utilizzando uno dei tre dataset sintetici (no privacy, privacy moderata, privacy strong), contenenti 202.944 osservazioni ciascuno.

2. **Valutazione su dati reali**: tutti i modelli vengono valutati su uno stesso set di holdout reale, composto da 50.736 pazienti, che non è stato utilizzato né durante

l'addestramento dei modelli generativi né durante quello dei modelli di classificazione. Questa scelta garantisce una misura oggettiva della capacità di generalizzazione dei modelli sui dati non visti.

3. **Confronto con baseline**: le prestazioni dei modelli addestrati sui dati sintetici vengono confrontate con quelle di modelli di riferimento addestrati e testati direttamente sui dati reali (training set e holdout set). Questo confronto serve come scenario ideale di riferimento: permette di simulare il limite massimo di performance e di capire quanto i dati sintetici siano in grado di avvicinarsi a tali prestazioni. In questo modo è possibile valutare in maniera quantitativa l'impatto della sintesi dei dati sulla capacità predittiva, evidenziando eventuali differenze dovute al compromesso tra privacy e utilità.

6.1.2 Modelli valutati

Per garantire robustezza e generalità dei risultati, sono stati valutati 4 algoritmi di machine learning rappresentativi di diverse famiglie metodologiche:

- **Modelli lineari**: Logistic Regression (in seguito LR)
- **Ensemble methods**: Random Forest (in seguito RF), XGBoost (in seguito XGB)
- **Neural networks**: Multi-Layer Perceptron (in seguito MLP)

Questa varietà permette di valutare come diverse assunzioni algoritmiche e capacità di modellazione interagiscano con le caratteristiche dei dati sintetici.

6.1.3 Metriche di performance

Le prestazioni vengono misurate attraverso cinque metriche complementari:

- **Accuracy**: proporzione di predizioni corrette sul totale, fornisce una misura globale ma può essere fuorviante in presenza di classi sbilanciate.
- **Precision**: proporzione di veri positivi tra tutte le predizioni positive, critica in contesti medici per evitare falsi allarmi.
- **Recall (Sensitivity)**: proporzione di veri positivi identificati tra tutti i casi positivi reali, fondamentale per non mancare diagnosi di pazienti malati.
- **F1-Score**: media armonica di precision e recall, bilancia i due aspetti fornendo una misura complessiva della qualità predittiva.
- **ROC-AUC**: area sotto la curva ROC, misura la capacità del modello di discriminare tra classi indipendentemente dalla soglia di decisione scelta, particolarmente robusta in presenza di sbilanciamento.

6.1.4 Gestione dello sbilanciamento delle classi

Il dataset analizzato presenta una forte sproporzione tra le classi: la stragrande maggioranza dei campioni appartiene a soggetti non diabetici (circa l'86%). In una situazione del genere, un modello di Machine Learning standard tenderebbe a massimizzare l'accuratezza semplicemente predicendo "non diabetico" per ogni paziente, ignorando di fatto la classe minoritaria. Per ovviare a questo problema e garantire che il sistema sia in grado di identificare correttamente i soggetti a rischio, abbiamo adottato tre diverse strategie di compensazione:

- Pesatura delle classi: per i modelli LR e RF, abbiamo utilizzato il parametro class_weight='balanced'. Questa tecnica non modifica i dati, ma agisce direttamente sulla funzione di costo del modello durante l'addestramento. In pratica, ogni errore commesso sulla classe minoritaria (i diabetici) viene "pagato" dal modello molto più caro rispetto a un errore sulla classe maggioritaria. Questo costringe l'algoritmo a prestare la stessa attenzione a entrambe le categorie, livellando l'importanza dei campioni indipendentemente dalla loro frequenza.

- Bilanciamento dinamico dei pesi: XGB gestisce lo sbilanciamento attraverso il parametro scale_pos_weight. A differenza di un valore statico, abbiamo scelto di calcolarlo dinamicamente come il rapporto tra il numero di esempi negativi e positivi ($N_{negativi}/N_{positivi}$). Poiché XGB è un algoritmo basato su alberi di decisione potenziati (boosting), questo coefficiente permette di dare un peso maggiore ai gradienti associati alla classe dei diabetici, assicurando che le iterazioni successive del modello si concentrino sul recupero degli errori fatti sui casi meno frequenti.

- Random OverSampling: il MLP, essendo una rete neurale, beneficia spesso di una gestione dello sbilanciamento a livello di dati piuttosto che di pesi. Abbiamo quindi applicato il Random OverSampling sul set di addestramento. Questa tecnica consiste nel duplicare casualmente i campioni della classe minoritaria finché non raggiungono lo stesso numero di quelli della classe maggioritaria. In questo modo, durante le epoche di addestramento, la rete neurale "vede" i casi di diabete con la stessa frequenza dei casi sani, imparando a riconoscerne i pattern specifici senza essere sopraffatta dalla prevalenza della classe 0.

