PANORAMICA DELLA SOLUZIONE

La piattaforma di dati sintetici di Aindo consente alle organizzazioni di sfruttare appieno il potenziale dei loro dati. Grazie a modelli generativi sviluppati specificamente per i dati strutturati, permette di:

- Sbloccare l'accesso e il riutilizzo di dataset soggetti a vincoli di privacy attraverso la sintesi e l'anonimizzazione/pseudonimizzazione;
- Migliorare la qualità dei dati tramite ribilanciamento, augmentation dei dati e imputazione dei valori mancanti;
- Creare modelli predittivi direttamente sui dati relazionali, senza necessità di utilizzare *feature engineering*;
- La generazione di scenari di *what-if*;
- Identificare e redarre le informazioni personali identificabili (PII) all'interno di dati testuali e documenti.

Le nostre soluzioni affrontano sfide tecniche chiave, come l'accessibilità limitata ai dati, la scarsità dei dati, la variabilità della qualità dei dati e la creazione di modelli di apprendimento automatico su dati strutturati, garantendo al contempo la conformità con rigorosi requisiti normativi come il GDPR e l'AI Act. Grazie a soluzioni avanzate di dati sintetici, le organizzazioni possono sfruttare con fiducia i propri dati per l'innovazione e il processo decisionale, trasformando gli ostacoli in opportunità.

Funzionalità

Generazione di dati sintetici

La nostra soluzione offre un'ampia gamma di generatori, in grado di soddisfare la maggior parte delle esigenze di mercato.

Generazione di dati sintetici relazionali – Una robusta libreria di generatori di dati per creare dataset sintetici di alta qualità, composti da una o più tabelle collegate tra loro tramite chiavi esterne. I generatori addestrano direttamente sui dati del cliente, senza uscire dall'ambiente di installazione. I nostri generatori preservano le statistiche delle tabelle originali, comprese le correlazioni tra tabelle diverse o tra righe all'interno della stessa tabella, supportando di *default* le serie temporali. Allo stesso tempo, i dati sintetici generati non possono essere collegati a nessuna persona reale, risultando quindi anonimi secondo il GDPR. Inoltre, preservano anche l'integrità delle chiavi e supportano diversi tipi di dati, tra cui:

- Categorici
- Booleano
- Data
- Ora
- Data e ora
- Interi
- Numero in virgola mobile
- Coordinate
- Testo libero
- Codice fiscale italiano

In particolare, anche la colonna di testo libero viene generata mantenendo le correlazioni con le altre colonne della stessa tabella.

I nostri generatori offrono un'eccellente protezione della privacy: supportano l'addestramento *differentially private* per garantire matematicamente il corretto livello di protezione delle informazioni. Inoltre, per proteggersi da fonti secondarie di perdita di privacy, l'utente ha la possibilità di generare dati sintetici privi di valori rari presenti nei dati originali.

Infine, la nostra soluzione include un modulo che consente all'utente di definire *business rules* o vincoli presenti nei dati, che devono essere rispettati sistematicamente dai dati sintetici. Attualmente gestiamo relazioni matematiche generiche tra colonne, espresse come equazioni di uguaglianza o disuguaglianza (espressioni come: "Colonna A / Colonna C + 3 > Colonna D") e mappature tra colonne.

Generazione parzialmente sintetica dei dati – La nostra soluzione permette di sintetizzare solo un sottoinsieme di colonne di un dataset, preservando le altre colonne dai dati originali. Questa opzione consente un controllo dettagliato sulle informazioni originali rivelate e sul livello di protezione della privacy.

Generatori pre-addestrati – La nostra soluzione consente di generare dati tabulari a partire da una descrizione in linguaggio naturale, utilizzando un generatore pre-addestrato su un vasto set di dati strutturati. Ciò consente la generazione di dati anche in assenza di esempi, per casi d'uso come il test del software in cui è necessario avere dei dati che non si avrebbero normalmente a disposizione. A differenza di altre soluzioni basate su LLM disponibili sul mercato, la nostra soluzione garantisce il formato delle righe e delle colonne, evitando errori di formattazione nei processamenti successivi.
Se ritenuto utile, la soluzione pre-addestrata può essere ulteriormente *fine-tuned*, un'opzione particolarmente utile quando l'utente dispone di un numero limitato di esempi. Questo fine-tuning può essere eseguito su più dataset di formati diversi, permettendo di apprendere da più sorgenti di dati contemporaneamente.
Il generatore pre-addestrato è anche in grado di ampliare un dataset aggiungendo colonne non originariamente presenti in esso.

