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 μU/mL (variabile continua, range: 0-846)
- **BMI**: indice di massa corporea, calcolato come peso in kg/(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 $G : \mathcal{Z} \to \mathcal{X}$ 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:

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

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

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`, $\varepsilon = 5.0$): Dopo la generazione tramite CTGAN, viene applicato rumore differenzialmente privato mediante il meccanismo di Laplace con dominio limitato, mediante l'utilizzo del parametro $\varepsilon = 5.0$.
3. **Privacy forte** (`synthetic_privacy_strong.csv`, $\varepsilon = 1.0$): Viene applicato un livello più rigoroso di protezione della privacy attraverso il medesimo meccanismo di Laplace ma con $\varepsilon = 1.0$, 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:

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

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:

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

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`):

$$l_f = \max(0, \min_i(x_i^f)), \quad u_f = \max_i(x_i^f) + 2 \tag{5}$$

  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`):

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

  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:

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

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:

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

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

$$outcome_i \leftarrow \text{clip}(\lfloor outcome_i + 0.5 \rfloor, 0, 1) \tag{9}$$

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:

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

- **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 $\varepsilon = 5.0$
- `synthetic_privacy_strong.csv`: 537 campioni sintetici con $\varepsilon = 1.0$

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:

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

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

"'latex

2.10. Risultati dell'analisi comparativa

I tre dataset sintetici generati sono stati sottoposti ad un'analisi statistica approfondita per quantificare l'impatto delle diverse garanzie di privacy sulla qualità dei dati. Le Figure 1-3 riportano i risultati comparativi per ciascuna configurazione, includendo le distribuzioni delle variabili chiave (glucosio e BMI) e le matrici di correlazione tra tutte le variabili.

Tabella 1. Statistiche descrittive - Dataset senza privacy

Feature	Real Mean	Synth Mean	Mean Err%	Real Std	Synth Std	Std Err%
pregnancies	3.9	3.7	4.9	3.4	3.4	0.0
glucose	121.6	126.1	3.7	30.0	28.4	5.2
blood_pressure	72.2	69.8	3.4	12.2	12.2	0.3
skin_thickness	28.9	29.3	1.4	8.3	6.7	18.9
insulin	138.1	133.3	3.5	79.7	45.5	43.0
bmi	32.6	33.3	2.1	6.7	6.2	7.7
diabetes_pedigree	0.5	0.4	7.4	0.3	0.4	7.7
age	33.6	32.8	2.3	11.8	9.7	17.7
outcome	0.3	0.5	36.4	0.5	0.5	4.8

2.10.1. Dataset senza protezione della privacy

Il dataset sintetico generato senza alcuna protezione della privacy (`synthetic_no_privacy.csv`) mostra un'eccellente preservazione delle proprietà statistiche del dataset originale. L'analisi delle statistiche descrittive (Tabella 1) evidenzia errori medi contenuti per tutte le variabili:

- Le variabili continue principali (`glucose`, `bmi`, `blood_pressure`) presentano errori sulla media inferiori al 4%, indicando una fedele riproduzione dei valori centrali delle distribuzioni.
- La deviazione standard è preservata in modo eccellente per la maggior parte delle variabili, con errori inferiori al 8% per `glucose`, `blood_pressure` e `bmi`.
- Si osserva una maggiore discrepanza per `insulin` (errore 43.0% sulla deviazione standard), attribuibile alla natura altamente variabile e multimodale di questa variabile nel dataset originale.

L'analisi delle correlazioni rivela una differenza media di Frobenius pari a 0.1030, con una discrepanza massima di 0.3817 su singole coppie di variabili. Come mostrato nella Figura 1, le matrici di correlazione reale e sintetica risultano visivamente molto simili, preservando le relazioni chiave come la correlazione positiva tra `glucose` e `outcome` ($\rho_{real} = 0.50$, $\rho_{synth} = 0.17$) e tra `age` e `pregnancies` ($\rho_{real} = 0.53$, $\rho_{synth} = 0.29$).

Le distribuzioni empiriche di `glucose` e `bmi` (pannelli superiori della Figura 1) mostrano una sovrapposizione quasi completa tra dati reali (blu) e sintetici (rosso), confermando la capacità di CTGAN di catturare accuratamente la forma delle distribuzioni univariate.

2.10.2. Dataset con privacy moderata ($\varepsilon = 5.0$)

L'applicazione del meccanismo di Laplace con $\varepsilon = 5.0$ introduce una perturbazione controllata che inizia a manifestare effetti misurabili sulla qualità statistica dei dati, come evidenziato nella Figura 2.