Tutti questi approcci hanno lo scopo finale di aumentare la recall, ovvero la capacità del sistema di non farsi sfuggire i casi reali di diabete, obiettivo primario in un contesto di screening medico.

6.2 Risultati sperimentali

La Tabella 7 presenta un'analisi comparativa delle performance di diversi modelli di classificazione addestrati su dati reali e su dati sintetici generati con differenti livelli di protezione della privacy. L'obiettivo dell'analisi è valutare in che misura i dati sintetici riescano a preservare l'utilità predittiva dei dati reali e come tale utilità vari all'aumentare delle garanzie di privacy.

Baseline – addestramento su dati reali
I modelli addestrati direttamente sui dati reali rappresentano il riferimento teorico di massima utilità. In questo scenario, LR, XGB e MLP mostrano valori di ROC-AUC comparabili e relativamente elevati (circa 0.82), indicando una buona capacità discriminante.

Tabella 7: Confronto performance modelli di classificazione

Modello	Accuracy	ROC-AUC	Precision	Recall	F1-Score
Baseline - Addestramento su dati reali					
LR	0.7317	0.8196	0.3108	0.7608	0.4413
RF	0.8577	0.7921	0.4685	0.1601	0.2387
XGB	0.7284	0.8218	0.3100	0.7745	0.4428
MLP	0.7139	0.8241	0.3013	0.7987	0.4376
Addestramento su dati sintetici NO PRIVACY					
LR	0.7673	0.8055	0.3307	0.6545	0.4394
RF	0.8486	0.7561	0.4139	0.2077	0.2766
XGB	0.7854	0.7858	0.3408	0.5784	0.4289
MLP	0.7771	0.8010	0.3380	0.6257	0.4389
Addestramento su dati sintetici PRIVACY MODERATA					
LR	0.7293	0.7993	0.3025	0.7222	0.4264
RF	0.8470	0.7521	0.3582	0.1238	0.1840
XGB	0.7825	0.7319	0.3017	0.4265	0.3534
MLP	0.7470	0.7502	0.2851	0.5412	0.3735
Addestramento su dati sintetici PRIVACY STRONG					
LR	0.7002	0.7351	0.2616	0.6316	0.3699
RF	0.8577	0.6757	0.2246	0.0088	0.0169
XGB	0.7736	0.6569	0.2480	0.3075	0.2746
MLP	0.7088	0.6978	0.2450	0.5236	0.3338

RF ottiene l'accuracy più alta, ma a fronte di un recall molto basso, evidenziando una strategia fortemente conservativa che penalizza la capacità di individuare correttamente la classe positiva.

Addestramento su dati sintetici senza privacy
Nel caso dei dati sintetici generati senza vincoli di privacy, le performance risultano complessivamente comparabili alla baseline, e in alcuni casi addirittura superiori in termini di accuracy. Tuttavia, la ROC-AUC tende a ridursi leggermente rispetto ai dati reali, indicando una perdita marginale di capacità discriminante. Questo comportamento è coerente con l'ipotesi che i dati sintetici non protetti riescano a catturare buona parte della struttura statistica dei dati originali, pur introducendo inevitabilmente una certa approssimazione. I valori di precision, recall e F1-score rimangono allineati alla baseline, suggerendo che l'utilità dei dati sintetici, in assenza di privacy, sia sostanzialmente preservata.

Addestramento su dati sintetici con privacy moderata
Con l'introduzione di un livello di privacy moderato, emerge un deterioramento più evidente delle prestazioni, in particolare per i modelli più complessi. La ROC-AUC diminuisce in modo più marcato per XGB e MLP, segnalando una riduzione della capacità di apprendere pattern discriminanti affidabili. Anche il recall tende a ridursi, soprattutto per RF e XGB, con un conseguente calo del F1-score. Ciononostante, LR mantiene performance relativamente stabili, suggerendo che modelli più semplici e lineari risultino maggiormente robusti alla distorsione statistica introdotta dai meccanismi di privacy.

Addestramento su dati sintetici con privacy strong
Lo scenario di privacy elevata mette in evidenza il trade-off più critico tra protezione

dei dati e utilità predittiva. Le performance mostrano un peggioramento generalizzato in termini di ROC-AUC e F1-score, con un impatto particolarmente marcato su RF, che presenta un recall quasi nullo. Ciò indica che l'elevato rumore introdotto per garantire una forte protezione della privacy può compromettere la capacità di apprendere pattern informativi, soprattutto per modelli ensemble. Al contrario, modelli più semplici o più robusti al rumore, come LR, XGB e MLP, mostrano una degradazione più graduale, mantenendo una discreta capacità predittiva.

Nel complesso, i risultati confermano l'esistenza di un chiaro trade-off privacy–utilità. I dati sintetici senza privacy riescono a simulare in modo efficace le prestazioni ottenibili sui dati reali, costituendo un buon proxy per la valutazione dei modelli. All'aumentare delle garanzie di privacy, l'utilità predittiva decresce progressivamente, con un impatto più marcato sui modelli complessi e sulle metriche sensibili alla classe minoritaria. Tali evidenze suggeriscono che, in contesti applicativi reali, la scelta del livello di privacy debba essere attentamente bilanciata rispetto agli obiettivi desiderati.

