TECNOLOGÍA DE MEJORA DE LA PRIVACIDAD (PET): PROPUESTA DE GUÍA SOBRE LA GENERACIÓN DE DATOS SINTÉTICOS

Traducido de la guía publicada por la Comisión de Protección de Datos Personales de Singapur (PDPC Singapore) publicada el 15 de julio de 2024

Número de versión 1.0

PRÓLOGO AEPD

El presente documento es la traducción de la guía básica elaborada por la Comisión de Protección de Datos Personales de Singapur (PDPC Singapore), que se puede descargar en su versión original en ingles en el siguiente enlace:

Proposed Guide to Synthetic Data Generation

La División de Innovación Tecnológica de la AEPD estimó el gran valor didáctico de esta guía por lo que, tras pedir las autorizaciones correspondientes al PDPC Singapore, se decidió a la traducción y publicación en su espacio de divulgación dentro de la página web de la AEPD.

Sin embargo, se ha de tener en cuenta que su contenido se ha desarrollado en un contexto normativo que no es el del RGPD y que hay determinadas afirmaciones o términos que pude diferir de o establecido en la normativa europea y por el Comité Europeo de Protección de Datos.

Este documento se considera de especial interés para responsables, encargados de tratamientos y delegados de protección de datos (DPD).

ÍNDICE DE CONTENIDO

I. Introducción a la tecnología de mejora de la privacidad (PET)

Las tecnologías de mejora de la privacidad (PET, por sus siglas en inglés *Privacy Enhancing Technology*) son un conjunto de herramientas y técnicas que permiten el tratamiento, el análisis y la extracción de información de los datos sin revelar los datos personales o comercialmente confidenciales subyacentes. Al incorporar PET, las empresas pueden mantener una ventaja competitiva en el mercado aprovechando sus activos de datos existentes para la innovación, al tiempo que cumplen con las regulaciones de protección de datos, reducen el riesgo de brechas de datos y demuestran un compromiso con la protección de datos. Las PET no son solo una medida defensiva; son un paso proactivo para fomentar una cultura de protección de datos y asegurar la reputación de una empresa en la era digital.

En general, las PET se pueden clasificar en tres categorías clave[1]: ofuscación de datos, tratamiento de datos cifrados y análisis federado. Las PET también pueden combinarse para abordar las diversas necesidades de las organizaciones. La siguiente **Tabla 1** muestra los tipos actuales de PET en el mercado y sus principales aplicaciones.

Tabla 1. Tipos de PET y sus aplicaciones

Categorías de PETs	PETs	Ejemplos de aplicaciones (lista no exhaustiva)
Ofuscación de datos	Técnicas de anonimización/seudonimización	• Almacenamiento seguro • Uso compartido y retención de datos • Pruebas de software
	Generación de datos sintéticos	• Aprendizaje automático de IA que preserva la privacidad • Compartición y análisis de datos • Pruebas de software
	Privacidad diferencial	• Ampliar las oportunidades de investigación • Compartición de datos
	Pruebas de conocimiento cero	• Verificación de información sin requerir divulgación (por ejemplo, verificación de edad)
Tratamiento de datos encriptados	Cifrado homomórfico	• Proteger los datos almacenados en la nube • Computación con datos privados que no se divulgan
	Cómputo multiparte (incluida la intersección de conjuntos privados)	• Computación con datos privados que no se divulgan
	Entornos de ejecución de confianza	• Computación con modelos que deben permanecer privados • Computación con datos privados que no se divulgan

[1] Adaptado del documento de la OCDE, "Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches", OECD Digital Economy Papers (OCDE, 2023)

Analítica federada	Aprendizaje federado	• Aprendizaje automático de IA que preserva la privacidad
	Análisis distribuido	

II. Datos sintéticos

Esta guía se centra en el uso de datos sintéticos[2] para generar datos estructurados. Si bien los datos sintéticos son generalmente datos ficticios que pueden no considerarse datos personales por sí mismos, no están intrínsecamente exentos de riesgos debido a los posibles riesgos de reidentificación[3]. Como tal, esta guía propone buenas prácticas que las organizaciones pueden adoptar para generar datos sintéticos con el fin de minimizar dichos riesgos para un conjunto de arquetipos de casos de uso comunes. La guía también incluye un conjunto de buenas prácticas y evaluaciones/consideraciones de riesgos para generar datos sintéticos, así como controles de gobernanza, procesos contractuales y medidas técnicas para mitigar los riesgos residuales.