Valutazione dei dati sintetici – La nostra soluzione offre un set completo di metriche per garantire il rispetto degli standard essenziali di privacy e utilità. Questo framework di valutazione è fondamentale per mantenere la qualità dei dati e la conformità normativa nei diversi casi d'uso. Le valutazioni sono disponibili in formato PDF, generato automaticamente ad ogni creazione di dato.

Strumenti di anonimizzazione

Oltre alle opzioni di sintesi, la nostra piattaforma offre strumenti aggiuntivi per una protezione avanzata della privacy.

Anonimizzazione o pseudonimizzazione classica – Un set di strumenti per l'anonimizzazione o la pseudonimizzazione dei dati tabulari, che combina diverse tecniche: generalizzazione, randomizzazione, permutazione, hashing, mock data. Questa funzione può essere combinata con la sintesi parziale per una grande flessibilità di configurazione.

Redazione delle PII nei testi e nei documenti * – Offriamo un avanzato strumento di elaborazione testuale per identificare e rimuovere le informazioni personali identificabili (PII) nei

testi e nei documenti. L'utente può scegliere di sostituire le informazioni personali con dati fittizi, preservando la struttura complessiva del documento.

** in fase di sviluppo*

Modellistica predittiva

Oltre alla generazione di dati, i nostri generatori possono essere utilizzati per eseguire compiti predittivi su dati strutturati.

Predizione su dati relazionali – La nostra soluzione consente di utilizzare i generatori di dati sintetici in modalità predittiva. In questa modalità, i generatori possono effettuare previsioni sul valore di determinate colonne di un dataset relazionale in base al valore di tutte le altre colonne. Questo processo avviene direttamente sui dati relazionali, senza necessità di *feature engineering*, mantenendo al contempo un livello di prestazioni molto elevato. Questa funzionalità può essere utilizzata anche come strumento di imputazione dei dati, per inferire informazioni mancanti presenti in un dataset.

Estrapolazione su dati relazionali – La nostra soluzione include un modulo in grado di effettuare estrapolazioni sui dati relazionali. In questa modalità, il sistema permette di prevedere tendenze future basandosi su dati storici. Questo modulo utilizza un generatore specifico addestrato su dati passati per fornire previsioni accurate e contestualmente rilevanti. Poiché la previsione viene eseguita su dati strutturati complessi, è possibile includere informazioni in formato variegato nelle estrapolazioni. Il modulo è anche in grado di generare scenari di *what-if*, consentendo di valutare le differenze nei risultati nel caso in cui si verifichi un determinato evento.

Installazione

La piattaforma di dati sintetici di Aindo è eseguita da container Docker orchestrati in un cluster Kubernetes, garantendo scalabilità e affidabilità per una varietà di casi d'uso. Può quindi essere utilizzata come SaaS, ospitata su cloud privati o installata *on-premise*.

① **SaaS**	② **Self hosted**	③ **On–Prem**
aindo		

Interfaccia utente, REST API, SDK – La piattaforma è progettata per supportare sia progetti di sviluppo che di produzione, offrendo diverse opzioni di interazione per soddisfare le esigenze dei clienti. La piattaforma dispone di un'interfaccia utente (UI) intuitiva che consente anche agli utenti meno esperti di sfruttare la maggior parte delle funzionalità disponibili. Le stesse funzionalità presenti nell'UI sono accessibili anche tramite una REST API per interazioni macchina-macchina. Inoltre, offriamo un SDK adatto agli sviluppatori. Le funzionalità disponibili nelle diverse opzioni di implementazione sono elencate nella seguente tabella.