In particolare, lo studio si concentra sull'analisi della ROC-AUC, che risulta fondamentale per comprendere l'utilità dei dati sintetici, perché non dipende dalla distribuzione delle classi o da soglie di classificazione specifiche. A differenza di metriche come accuracy o precision, la ROC-AUC valuta in modo complessivo quanto il modello sia in grado di discriminare correttamente tra le classi, anche in presenza di dati sbilanciati. Focalizzarsi sulla ROC-AUC è quindi utile per capire quanto i dati sintetici mantengano le informazioni essenziali necessarie a costruire modelli predittivi affidabili. Se un modello addestrato su dati sintetici mantiene un ROC-AUC vicino a quello ottenuto sui dati reali, significa che i dati sintetici conservano gran parte della capacità discriminativa, garantendo così utilità elevata pur rispettando vincoli di privacy.

I seguenti grafici mostrano l'andamento della metrica ROC-AUC per i quattro modelli considerati in funzione del tipo di dati utilizzati per l'addestramento: dati reali, dati sintetici senza privacy, dati sintetici con privacy moderata e dati sintetici con privacy strong. La linea tratteggiata rossa in ciascun grafico rappresenta il valore di ROC-AUC ottenuto dal modello addestrato sui dati reali, fungendo da baseline.

Per LR, la ROC-AUC rimane relativamente stabile passando dai dati reali (0.8196) ai dati sintetici senza privacy (0.8055) e con privacy moderata (0.7993). Solo con privacy strong si osserva un calo più marcato (0.7351), ma il modello mantiene comunque una buona capacità discriminante. Questo indica che LR è robusto ai dati sintetici e alle limitazioni imposte dalla privacy.

Per RF, si nota un trend decrescente simile, ma il modello perde maggiore discriminatività con la privacy strong, passando da 0.7921 dei dati reali a 0.6757. Ciò suggerisce che RF è più sensibile alla degradazione delle informazioni introdotta dalla privacy rispetto a LR. XGB mostra buone performance sui dati reali (0.8218), ma il calo con l'introduzione della privacy è più evidente: con privacy moderata la ROC-AUC scende a 0.7319 e con privacy strong a 0.6569. Questo indica che XGB, pur performando bene sui dati originali, è sensibile alla perturbazione dei dati dovuta alle tecniche di privacy.

Infine, MLP ha un comportamento simile a LR in termini di ROC-AUC, mantenendo valori relativamente alti fino alla privacy moderata (0.7502) e registrando un calo maggiore solo con privacy strong (0.6978). Anche in questo caso, il modello mostra una certa robustezza rispetto a modelli più complessi come RF e XGB.