El público objetivo de esta guía son CIO (el mánager de información), CTO (el mánager de tecnología), CDO (el mánager de datos), científicos de datos, profesionales de la protección de datos y responsables de la toma de decisiones técnicas que pueden estar directa o indirectamente involucrados en la generación y el uso de datos sintéticos.

Los datos sintéticos son una tecnología que se está investigando y desarrollando activamente en el momento de la publicación. Por lo tanto, esta guía no pretende proporcionar una revisión exhaustiva o en profundidad de la tecnología o sus métodos de evaluación. La guía pretende ser un documento vivo y se actualizará para garantizar que sus recomendaciones sigan siendo pertinentes

¿Qué son los datos sintéticos?

Los datos sintéticos se conocen comúnmente como datos artificiales que se han generado utilizando un modelo matemático especialmente diseñado (incluidos los modelos de inteligencia artificial (IA)/aprendizaje automático (ML, por sus siglas en inglés)) o algoritmo. Se puede derivar entrenando un modelo (o algoritmo) con un conjunto de datos de origen para imitar las características y la estructura de los datos de origen. Los datos sintéticos de buena calidad pueden conservar en gran medida las propiedades estadísticas y los patrones de los datos de origen. Como resultado, la realización del análisis en datos sintéticos puede producir resultados similares a los obtenidos con los datos de origen.

Características de los datos sintéticos

En la **Figura 1** se muestra un ejemplo de cómo pueden ser los datos sintéticos en comparación con los datos de origen. El conjunto de los datos sintéticos generados

[2] Hay dos tipos de datos sintéticos: datos totalmente sintéticos y datos parcialmente sintéticos. En esta guía se analiza el uso de datos totalmente sintéticos.

[3] En esta guía, generalmente nos referimos a los riesgos de privacidad como riesgos de reidentificación.

generalmente tendrá datos diferentes de los datos de origen, como se ve en las tablas de datos. Sin embargo, los datos sintéticos tendrán propiedades estadísticas cercanas a las de los datos de origen, es decir, capturarán la distribución y la estructura de los datos de origen como se ve en las líneas de tendencia de la **Figura 1.**

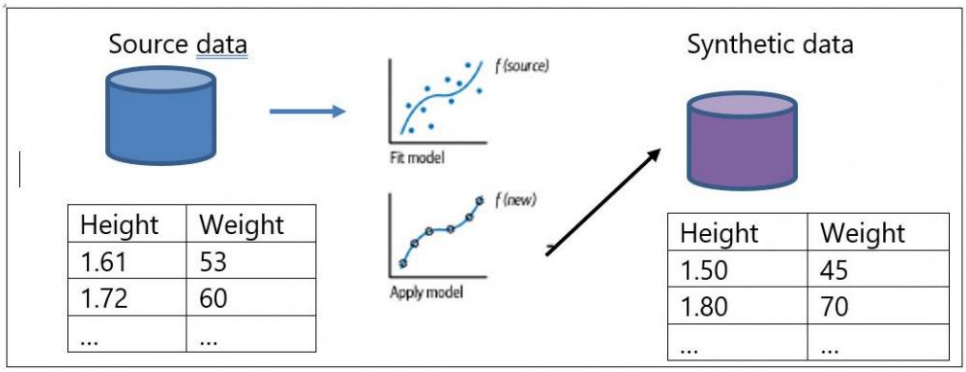


Figura 1: Datos de origen frente a datos sintéticos.[4]

Como tal, es posible que los datos sintéticos no siempre estén inherentemente libres de riesgos, ya que la información sobre un individuo en el conjunto de datos de origen, o los datos confidenciales, aún pueden filtrarse debido a la semejanza de los datos sintéticos con los datos de origen. También habrá que encontrar el equilibrio[5] entre la utilidad de los datos y los riesgos de protección de datos en la generación de datos sintéticos. Sin embargo, estos riesgos pueden minimizarse teniendo en cuenta la protección de datos durante el proceso de generación de datos sintéticos.