Tabella 2. Statistiche descrittive - Dataset con privacy moderata ($\varepsilon = 5.0$)

Feature	Real Mean	Synth Mean	Mean Err%	Real Std	Synth Std	Std Err%
pregnancies	3.9	5.5	40.6	3.4	4.0	18.1
glucose	121.6	125.0	2.8	30.0	33.7	12.5
blood_pressure	72.2	73.8	2.2	12.2	20.1	65.2
skin_thickness	28.9	31.1	7.6	8.3	11.3	36.5
insulin	138.1	214.1	55.1	79.7	144.5	81.3
bmi	32.6	35.7	9.4	6.7	9.6	42.8
diabetes_pedigree	0.5	0.7	50.4	0.3	0.4	33.0
age	33.6	43.9	31.0	11.8	14.7	24.5
outcome	0.3	0.5	36.4	0.5	0.5	4.8

Gli effetti principali includono:

- **Incremento della variabilità**: l'iniezione di rumore causa un aumento sostanziale della deviazione standard per diverse variabili. In particolare, `insulin` mostra un incremento dell'81.3% ($\sigma_{real} = 79.7$, $\sigma_{synth} = 144.5$), mentre `blood_pressure` registra un aumento del 65.2%.
- **Spostamento delle medie**: alcune variabili presentano uno shift sistematico dei valori medi. Particolarmente significativo è l'incremento di `insulin` (+55.1%, da 138.1 a 214.1) e `age` (+31.0%, da 33.6 a 43.9 anni), suggerendo un bias introdotto dal processo di perturbazione.
- **Degradazione delle correlazioni**: la distanza di Frobenius tra le matrici di correlazione aumenta a 0.1434 (media) con un massimo di 0.5396. Dalla Figura 2 si osserva una perdita parziale delle strutture di dipendenza, con correlazioni che tendono ad attenuarsi rispetto al caso reale.

Nonostante queste alterazioni, le distribuzioni di `glucose` e `bmi` mantengono una forma riconoscibile, sebbene con code più pesanti e maggiore dispersione, particolarmente evidente nella distribuzione del BMI dove compare un picco anomalo attorno al valore 50.

2.10.3. Dataset con privacy forte ($\varepsilon = 1.0$)

La configurazione con $\varepsilon = 1.0$ rappresenta il livello più stringente di protezione della privacy analizzato, e i suoi effetti sulla qualità dei dati sono marcati, come mostrato nella Figura 3.

Tabella 3. Statistiche descrittive - Dataset con privacy forte ($\varepsilon = 1.0$)

Feature	Real Mean	Synth Mean	Mean Err%	Real Std	Synth Std	Std Err%
pregnancies	3.9	7.9	104.6	3.4	4.8	43.5
glucose	121.6	130.0	6.9	30.0	39.6	32.2
blood_pressure	72.2	72.8	0.9	12.2	23.0	88.7
skin_thickness	28.9	32.2	11.5	8.3	13.5	63.6
insulin	138.1	319.8	131.6	79.7	169.1	112.1
bmi	32.6	38.6	18.4	6.7	11.3	67.6
diabetes_pedigree	0.5	1.0	108.4	0.3	0.6	72.0
age	33.6	48.0	43.2	11.8	16.8	42.6
outcome	0.3	0.5	36.4	0.5	0.5	4.8

Le principali caratteristiche sono:

- **Distorsione significativa delle medie**: gli errori percentuali sulle medie raggiungono valori critici per diverse variabili: `pregnancies` +104.6% (da 3.9 a 7.9), `insulin` +131.6% (da 138.1 a 319.8), `diabetes_pedigree` +108.4% (da 0.5 a 1.0). Questi spostamenti indicano una perdita sostanziale di fedeltà rispetto alla distribuzione originale.
- **Amplificazione della varianza**: la deviazione standard subisce incrementi drammatici: `insulin` +112.1%, `blood_pressure` +88.7%, `bmi` +67.6%, `diabetes_pedigree` +72.0%. L'elevata quantità di rumore introdotto rende i dati sintetici significativamente più dispersi rispetto agli originali.
- **Collasso delle strutture di correlazione**: la differenza media di Frobenius aumenta a 0.1671 con un massimo di 0.5566. La matrice di correlazione sintetica (Figura 3, pannello inferiore destro) mostra valori drasticamente ridotti rispetto alla matrice reale, con la maggior parte delle correlazioni che convergono verso zero. Questo fenomeno indica una sostanziale perdita di informazione sulle dipendenze tra variabili.
- **Distorsione delle distribuzioni**: le distribuzioni empiriche di `glucose` e `bmi` (pannelli superiori della Figura 3) presentano forme marcatamente diverse da quelle reali. In particolare, la distribuzione del BMI mostra una forma quasi uniforme con un picco anomalo estremo, indicando una perdita quasi completa della struttura della distribuzione originale.