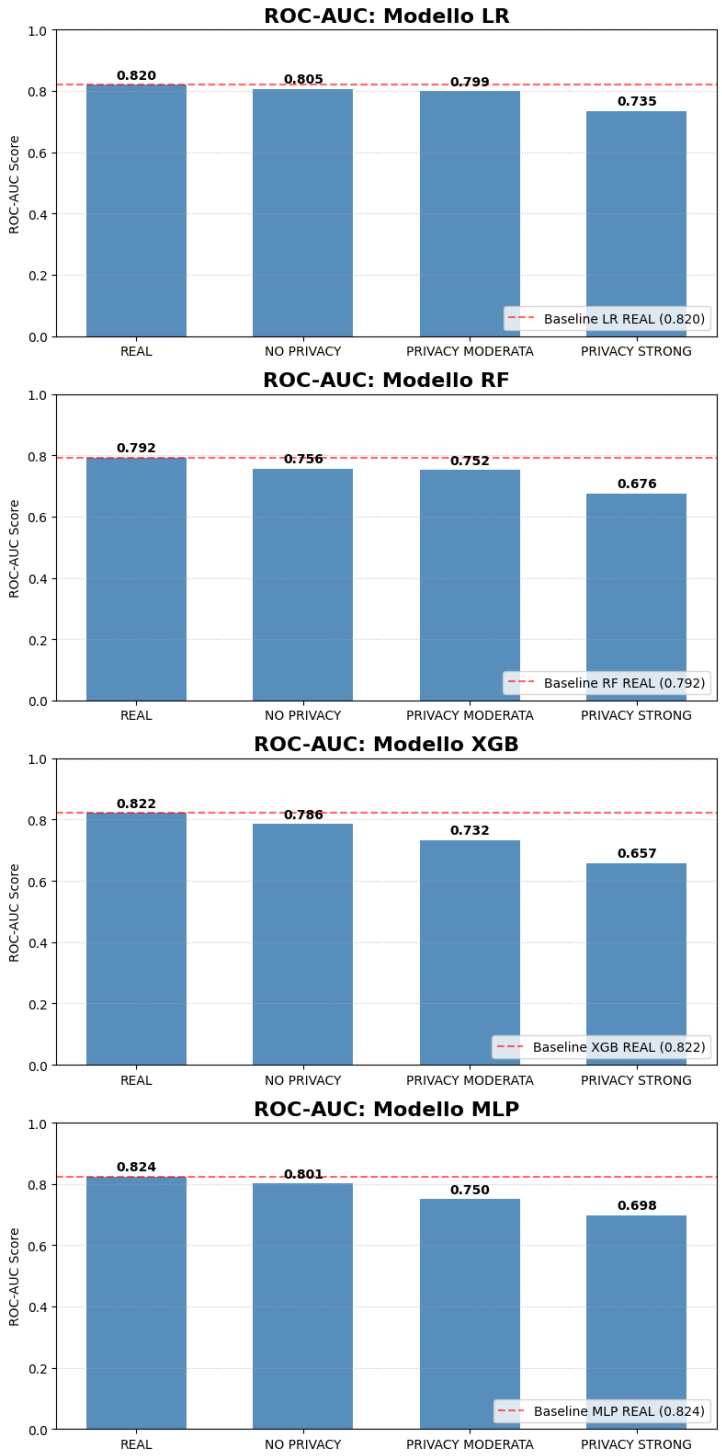

Figura 7: Variazione della ROC-AUC al variare del livello di privacy dei modelli generativi

In sintesi, i grafici evidenziano chiaramente che LR e MLP sono i modelli più stabili in termini di ROC-AUC quando si passa dai dati reali a quelli sintetici e quando si applicano livelli crescenti di privacy. Al contrario, RF e XGB, pur avendo ottime prestazioni sui dati reali, subiscono cali più significativi, indicando una maggiore sensibilità alla riduzione di informazioni rilevanti.

7 Valutazione della resistenza agli attacchi sulla privacy

Mentre le sezioni precedenti hanno quantificato l'utilità pratica dei dataset sintetici attraverso metriche statistiche e prestazioni predittive, è fondamentale verificare empiricamente le garanzie di privacy offerte. Un dataset sintetico di alta qualità statistica potrebbe involontariamente codificare informazioni che permettono di identificare pazienti specifici del training set originale. Questa sezione presenta un'analisi sistematica della resistenza dei dataset sintetici agli attacchi alla privacy, verificando se le garanzie teoriche della privacy differenziale si traducono in protezione empirica misurabile. Per verificare il grado di protezione dei dati e la robustezza del sistema contro potenziali fughe di informazioni, è stato condotto un Membership Inference Attack (MIA). Si tratta di una classe di attacchi alla privacy che mira a determinare se un determinato record individuale abbia fatto parte del dataset di addestramento di un modello o, nel caso dei dati sintetici, del dataset reale utilizzato per generare tali dati. Nel contesto dei dataset sanitari, questo tipo di attacco è particolarmente rilevante, poiché la capacità di inferire l'appartenenza di un paziente al training set può comportare una violazione diretta della riservatezza individuale, anche in assenza di una ricostruzione esplicita dei dati originali.

L'intuizione alla base del MIA è che un generatore sintetico che "memorizza" eccessivamente il dataset di addestramento produca campioni che rappresentano in modo più accurato i record visti durante il training rispetto a quelli mai osservati. Di conseguenza, un attaccante può sfruttare differenze statistiche o geometriche tra record membri e non-membri per inferirne l'appartenenza.

Nel presente lavoro, il MIA è utilizzato come strumento empirico di valutazione della privacy, con l'obiettivo di verificare se e in che misura le garanzie teoriche fornite dai meccanismi di privacy differenziale si traducano in una riduzione effettiva del rischio di inferenza dell'appartenenza.

7.1 Metodologia dell'attacco

L'attacco è formulato come un problema di classificazione supervisionata binaria, in cui l'attaccante cerca di predire, per ciascun record reale x, se esso appartenga al dataset di addestramento del generatore sintetico (membro) oppure a un insieme di dati mai utilizzati durante il training (non-membro).

- Membri: record appartenenti al dataset reale di training (diabetes_train.csv).
- Non-membri: record appartenenti al dataset reale di holdout (diabetes_holdout.csv).

Per ciascun generatore sintetico considerato (senza privacy, privacy moderata, privacy strong), viene costruito un dataset di attacco indipendente, consentendo un confronto diretto della resistenza alla privacy nei diversi scenari.

7.1.1 Feature estratte

L'attaccante non ha accesso diretto al processo di generazione, ma osserva esclusivamente il dataset sintetico prodotto. A partire da esso, vengono estratte per ogni record reale (train e holdout) una serie di feature che quantificano quanto bene tale record è rappresentato dal dataset sintetico. In particolare, vengono utilizzate tre famiglie di feature complementari:

1. **Distanze k-NN (Nearest Neighbors)**: Per ogni record reale, si calcola la distanza euclidea rispetto ai campioni sintetici più vicini nello spazio delle feature, dopo standardizzazione. In particolare:

   - la distanza minima dal campione sintetico più vicino;
   - la distanza media dai $k = 5$ vicini più prossimi.