¿En qué circunstancias son útiles los datos sintéticos?

Los datos sintéticos se pueden utilizar en una variedad de casos de uso, que van desde la generación de conjuntos de datos de entrenamiento para modelos de IA hasta el análisis de datos y la colaboración. El uso de datos sintéticos no solo puede acelerar la investigación, la innovación, la colaboración y la toma de decisiones, sino que también puede mitigar las preocupaciones sobre incidentes de ciberseguridad y brechas de datos, lo que permite un mejor cumplimiento de las regulaciones de protección de datos/privacidad. En la **Tabla 2** se analizan algunos arquetipos de casos de uso comunes, sus principales beneficios y las buenas prácticas en las que las organizaciones pueden centrarse a la hora de generar datos sintéticos.

Tipos de casos de uso	Principales ventajas	Buenas prácticas para generar datos sintéticos
Arquetipo de caso de uso 1: Generación de conjuntos de datos de entrenamiento para modelos		

[4] Diagrama tomado con modificaciones de Khaled El Emam, Lucy Mosquera y Richard Hoptroff, Practical Synthetic Data Generation (O'Reilly Media, Inc, 2020).

[5] El equilibrio entre los riesgos de la utilidad de los datos y la protección de datos se analiza con más detalle en el Anexo A: Paso 1 y Paso 3 de esta guía.

Incrementar la cantidad de muestras[6] disponibles para modelos de IA/ML	• Los datos sintéticos abordan el desafío de que el usuario tenga que obtener grandes volúmenes de muestras etiquetadas necesarios para entrenar y probar modelos de IA/ML debido a los costos, las regulaciones legales y los derechos de propiedad. • Aumentar los conjuntos de datos de entrenamiento con datos etiquetados generados sintéticamente puede ser ***más rentable***, especialmente cuando los conjuntos de datos de origen son escasos.	• Agregar ruido* o reducir la granularidad de los datos de las muestras sintéticas. • Por lo general, estas nuevas muestras de datos ficticios no se considerarán datos personales. * Si las propiedades/ características estadísticas de los datos sintéticos son representativas de la población en cuestión y no están significativamente sesgadas hacia un individuo/grupo específico de individuos utilizados como fuente de datos de entrenamiento, es posible que no sea necesario agregar ruido, ya que los riesgos de reidentificación son generalmente bajos.
Incremento de la diversidad de datos para modelos de IA/ML	• Los datos sintéticos se pueden utilizar para simular eventos extraordinarios o aumentar grupos infrarrepresentados en el entrenamiento de modelos de IA. • Diversos conjuntos de datos pueden ser útiles para ***mejorar el rendimiento*** de los modelos de IA/ML	
Arquetipo de caso de uso 2: Análisis de datos y colaboración		
Compartición y análisis de datos	• Las tendencias o patrones subyacentes y los sesgos de los datos son útiles para el análisis de datos, independientemente de si la fuente de datos es real o sintética. • Los datos sintéticos pueden permitir el intercambio de datos para el análisis, especialmente en industrias y sectores, por ejemplo, la atención médica, ***donde los datos de origen pueden ser sensibles***.	• Equilibrar las compensaciones entre la utilidad de los datos y la protección de datos mediante la incorporación de medidas de protección de datos a lo largo del proceso de generación de datos sintéticos, por ejemplo: <u>Preparación de datos</u> • Eliminar valores atípicos de los datos de origen • Seudonimizar los datos de origen mediante la minimización de datos
Vista previa de datos para colaboración	• Los datos sintéticos se pueden utilizar en la exploración, el análisis y la colaboración de datos para proporcionar a las partes interesadas una vista previa representativa de los datos	

[6] N.T.: Se utiliza la palabra "muestra" para los distintos ejemplos que son necesarios para el entrenamiento, cada "muestra" contendrá distintos datos. Por ejemplo, sintéticamente se puede crear datos de un nuevo individuo ficticio, que será una nueva muestra, y que contendrá datos sintéticos como edad, peso, altura, etc.