2.10.4. Sintesi comparativa

La Tabella 4 riassume quantitativamente l'impatto progressivo delle garanzie di privacy sulla qualità statistica dei dataset sintetici.

Tabella 4. Confronto quantitativo tra i tre livelli di privacy

Metrica	No Privacy	$\varepsilon = 5.0$	$\varepsilon = 1.0$
Errore medio medie (%)	7.0	26.0	63.2
Errore medio std. dev. (%)	11.7	36.3	69.1
Dist. Frobenius media	0.103	0.143	0.167
Dist. Frobenius max	0.382	0.540	0.557

I risultati evidenziano chiaramente il trade-off fondamentale tra privacy e utilità: all'aumentare delle garanzie di privacy (epsilon decrescente), si osserva un degrado progressivo e non lineare della qualità statistica dei dati. Il passaggio da assenza di privacy a privacy moderata introduce perturbazioni gestibili, mentre il livello di privacy forte causa alterazioni così marcate da compromettere potenzialmente l'utilità dei dati per applicazioni di machine learning.

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?

3. 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.

3.1. Metodologia sperimentale

3.1.1. 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 viene addestrato utilizzando uno dei tre dataset sintetici (no privacy, privacy moderata $\varepsilon = 5.0$, privacy forte $\varepsilon = 1.0$), contenenti 537 osservazioni ciascuno.
2. **Valutazione su dati reali**: tutti i modelli vengono testati sullo stesso holdout set reale di 231 pazienti, mai utilizzato durante l'addestramento. Questo garantisce una misura oggettiva della capacità di generalizzazione.
3. **Confronto con baseline**: le prestazioni vengono confrontate con modelli di riferimento addestrati direttamente sui dati reali (training set di 537 osservazioni), rappresentando il limite superiore teorico delle prestazioni ottenibili.

3.1.2. Modelli valutati

Per garantire robustezza e generalità dei risultati, sono stati valutati 11 algoritmi di machine learning rappresentativi di diverse famiglie metodologiche:

- **Modelli lineari**: Logistic Regression (LR)

- **Ensemble methods**: Random Forest (RF), Gradient Boosting (GB), AdaBoost (ADA), XGBoost (XGB), LightGBM (LGBM)
- **Tree-based**: Decision Tree (DT)
- **Instance-based**: K-Nearest Neighbors (KNN)
- **Probabilistic**: Naive Bayes (NB)
- **Kernel methods**: Support Vector Machine (SVM)
- **Neural networks**: Multi-Layer Perceptron (MLP)

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

3.1.3. 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.

3.2. Risultati sperimentali

La Tabella 5 riporta i risultati completi di tutti gli 11 algoritmi testati su ciascuno dei tre dataset sintetici. Ogni modello è stato addestrato esclusivamente sui dati sintetici e valutato sull'holdout set reale di 231 pazienti.

3.2.1. Analisi per livello di privacy

Dataset senza privacy (CTGAN)

I modelli addestrati su dati sintetici senza protezione della privacy mostrano prestazioni complessivamente buone, sebbene inferiori alla baseline reale. L'accuracy media si attesta a 0.681, circa 6 punti percentuali sotto la baseline (0.742), mentre il ROC-AUC medio raggiunge 0.725 contro 0.825 della baseline.

Analizzando i risultati per singolo modello, emergono pattern significativi:

- **Support Vector Machine** ottiene le prestazioni migliori con accuracy 0.710 e ROC-AUC 0.758, dimostrando che i kernel methods riescono a catturare efficacemente le relazioni nei dati sintetici.
- **Naive Bayes** raggiunge il miglior F1-Score (0.600) con un eccellente bilanciamento tra precision (0.573) e recall (0.630), suggerendo che le assunzioni di indipendenza condizionale sono compatibili con la struttura dei dati sintetici.
- **Random Forest** ottiene il recall più elevato (0.667), identificando correttamente i due terzi dei casi positivi, superiore anche alla baseline reale (0.519). Questo comportamento indica che i dati sintetici potrebbero enfatizzare alcuni pattern discriminativi.
- **Multi-Layer Perceptron** mostra le prestazioni più deboli (ROC-AUC 0.643), suggerendo che l'ottimizzazione delle reti neurali su dataset sintetici di dimensioni limitate risulta problematica.