   L'ipotesi è che i record membri risultino mediamente più vicini ai dati sintetici rispetto ai non-membri.

2. **Local density** (radius=1): viene misurato il numero di campioni sintetici presenti entro un raggio fissato nello spazio delle feature. Una densità locale elevata indica che il generatore produce molti campioni in prossimità del record reale, suggerendo una possibile memorizzazione implicita.

3. **Reconstruction error**: per stimare quanto bene il generatore catturi le relazioni interne tra le feature, viene addestrato un insieme di regressori sui dati sintetici, ciascuno incaricato di predire una feature a partire dalle restanti. L'errore medio di ricostruzione sui record reali fornisce una misura indiretta della loro compatibilità con la distribuzione sintetica: errori più bassi indicano record meglio rappresentati.

Intuizione: Record che erano nel training originale dovrebbero essere meglio rappresentati dal sintetico (distanze minori, densità maggiore, errore di ricostruzione minore).

7.1.2 Modello dell'attaccante

Sulle feature così costruite viene addestrato un classificatore XGBoost, scelto per la sua capacità di modellare relazioni non lineari e combinare efficacemente feature eterogenee. Il dataset di attacco viene suddiviso in training e test set mantenendo il bilanciamento tra membri e non-membri. Il classificatore viene quindi addestrato per predire:

- $y = 1$: record era nel training del modello che ha generato i dati sintetici (membro)
- $y = 0$: record era nell'holdout (non-membro)

Le prestazioni dell'attacco sono valutate tramite:

- Accuracy, che misura la correttezza complessiva della predizione;
- ROC-AUC, che rappresenta la metrica principale, in quanto indipendente dalla soglia di decisione.

Un valore di ROC-AUC prossimo a 0.5 indica un attacco non migliore del caso casuale, mentre valori significativamente superiori segnalano una perdita di privacy misurabile.

La perdita di privacy viene definita come: $PrivacyLoss = AUC - 0.5$ che rappresenta il vantaggio informativo dell'attaccante rispetto a una scelta casuale. Valori positivi indicano fuga di informazioni, di conseguenza, questa formulazione consente di confrontare in modo diretto i diversi generatori sintetici e di verificare empiricamente se l'introduzione di meccanismi di privacy differenziale riduca effettivamente la vulnerabilità agli attacchi di membership inference.

7.2 Risultati dell'attacco

Tabella 8: Risultati Membership Inference Attack

Generatore	Accuracy	ROC-AUC	Privacy Loss
No Privacy	0.694	0.808	0.308
Privacy Moderata ($\varepsilon = 4.0$)	0.613	0.698	0.198
Privacy Strong ($\varepsilon = 2.0$)	0.493	0.492	-0.008

La Tabella 8 riporta i risultati del Membership Inference Attack applicato ai dataset sintetici generati con diversi livelli di protezione della privacy. L'analisi consente di valutare in modo empirico la vulnerabilità dei generatori sintetici all'inferenza di appartenenza e di verificare se l'introduzione della privacy differenziale riduca effettivamente il rischio di leakage informativo.

1. CTGAN senza privacy (ROC-AUC 0.808)
Nel caso del generatore privo di meccanismi di privacy, l'attacco ottiene valori elevati sia di accuracy (0.694) sia di ROC-AUC (0.808), con una privacy loss pari a 0.308. Tali risultati indicano che l'attaccante è in grado di distinguere in modo significativo tra record membri e non-membri, ben oltre il livello casuale. Questo comportamento suggerisce che il generatore sintetico, pur producendo dati di buona qualità statistica, conserva informazioni sufficienti a rendere riconoscibili i campioni appartenenti al training set originale, evidenziando una sostanziale vulnerabilità dal punto di vista della privacy.

2. Privacy Moderata (ROC-AUC 0.698)
Con l'introduzione di un livello moderato di privacy differenziale, si osserva una riduzione consistente delle prestazioni dell'attacco. La ROC-AUC scende a 0.698 e la privacy loss si riduce a 0.198, indicando un vantaggio informativo dell'attaccante sensibilmente inferiore rispetto al caso senza privacy. Tuttavia, l'attacco rimane comunque efficace, segnalando che, sebbene il rumore introdotto attenui la memorizzazione dei dati reali, una parte dell'informazione utile all'inferenza dell'appartenenza è ancora presente. Questo risultato è coerente con l'idea che la privacy moderata rappresenti un compromesso tra protezione e utilità, riducendo ma non annullando il rischio di leakage.

3. Privacy Strong (ROC-AUC 0.492)
Nel caso di privacy elevata, le prestazioni dell'attacco collassano verso il livello casuale. L'accuracy e il ROC-AUC assumono valori intorno al 0.49 prossima a 0.5, con una privacy loss leggermente negativa. Ciò indica che l'attaccante non è in grado di distinguere in modo affidabile tra membri e non-membri, e che il dataset sintetico non fornisce informazioni sfruttabili per l'inferenza di appartenenza. Questo risultato conferma empiricamente l'efficacia delle garanzie di privacy differenziale strong nel mitigare il rischio di Membership Inference Attack. A questo livello, il dataset sintetico non rivela informazioni sui membri del training originale.