	de origen sin exponer información confidencial. • Esto permite a las partes interesadas explorar y comprender la estructura, las relaciones y los posibles conocimientos dentro de los datos para ***obtener garantías de la calidad de los datos*** antes de finalizar cualquier acuerdo o colaboración.	y generalizar los datos granulares <u>Generación de datos sintéticos</u> • Añadir ruido antes o después de la generación de datos sintéticos <u>Generación de datos postsintéticos</u> • Incorporar medidas técnicas, contractuales y de gobernanza para mitigar los riesgos residuales de reidentificación
Arquetipo de caso de uso 3: Pruebas de software		
Desarrollo de sistemas/ pruebas de software	• Las organizaciones pueden utilizar datos sintéticos en lugar de datos de producción para facilitar el desarrollo de software. • El uso de datos sintéticos puede ayudar a las organizaciones a ***evitar brechas de datos*** en caso de que el entorno de desarrollo se vea comprometido.	• Generando datos sintéticos que sigan la semántica, por ejemplo, el formato, los valores mínimos y máximos y las categorías de los datos de origen en lugar de únicamente las características y propiedades estadísticas.

Consulte el **Anexo A** para conocer las consideraciones propuestas y las buenas prácticas para generar datos sintéticos.

Casos de estudio

(A) Entrenamiento de un modelo de IA para la detección de fraudes en el sector financiero[7]

Problema: Dado que el número de transacciones fraudulentas en los datos de origen es pequeño en comparación con las transacciones normales no fraudulentas, los datos de origen no entrenan adecuadamente los modelos para la detección de fraudes.

Solución: J.P. Morgan utilizó con éxito datos sintéticos para el entrenamiento de modelos de detección de fraude. Se proporcionaron modelos de IA con muestras de transacciones normales y fraudulentas para comprender los signos reveladores de transacciones sospechosas.

Beneficio: Los datos sintéticos demostraron ser más efectivos en términos de modelos de entrenamiento para detectar comportamientos anómalos. Esto se debe a que los datos sintéticos utilizados fueron diseñados para contener un mayor porcentaje de transacciones fraudulentas.

(B) Entrenamiento del modelo de IA para investigación acerca del sesgo de la IA[8]

Problema: Los modelos de clasificación y regresión *multi-label* (multi etiquetado) se utilizan con frecuencia en Mastercard para diversas aplicaciones, incluida la prevención del fraude, la lucha contra el lavado de dinero y los casos de uso de marketing para la optimización de la cartera. Estos modelos, aunque potentes, requieren una atención cuidadosa a los *proxies* de los atributos demográficos dentro de sus datos de entrenamiento, que podrían aprender sesgos no deseados. Garantizar la precisión y la equidad de estos modelos es complejo debido a su configuración de multi etiquetado, la confidencialidad de los atributos demográficos y los desafíos para acceder al conjunto de datos de entrenamiento para el desarrollo de modelos.

Solución: Mastercard se asoció con investigadores para desarrollar nuevos métodos de prueba de sesgo de IA adaptados a configuraciones de multi etiquetado. Para proteger la privacidad de los datos compartidos externamente, se crearon datos sintéticos para respaldar el entrenamiento de modelos y la investigación metodológica en modelos justos de multi etiquetado.

Beneficio: Se midió que los datos sintéticos eran lo suficientemente privados como para ser compartidos con investigadores externos, al tiempo que se capturaban las relaciones reales dentro de los datos de origen. Los datos sintéticos permitieron obtener nuevos conocimientos que no habrían sido posibles sin las características de protección de la privacidad inherentes a los datos sintéticos.

---

[7] J. P. Morgan, "Synthetic Data for Real Insights", Technology Blog, n.d., https://www.jpmorgan.com/technology/technology-blog/synthetic-data-for-real-insights

[8] Contribución de Mastercard

(C) Salvaguardar los datos de los pacientes para el análisis de datos[9]

Problema: Antes de utilizar datos sintéticos, Johnson & Johnson (J&J) permitía a investigadores o consorcios externos acceder a datos de atención médica para propuestas de investigación validadas por J&J. Para salvaguardar la privacidad del paciente, los datos se transformaron en datos sanitarios anonimizados. Sin embargo, los comentarios recibidos indicaban que la utilidad general de los datos anonimizados, que se basaban en técnicas tradicionales de anonimización, no siempre era satisfactoria y no siempre cumplía los requisitos de los investigadores o consorcios.