Il recall medio elevato (0.616) rispetto alla baseline (0.525) rappresenta una caratteristica distintiva: i modelli addestrati su dati sintetici tendono a essere più sensibili nell'identificazione dei casi positivi, a scapito di una precision leggermente inferiore (0.542 vs 0.669 baseline). Questo trade-off potrebbe essere vantaggioso in applicazioni di screening dove è prioritario minimizzare i falsi negativi.

Dataset con privacy moderata ($\varepsilon = 5.0$)

I risultati per il livello di privacy moderata rivelano un pattern controintuitivo: nonostante l'introduzione di rumore differenziale, il ROC-AUC medio (0.728) risulta superiore sia al dataset senza privacy (0.725) che competitivo rispetto alla baseline (0.825).

L'analisi dettagliata rivela eterogeneità significativa tra algoritmi:

- **Logistic Regression** raggiunge prestazioni eccezionali con ROC-AUC 0.812, appena 2 punti sotto la baseline. La capacità dei modelli lineari di resistere al rumore suggerisce che le relazioni lineari tra features sono robustamente preservate anche con $\varepsilon = 5.0$.
- **SVM, AdaBoost e Naive Bayes** mantengono ROC-AUC superiori a 0.78, confermando che modelli con buona capacità di generalizzazione non soffrono eccessivamente della perturbazione moderata.
- **Multi-Layer Perceptron** mostra il comportamento più anomalo: accuracy bassa (0.567) ma recall altissimo (0.753), il più elevato tra tutte le configurazioni. La rete neurale ha appreso un classificatore estremamente sensibile ma poco specifico, probabilmente a causa di difficoltà nell'ottimizzazione su dati rumorosi.
- **Decision Tree** registra il crollo più marcato (ROC-AUC 0.569), indicando vulnerabilità dei modelli a singolo albero al rumore.

Il fenomeno critico è il crollo del recall medio a 0.408, meno della metà rispetto al dataset senza privacy (0.616) e significativamente inferiore alla baseline (0.525). Gradient Boosting è l'unica eccezione con recall 0.543, mantenendo un bilanciamento accettabile (F1-Score 0.561). Questo pattern indica che il rumore differenziale introduce bias verso predizioni negative, rendendo i modelli più conservativi.

Dataset con privacy forte ($\varepsilon = 1.0$)

Il livello di privacy più stringente causa un degrado critico delle prestazioni. L'accuracy media (0.644) scende di soli 3 punti rispetto alla privacy moderata, ma il ROC-AUC crolla a 0.591, appena sopra il random guessing (0.5).

L'analisi per modello rivela un collasso quasi uniforme:

- **Il recall medio crolla a 0.129**, con 8 modelli su 11 sotto 0.21. Questo significa che i modelli identificano correttamente meno del 20% dei casi positivi reali, rendendo i dataset praticamente inutili per applicazioni di screening o diagnosi.
- **Naive Bayes e AdaBoost** raggiungono precision altissime (0.800 e 0.778) ma con recall rispettivamente 0.049 e 0.086, configurando classificatori estremamente conservativi che predicono positivo solo in casi di altissima certezza.
- **K-Nearest Neighbors** mostra la migliore robustezza con ROC-AUC 0.675 e recall 0.210, suggerendo che metodi non parametrici basati su similarità locale sono meno sensibili al rumore globale.
- **Multi-Layer Perceptron** registra il peggior ROC-AUC (0.466), addirittura sotto il random guessing, indicando che la rete ha appreso correlazioni spurie dal rumore.

Il collasso delle prestazioni non è uniforme: modelli come LR, SVM, NB mantengono accuracy ragionevole (0.65-0.67) ma con recall vicino a zero, mentre modelli come DT, RF, KNN preservano un minimo di recall (0.12-0.21) a costo di accuracy leggermente inferiore. Questa dicotomia riflette diverse strategie di gestione dell'incertezza introdotta dal rumore massiccio.