Nel complesso, i risultati mostrano una chiara relazione tra il livello di privacy e la resistenza agli attacchi di membership inference. All'aumentare delle garanzie di privacy, il vantaggio informativo dell'attaccante diminuisce progressivamente fino ad annullarsi. Tali evidenze rafforzano le conclusioni emerse dalle analisi di utilità: la protezione strong garantisce un'elevata robustezza agli attacchi sulla privacy, ma a costo di una riduzione delle prestazioni predittive, mentre la privacy moderata offre un compromesso intermedio, in cui parte dell'utilità è preservata a fronte di un rischio di privacy ancora non nullo.

8 Trade-off Privacy-Utilità per identificare il bilanciamento ottimale

Per questa analisi del trade-off tra Privacy-Utilità è stato scelto il modello di LR in quanto si è dimostrato il più robusto tra le architetture testate. Come emerge dai risultati sperimentali, la LR non solo garantisce le performance migliori in termini di equilibrio tra precisione e recall, ma presenta soprattutto una degradazione delle ROC-AUC molto più contenuta all'aumentare dei vincoli di privacy.
Il grafico 8 illustra l'analisi del trade-off tra utilità clinica dei dati sintetici e rischio per la privacy per il modello LR, fornendo una sintesi visiva fondamentale per la validazione dei dati sintetici generati.

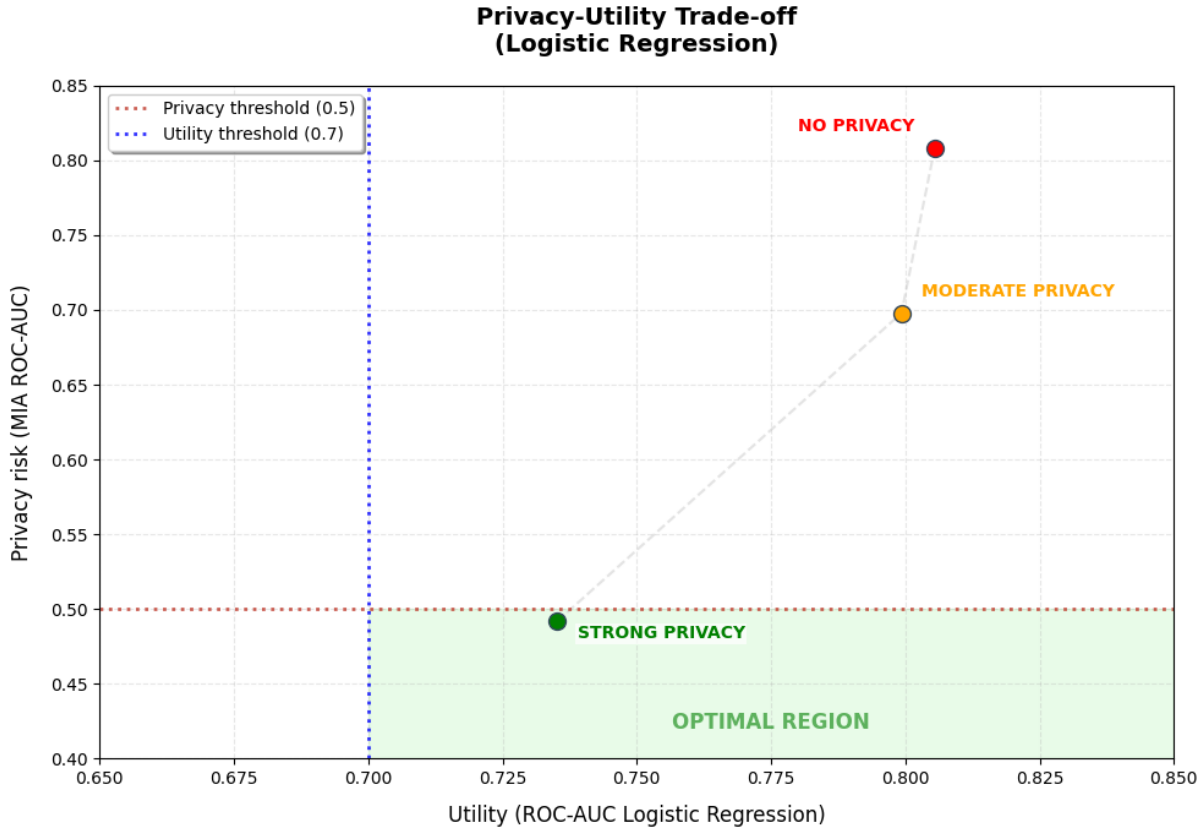

Figura 8: Trade-off Privacy-Utilità

L'immagine mette in relazione due metriche contrapposte: sull'asse delle ascisse è riportata l'utilità medica, misurata tramite la ROC-AUC del classificatore, mentre sull'asse delle ordinate è indicato il rischio di violazione della privacy, quantificato dall'efficacia di un Membership Inference Attack (ROC-AUC dell'attaccante).