ITEM	SDK	UI / API

Generazione di dati sintetici relazionali	✅	✅
Allenamento differentially private	✅	✅
Protezione valori rari	✅	✅
Gestione vincoli	✅	
Generazione dati parzialmente sintetici	✅	✅
Generatori pre-addestrati	✅	
Valutazione dei dati sintetici	✅	✅
Data Curation	✅	✅
Anonimizzazione o pseudonimizzazione classica	✅	✅
Redazione delle PII nei testi e nei documenti	in fase di sviluppo	
Predizione su dati relazionali	✅	
Estrapolazione su dati relazionali	✅	

Scalabilità – La soluzione può essere eseguita su CPU, ma supporta l'addestramento multi-GPU per i generatori, consentendo la sintesi di dati altamente complessi e voluminosi.

Connettori – La piattaforma offre diversi connettori per l'input/output dei dati. Attualmente, abbiamo implementato i seguenti connettori per database:

- PostgreSQL
- MySQL
- MariaDB
- Google Big Query
- Microsoft SQL Server
- Oracle Database

Supportiamo inoltre le connessioni ai seguenti sistemi di storage a oggetti:

- Amazon S3
- Google Cloud Storage

Crittografia – La nostra piattaforma cripta tutti i dati sia a riposo che in transito, garantendo alti livelli di sicurezza e impedendo l'accesso non autorizzato ai dati.

Certificazioni

Siamo la prima azienda di dati sintetici a ottenere la certificazione EuroPrivacy in Europa. Questa certificazione attesta la conformità al GDPR dei nostri protocolli di generazione di dati sintetici nel settore sanitario. Aindo possiede inoltre le certificazioni ISO 9001 e ISO 27001.

Supporto

Forniamo servizi di supporto ai clienti con un contratto attivo (manutenzione o abbonamento annuale). Il supporto è disponibile in italiano e inglese.

Casi d’uso

Sviluppo di modelli AI/ML e analisi – Generazione di una versione anonima di un dataset di input che preserva i pattern dei dati, rimuovendo gli identificatori personali e garantendo la conformità alla privacy. Miglioramento della qualità del dataset sintetico per massimizzarne il valore per analisi successive, utilizzando funzionalità come il riequilibrio per garantire rappresentatività o l'imputazione intelligente per colmare lacune nei dati.

> *Esempio*: Anonimizzazione di dati reali per la ricerca medica.

Condivisione dei dati per la ricerca e lo sviluppo collaborativo – Permette alle organizzazioni di condividere versioni sintetiche realistiche di dataset sensibili con collaboratori esterni o stakeholder senza esporre informazioni personali. Supporta la collaborazione aperta in settori come la ricerca medica o gli studi accademici, dove le preoccupazioni sulla privacy spesso limitano la condivisione dei dati.

> *Esempio*: Condivisione sicura di dati con ricercatori esterni, fornitori e partner.

Aumento e riequilibrio dei dati – Riequilibrio delle categorie presenti in un dataset per ridurre i bias e aumentare l'equità dei dati. Aumento dei casi rari o marginali in un dataset per migliorare l'addestramento dei modelli di machine learning.

> *Esempio*: Riequilibrare il numero di uomini e donne in un dataset per migliorare l'equità di un algoritmo di machine learning.

Modellazione predittiva – Previsione delle probabili tendenze future di un individuo basandosi su dati storici. Analizzando pattern appresi da dati simili, la nostra piattaforma fornisce previsioni personalizzate, utili per applicazioni come il monitoraggio della salute dei pazienti, la previsione del comportamento dei clienti o l'analisi delle tendenze finanziarie.

> *Esempio*: Stima del rischio di sviluppare malattie per ottimizzare le campagne di screening o valutazione del rischio di eventi avversi per una pianificazione ottimale delle risorse sanitarie.

Redazione delle PII in testi e documenti* – Rilevamento e protezione delle informazioni personali identificabili (PII) all'interno di dati testuali e documenti. Gli utenti possono scegliere di mascherare le PII per motivi di sicurezza o sostituirle con equivalenti sintetici realistici, consentendo una condivisione sicura e un'elaborazione senza compromettere la privacy.

> *Esempio*: Rimozione di tutte le informazioni PII dalle cartelle cliniche elettroniche.

** in fase di sviluppo.*