3.3. Analisi comparativa e discussione

3.3.1. Trade-off privacy-utilità

I risultati quantificano chiaramente il trade-off tra protezione della privacy e utilità pratica. Definendo il *gap di utilità* come la differenza percentuale di ROC-AUC rispetto alla baseline:

$$\text{Utility Gap} = \frac{\text{ROC-AUC}_{baseline} - \text{ROC-AUC}_{synthetic}}{\text{ROC-AUC}_{baseline}} \times 100\% \tag{12}$$

Si ottiene:

- No privacy: 12.1% utility gap (ROC-AUC 0.725 vs 0.825)
- Privacy moderata ($\varepsilon = 5.0$): 11.8% utility gap (ROC-AUC 0.728 vs 0.825)
- Privacy forte ($\varepsilon = 1.0$): 28.4% utility gap (ROC-AUC 0.591 vs 0.825)

La progressione rivela una transizione non lineare: il passaggio da assenza di privacy a privacy moderata comporta un degrado minimo (+0.3 punti percentuali), mentre il salto verso privacy forte causa un crollo sproporzionato (+16.6 punti aggiuntivi). Questo suggerisce l'esistenza di una soglia critica di rumore oltre la quale l'informazione predittiva viene irreparabilmente compromessa.

Analizzando l'F1-Score medio, la transizione appare ancora più drammatica:

- No privacy: 0.572 (degradazione 3% vs baseline 0.591)
- Privacy moderata: 0.474 (degradazione 20% vs baseline)
- Privacy forte: 0.211 (degradazione 64% vs baseline)

Mentre ROC-AUC misura la capacità discriminativa teorica, F1-Score cattura l'utilità pratica bilanciando precision e recall. Il crollo dell'F1-Score per privacy forte (da 0.474 a 0.211, -55%) indica che il classificatore diventa praticamente inutilizzabile per applicazioni reali.

3.3.2. Variabilità tra modelli

L'analisi della deviazione standard delle prestazioni tra algoritmi quantifica la robustezza dei dataset:

La variabilità crescente con l'aumentare della privacy (da 0.013 a 0.076) indica che il rumore differenziale impatta diversamente algoritmi con diverse capacità di generalizzazione. Questo fenomeno rivela insight sulla natura dei dati:

- **Modelli lineari** (LR, SVM) mostrano robustezza superiore al rumore, con ROC-AUC che rimane elevato anche con $\varepsilon = 5.0$ (0.812 e 0.785 rispettivamente). Questo conferma che relazioni lineari tra features vengono preservate efficacemente dal processo generativo CTGAN combinato con privacy differenziale limitata.
- **Ensemble methods** (RF, XGB, LGBM, GB) mantengono prestazioni discrete con privacy assente (ROC-AUC 0.73-0.75) ma degradano significativamente con $\varepsilon = 5.0$ (0.72-0.76) e collassano con $\varepsilon = 1.0$ (0.52-0.59). Questo indica che pattern non-lineari complessi e interazioni di ordine superiore tra variabili vengono progressivamente corrotti dal rumore.
- **Decision Tree** singolo mostra vulnerabilità estrema (ROC-AUC da 0.646 a 0.569 a 0.622), confermando che modelli ad alta varianza senza meccanismi di regolarizzazione sono inadatti per dati rumorosi.
- **Neural networks** (MLP) esibiscono comportamento erratico: prestazioni mediocri su tutti i livelli (ROC-AUC 0.64-0.66), con picco anomalo di recall su privacy moderata (0.753). L'ottimizzazione gradient-based su dataset sintetici di dimensioni limitate (537 campioni) sembra convergere verso minimi locali subottimali.
- **K-Nearest Neighbors** mostra resilienza sorprendente su privacy forte (ROC-AUC 0.675, migliore in quella categoria), suggerendo che metodi basati su similarità locale nel feature space sono meno sensibili a perturbazioni globali distribuite uniformemente.

3.3.3. Implicazioni pratiche per la ricerca medica

I risultati sperimentali forniscono linee guida concrete per ricercatori che devono selezionare dataset sintetici:

Scenario 1: Sviluppo e validazione di algoritmi

Per ricerca metodologica dove l'obiettivo è sviluppare nuovi algoritmi o validare approcci:

- **Raccomandazione**: Dataset senza privacy o privacy moderata
- **Rationale**: Utility gap limitato (11-12%) permette sviluppo iterativo efficiente. Modelli validati su dati sintetici richiedono fine-tuning minimo su dati reali.

- **Modelli consigliati**: Ensemble methods (RF, GB, XGB) per catturare pattern complessi; SVM per robustezza.