Per guidare l'interpretazione dei risultati, sono state stabilite due linee di demarcazione strategiche:

- **Soglia di utilità accettabile (ROC-AUC = 0.70)**: rappresenta il limite minimo affinché il modello sia considerato clinicamente rilevante. Un valore superiore a questa soglia indica che i dati sintetici preservano i segnali necessari per uno screening efficace.

- **Soglia di privacy ottimale (ROC-AUC = 0.50)**: corrisponde alla linea di "casualità" per un attacco MIA. Un valore prossimo a 0.50 indica che un utente malintenzionato non ha capacità superiori al caso nel distinguere tra soggetti reali e sintetici, garantendo la massima protezione dell'anonimato.

Il grafico mostra chiaramente come l'aumento delle restrizioni di privacy sposti le performance lungo una curva discendente:

- No privacy: garantisce un'ottima utilità (ROC-AUC > 0.80), ma espone i dati a un rischio di violazione della privacy critico (MIA ROC-AUC ≈ 0.80).
- Privacy moderata: rappresenta un punto in cui l'utilità rimane eccellente, mentre il rischio MIA subisce una drastica riduzione, pur rimanendo sopra la soglia di sicurezza assoluta.
- Privacy strong: si posiziona all'interno dell'area ottimale (evidenziata in verde). In questa configurazione, il rischio di privacy crolla sotto lo 0.50, annullando l'efficacia degli attacchi di inferenza, mentre l'utilità clinica si mantiene stabilmente sopra lo 0.73.

La scelta di evidenziare l'*optimal region* nel quadrante in basso a destra serve a identificare la configurazione di successo per la domanda di ricerca. La conclusione principale desumibile dall'immagine è che la configurazione privacy strong permette di soddisfare simultaneamente i requisiti etici di protezione del paziente e i requisiti tecnici di accuratezza diagnostica. Nonostante l'inevitabile degradazione delle performance rispetto ai dati reali, il modello mantiene una capacità predittiva solida e sicura, rendendo il dataset sintetico idoneo per l'utilizzo in contesti di ricerca medica regolamentati.

È fondamentale sottolineare che i risultati presentati derivano specificamente dal dataset utilizzato in questa analisi e che ogni scenario applicativo richiede una valutazione puntuale e dedicata. Di seguito sono riportate alcune considerazioni essenziali per contestualizzare correttamente questi esiti:

- **Relatività del dataset**: i risultati ottenuti, inclusi i livelli di resilienza dei modelli alla Differential Privacy, sono legati alla distribuzione e alle caratteristiche del nostro dataset di riferimento. Non è possibile assumere che le medesime configurazioni di privacy producano lo stesso equilibrio tra utilità e rischio su dataset con diverse cardinalità, numero di feature o livelli di rumore intrinseco.
- **Variabilità della soglia "accettabile"**: sebbene nel presente studio la soglia di utilità accettabile sia stata fissata a 0.70, questo valore non è universale. Quindi anche la definizione di *optimal region* non è statica, ma rappresenta un compromesso dinamico che deve essere ricalibrato in base alla tolleranza al rischio (quanto è grave un errore di predizione?) e ai requisiti legali/etici di confidenzialità (quanto è sensibile l'informazione trattata?).

Pertanto, l'approccio qui descritto funge da framework metodologico piuttosto che da regola fissa, evidenziando la necessità di una supervisione umana esperta nella scelta finale dei parametri di privacy e utilità.

9 Conclusioni

9.1 Risposte alle domande di ricerca

RQ1: Quali modelli di machine learning mantengono le migliori prestazioni quando addestrati su dati sintetici, rispetto alla baseline ottenuta con dati reali?
L'analisi su dataset ha prodotto risultati chiari:

1. **LR** emerge come modello più robusto:
   - Con $\varepsilon = 4.0$: ROC-AUC 0.7993 (-2.5% vs baseline)
   - Con $\varepsilon = 2.0$: ROC-AUC 0.7351 (-10.3% vs baseline)
   - Resilienza superiore grazie a assunzioni lineari semplici
2. **XGB** mantiene prestazioni discrete ma degrada più rapidamente:
   - Con $\varepsilon = 4.0$: ROC-AUC 0.7319 (-10.9% vs baseline)
   - Con $\varepsilon = 2.0$: ROC-AUC 0.6569 (-20.1% vs baseline)
3. **RF** mostra vulnerabilità critica:
   - Recall crolla già con $\varepsilon = 4.0$ (0.1238)
   - Praticamente inutilizzabile con $\varepsilon = 2.0$ (recall 0.0088)
4. **MLP** mostra un comportamento intermedio:
   - Con $\varepsilon = 4.0$: ROC-AUC pari a 0.7502 (–9.0% rispetto alla baseline)
   - Con $\varepsilon = 2.0$: ROC-AUC pari a 0.6978 (–15.3% rispetto alla baseline)
   - Pur subendo una degradazione progressiva delle prestazioni, il modello riesce a preservare una capacità predittiva non trascurabile, grazie alla sua flessibilità, risultando tuttavia meno robusto rispetto a LR.