Scenario 2: Studi epidemiologici e analisi esplorative
Per ricerca descrittiva dove correlazioni e trend sono più importanti di predizioni precise:

- **Raccomandazione**: Privacy moderata ($\varepsilon = 5.0$)
- **Rationale**: Preserva struttura di correlazione (differenza Frobenius 0.143) con garanzie formali. ROC-AUC medio 0.728 indica che relazioni principali sono preservate.
- **Limitazioni**: Recall ridotto (0.408) richiede cautela nell'interpretazione di prevalenze e incidenze.

Scenario 3: Condivisione pubblica per data challenges
Per rilascio pubblico dove protezione della privacy è prioritaria:

- **Raccomandazione**: Privacy forte ($\varepsilon = 1.0$) solo per task che non richiedono predizione accurata
- **Rationale**: Garanzie formali robuste (vedi Sezione 4) ma utilità predittiva gravemente compromessa (F1 0.211)
- **Use case appropriati**: Benchmarking di algoritmi robusti al rumore, didattica, sviluppo di pipeline di preprocessing

Considerazioni sul bilanciamento Precision-Recall
Un'osservazione critica emerge dall'analisi del trade-off precision-recall:

- Dataset **senza privacy**: Recall alto (0.616), precision moderata (0.542) → appropriato per screening dove falsi negativi sono costosi
- Dataset **privacy moderata**: Precision alta (0.586), recall basso (0.408) → appropriato per diagnosi confermativa dove falsi positivi sono costosi
- Dataset **privacy forte**: Entrambi compromessi → inadatto per applicazioni cliniche

Questa dicotomia suggerisce che la scelta del livello di privacy dovrebbe considerare non solo l'utility gap aggregato ma anche il profilo rischio-costo specifico dell'applicazione target.

4. 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.

5. Conclusioni

Il presente lavoro ha investigato sistematicamente il compromesso tra privacy e utilità nell'impiego di dati sanitari sintetici per la ricerca medica, rispondendo alle domande di ricerca iniziali attraverso un'analisi quantitativa multi-fase.

5.1. Risposte alle domande di ricerca

RQ1: Quali modelli mantengono le prestazioni più elevate quando addestrati su dati sintetici rispetto alla baseline con dati reali?

L'analisi sperimentale condotta su 11 algoritmi di machine learning ha evidenziato notevoli differenze nella capacità di generalizzare dai dati sintetici. In assenza di privacy differenziale ($\varepsilon \to \infty$), i modelli *Support Vector Machine* (accuracy: 0.710, ROC-AUC: 0.758) e *Naive Bayes* (F1-Score: 0.600) hanno raggiunto le prestazioni più elevate, mostrando una resilienza intrinseca alle caratteristiche specifiche dei dati sintetici generati da CTGAN. Per il livello di privacy moderata ($\varepsilon = 5.0$), la *Logistic Regression* ha dimostrato eccezionale robustezza, mantenendo un ROC-AUC di 0.812, degradando appena del 2% rispetto alla baseline (0.8345). Questa performance suggerisce che le relazioni lineari tra variabili predittive sono preservate efficacemente anche in presenza di perturbazione differenzialmente privata moderata. Al contrario, algoritmi complessi e ad alta varianza come *Decision Tree* e *Multi-Layer Perceptron* hanno mostrato vulnerabilità significativa al rumore introdotto, specialmente per $\varepsilon = 1.0$. Il modello *K-Nearest Neighbors* si è distinto per resilienza in condizioni di privacy forte (ROC-AUC: 0.675), indicando che metodi basati su similarità locale nel feature space sono meno sensibili a perturbazioni globali.

La variabilità inter-modello, quantificata dalla deviazione standard del ROC-AUC (Tabella 6), è aumentata progressivamente con l'introduzione di privacy: da 0.013 (baseline reale) a 0.039 (nessuna privacy) fino a 0.076 (privacy moderata). Questo fenomeno indica che il rumore differenziale amplifica la sensibilità degli algoritmi alle specifiche assunzioni di modellazione, rendendo la scelta del modello critica in contesti ad alta protezione.

RQ2: È possibile conciliare privacy e utilità in un dataset sintetico, o devono essere accettati compromessi significativi?

I risultati dimostrano che un compromesso tra privacy e utilità è raggiungibile, ma entro limiti ben definiti. Per $\varepsilon = 5.0$ (privacy moderata), il gap di utilità medio, definito come la riduzione percentuale del ROC-AUC rispetto alla baseline, è risultato dell'11.8%, una penalità accettabile per molte applicazioni di ricerca. In particolare, la struttura delle correlazioni lineari (preservata al 86% secondo la distanza di Frobenius) e la capacità discriminativa dei modelli lineari (Logistic Regression ROC-AUC: 0.812) rimangono adeguate per analisi epidemiologiche e sviluppo di modelli predittivi.

Tuttavia, il passaggio a $\varepsilon = 1.0$ (privacy forte) comporta un compromesso sostanziale: il gap di utilità aumenta al 28.4%, con un crollo del F1-Score medio da 0.474 a 0.211 (-55%). Questo livello di protezione, pur offrendo garanzie formali robuste secondo la definizione di ε-privacy differenziale, compromette irrimediabilmente l'utilità predittiva per la diagnosi del diabete. Il recall medio crolla a 0.129, rendendo i classificatori praticamente inutili per applicazioni di screening.

Il trade-off ottimale emerge quindi per $\varepsilon = 5.0$, dove le garanzie di privacy sono formalmente dimostrabili (la probabilità di inferire la presenza di un individuo nel dataset è limitata da un fattore $e^5 \approx 148$) mentre l'utilità predittiva si mantiene entro il 12% dalla baseline. Questo punto di equilibrio permette di conciliare esigenze contrapposte: da un lato, la protezione dei pazienti da attacchi di re-identification e membership inference; dall'altro, la preservazione di relazioni statistiche clinicamente rilevanti per la ricerca.

5.2. Limiti dello studio e direzioni future

La presente analisi presenta alcune limitazioni che delineano percorsi per ricerche future:

1. **Generalizzazione ad altri dataset**: L'utilizzo esclusivo dell'UCI Diabetes Dataset, sebbene rappresentativo, limita la generalizzabilità dei risultati. Studi futuri dovrebbero validare queste conclusioni su dataset sanitari più complessi, con dimensionalità maggiore, tipologie di variabili eterogenee (testi medici, immagini) e distribuzioni di patologie diverse.
2. **Parametrizzazione del rumore**: L'applicazione della privacy differenziale attraverso il meccanismo di Laplace con dominio limitato rappresenta una scelta specifica. Approcci alternativi, come il meccanismo Gaussiano, l'exponential mechanism per variabili categoriche, o tecniche di privacy locale, potrebbero offrire trade-off diversi meritevoli di esplorazione.

3. **Valutazione multi-obiettivo**: Lo studio ha adottato metriche di utilità incentrate sulle prestazioni predittive. Una valutazione più completa dovrebbe incorporare metriche specifiche per dominio medico, come l'aderenza a linee guida cliniche, la riproducibilità di associazioni rischio-fattore note in letteratura, e l'impatto su decisioni terapeutiche simulate.
4. **Attacchi avanzati alla privacy**: La valutazione della resistenza agli attacchi si è concentrata sul membership inference. Attacchi più sofisticati, come i modelli di linkage attack che combinano dati sintetici con dataset ausiliari pubblici, o attacchi basati su modelli generativi avversari, necessitano di indagini specifiche.
5. **Aspetti regolatori ed etici**: La ricerca futura dovrebbe integrare l'analisi tecnica con considerazioni sul quadro normativo (GDPR, HIPAA) e su framework etici per determinare soglie di ε accettabili in diversi contesti applicativi.

5.3. Implicazioni pratiche per la ricerca biomedica

I risultati di questo studio forniscono indicazioni operative per ricercatori e istituzioni sanitarie che intendono utilizzare dati sintetici:

- Per lo **sviluppo e benchmarking di algoritmi**, dove l'obiettivo è metodologico, i dataset sintetici senza privacy o con privacy moderata ($\varepsilon = 5.0$) offrono un ambiente realistico ed etico, con un degradamento controllato delle prestazioni (10-12%).
- Per gli **studi osservazionali e epidemiologici**, finalizzati all'identificazione di correlazioni e fattori di rischio, il livello $\varepsilon = 5.0$ preserva sufficientemente la struttura delle relazioni multivariate, garantendo al contempo protezione formale.
- Per la **condivisione pubblica di dataset a fini di riproducibilità e open science**, il livello $\varepsilon = 1.0$ fornisce garanzie di privacy solide, sebbene a scapito di una significativa riduzione dell'utilità analitica. Il suo impiego è consigliabile solo per task che non richiedono alta precisione predittiva.
- La scelta dell'**algoritmo di machine learning** deve essere allineata al livello di privacy: modelli lineari (Logistic Regression) per privacy moderata, ensemble methods per privacy assente, e metodi non parametrici (K-NN) in condizioni di rumore elevato.