Alla luce dei risultati ottenuti, si raccomanda l'impiego di modelli a complessità contenuta, in particolare la LR, quando si opera su dati sintetici generati con meccanismi di privacy differenziale. Tali modelli dimostrano una maggiore robustezza al rumore introdotto per la protezione della privacy, garantendo un compromesso più equilibrato tra utilità predittiva e tutela della riservatezza. Al contrario, l'utilizzo di modelli ensemble o fortemente non lineari dovrebbe essere valutato con cautela in scenari di privacy elevata, poiché la perdita di informazione strutturale può comprometterne significativamente l'efficacia. In contesti applicativi reali, la scelta del modello e del livello di privacy dovrebbe pertanto essere guidata da un'analisi congiunta dei requisiti di protezione dei dati e delle prestazioni attese, privilegiando soluzioni che assicurino una protezione empiricamente verificabile senza annullare l'utilità del dato sintetico.

RQ2: È possibile conciliare tutela della privacy e utilità dei dati sintetici al punto da renderli un'alternativa affidabile ai dati clinici reali nella ricerca medica e nelle applicazioni di machine learning?
La risposta è **sì, con calibrazione accurata del budget** ε.

9.2 Conclusione

L'adozione di dati sanitari sintetici rappresenta una promettente via per conciliare l'imperativo etico della protezione della privacy con le esigenze scientifiche dell'analisi dati nella ricerca medica. Questo lavoro dimostra che, attraverso una calibrata applicazione della privacy differenziale, è possibile generare dataset che offrono garanzie formali di riservatezza pur mantenendo un'utilità analitica accettabile. Il compromesso ottimale individuato costituisce un punto di riferimento per i ricercatori, evidenziando che la scelta tra privacy e utilità non è necessariamente binaria, ma può essere modulata lungo un continuum dove livelli intermedi offrono il miglior bilanciamento di valori concorrenti. La strada verso dataset sintetici clinicamente validi e eticamente solidi potrebbe richiedere ulteriori ricerche interdisciplinari, ma i risultati presentati confermano il potenziale trasformativo di questo approccio per il futuro della ricerca biomedica basata sui dati.

10 Studi esaminati

10.1 Avatar Method (Nature Digital Medicine, 2023)

Link: `https://www.nature.com/articles/s41746-023-00771-5`

Avatar genera dati sintetici patient-centric usando modelli locali invece di un modello globale, permettendo metriche di privacy aggiuntive come local cloaking. Il metodo riduce il rischio di re-identificazione sotto il 5% anche in attacchi sofisticati, risultando particolarmente efficace per dati longitudinali e time-series.

10.2 DP-CTGAN (Fang et al., 2022)

Link: `https://doi.org/10.1007/978-3-031-09342-5_17`

Questo studio incorpora la privacy differenziale in CTGAN (Conditional Tabular GAN) per generare dati medici tabulari sintetici. DP-CTGAN supera i metodi state-of-the-art come PATE-GAN e DPGAN con lo stesso budget di privacy su diversi dataset medici. L'approccio è stato esteso anche al federated learning (FDP-CTGAN) per una generazione più sicura senza necessità di centralizzare i dati.

10.3 Membership Inference Attacks (Zhang et al., 2022)

Link: `https://doi.org/10.1016/j.jbi.2021.103977`

Lo studio introduce un framework per attacchi di membership inference contro dati sintetici sanitari utilizzando contrastive representation learning. I risultati dimostrano che i dati parzialmente sintetici sono altamente vulnerabili (alto tasso di successo dell'attacco), mentre i dati completamente sintetici mostrano vulnerabilità marginale, suggerendo un livello di protezione adeguato nella maggior parte dei casi.

10.4 Synthetic Data for Clinical Risk Prediction (Qian et al., 2024)

Link: `https://doi.org/10.1038/s41598-024-72894-y`

Questo studio dimostra che i dati sintetici generati con tecniche all'avanguardia possono essere utilizzati efficacemente nell'intero pipeline di sviluppo di modelli prognostici

clinici, anche senza accesso finale ai dati reali. Utilizzando UK Biobank per modelli di rischio di cancro al polmone, lo studio conferma che i dati sintetici possono sostituire i dati reali mantenendo performance predittive comparabili con garanzie di privacy.

10.5 HealthGAN (Yale et al., 2020)

Link: `https://doi.org/10.1016/j.neucom.2020.01.041`

HealthGAN introduce un workflow end-to-end per la generazione di dati sanitari sintetici che preservano la privacy, sviluppando metriche innovative per misurare resemblance, privacy, utility e footprint. Il metodo supera approcci precedenti come medGAN su dataset MIMIC-III, dimostrando l'efficacia delle GAN nel dominio sanitario quando adeguatamente progettate.