5.4. 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 per molte applicazioni. Il compromesso ottimale individuato ($\varepsilon = 5.0$) costituisce un punto di riferimento per ricercatori e policymaker, 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 richiederà ulteriori ricerche interdisciplinari, ma i risultati presentati confermano il potenziale trasformativo di questo approccio per il futuro della ricerca biomedica basata sui dati.

Tabella 5. Prestazioni complete di tutti i modelli predittivi su dati sintetici

Modello	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline - Addestramento su Dati Reali					
Logistic Regression	0.7446	0.6719	0.5309	0.5931	0.8345
Random Forest	0.7403	0.6667	0.5185	0.5833	0.8159
Dataset Sintetico - Senza Privacy (CTGAN)					
Logistic Regression	0.6883	0.5495	0.6173	0.5814	0.7612
Random Forest	0.6926	0.5510	0.6667	0.6034	0.7461
Gradient Boosting	0.6840	0.5408	0.6543	0.5922	0.7360
AdaBoost	0.6580	0.5100	0.6296	0.5635	0.7453
Decision Tree	0.6450	0.4950	0.6173	0.5495	0.6464
SVM	0.7100	0.6000	0.5185	0.5563	0.7581
K-Nearest Neighbors	0.6926	0.5568	0.6049	0.5799	0.7113
Naive Bayes	0.7056	0.5730	0.6296	0.6000	0.7505
Multi-Layer Perceptron	0.6537	0.5057	0.5432	0.5238	0.6430
XGBoost	0.6797	0.5354	0.6543	0.5889	0.7415
LightGBM	0.6840	0.5417	0.6420	0.5876	0.7319
Media	*0.6812*	*0.5417*	*0.6162*	*0.5715*	*0.7250*
Dataset Sintetico - Privacy Moderata ($\varepsilon = 5.0$)					
Logistic Regression	0.7100	0.6750	0.3333	0.4463	0.8116
Random Forest	0.6753	0.5789	0.2716	0.3697	0.7562
Gradient Boosting	0.7013	0.5789	0.5432	0.5605	0.7374
AdaBoost	0.7013	0.6500	0.3210	0.4298	0.7931
Decision Tree	0.6190	0.4545	0.4321	0.4430	0.5691
SVM	0.7143	0.6744	0.3580	0.4677	0.7854
K-Nearest Neighbors	0.6407	0.4881	0.5062	0.4970	0.6460
Naive Bayes	0.7013	0.6429	0.3333	0.4390	0.7861
Multi-Layer Perceptron	0.5671	0.4326	0.7531	0.5495	0.6439
XGBoost	0.7013	0.6154	0.3951	0.4812	0.7604
LightGBM	0.6753	0.5600	0.3457	0.4275	0.7215
Media	*0.6734*	*0.5864*	*0.4084*	*0.4737*	*0.7282*
Dataset Sintetico - Privacy Forte ($\varepsilon = 1.0$)					
Logistic Regression	0.6537	0.6667	0.0247	0.0476	0.6113
Random Forest	0.6537	0.5263	0.1235	0.2000	0.5521
Gradient Boosting	0.6407	0.4706	0.1975	0.2783	0.5537
AdaBoost	0.6710	0.7778	0.0864	0.1556	0.6724
Decision Tree	0.6104	0.3953	0.2099	0.2742	0.6221
SVM	0.6450	0.4545	0.0617	0.1087	0.5584
K-Nearest Neighbors	0.6667	0.5667	0.2099	0.3063	0.6748
Naive Bayes	0.6623	0.8000	0.0494	0.0930	0.6677
Multi-Layer Perceptron	0.5931	0.3415	0.1728	0.2295	0.4663
XGBoost	0.6407	0.4615	0.1481	0.2243	0.5924
LightGBM	0.6407	0.4583	0.1358	0.2095	0.5264
Media	*0.6435*	*0.5563*	*0.1291*	*0.2115*	*0.5907*

Tabella 6. Variabilità delle prestazioni (deviazione standard del ROC-AUC tra modelli)

Dataset	Std(ROC-AUC)
Baseline (Reali)	0.013
No Privacy	0.039
Privacy Moderata ($\varepsilon = 5.0$)	0.076
Privacy Forte ($\varepsilon = 1.0$)	0.062