Solución: J&J ha introducido datos sintéticos de alta calidad generados por IA como una opción adicional para procesar sus datos de atención médica.

Beneficio: Los investigadores y los clientes han experimentado una mejora significativa en el análisis. Cuando se emplea correctamente, esta forma de datos sintéticos puede representar eficazmente a la población objetivo y ofrecer diversos beneficios analíticos y científicos.

(D) Facilitar la colaboración de datos[10]

Problema: Una compañía farmacéutica quería comprar datos de salud relativos a cardiología a un instituto de investigación para probar una nueva hipótesis. Los datos de salud, que fueron recopilados por el instituto de investigación de personas que dieron su consentimiento, se alojaron en un entorno altamente regulado, como se requiere en el sector de la salud. Sin embargo, esto presenta desafíos significativos para fomentar actividades con otras entidades de explotación de los datos.

Solución: La empresa farmacéutica contrató a A*STAR para construir un conversor que crea copias sintéticas a partir los datos reales, que luego se pueden llevar fuera de este entorno regulado.

Beneficio: Esto permitió a la empresa farmacéutica obtener una vista previa de los datos y estar segura de la calidad de los datos antes de la compra de alto valor y el acceso a los datos reales.

[9] Contribución de Johnson & Johnson (J&J)
[10] Contribución de A*STAR

III. Recomendaciones

Los datos sintéticos tienen el potencial de impulsar el crecimiento de la IA/ML al permitir el entrenamiento de modelos de IA y proteger los datos personales subyacentes. También aborda los desafíos relacionados con los conjuntos de datos para el entrenamiento de modelos de IA, como los datos insuficientes y sesgados, al permitir el aumento y la extensión de la diversidad de los conjuntos de datos de entrenamiento.

Además, los datos sintéticos se pueden utilizar para facilitar y respaldar las necesidades de análisis de datos, colaboración y desarrollo de software de las organizaciones. Un beneficio adicional de usar datos sintéticos en lugar de datos de producción para facilitar el desarrollo de software es que se pueden evitar brechas de datos personales en caso de que el entorno de desarrollo se vea comprometido.

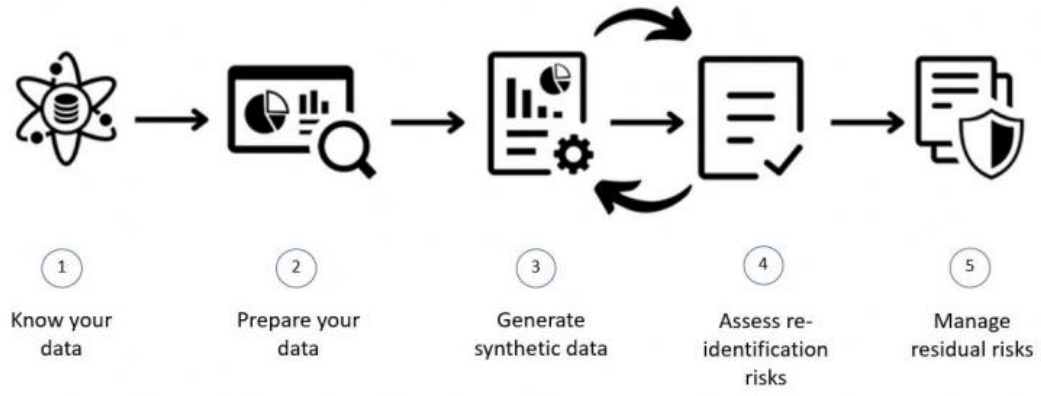
El PDPC recomienda un conjunto de buenas prácticas y evaluaciones/consideraciones de riesgos para generar datos sintéticos y reducir los riesgos residuales de la reidentificación a través de controles de gobernanza, procesos contractuales y medidas técnicas (consulte el **Anexo A**)

Anexo A: Manual sobre Consideraciones Clave y Mejores Prácticas en la Generación de Datos Sintéticos

En esta guía describimos las consideraciones clave y las mejores prácticas para que las organizaciones reduzcan los riesgos de reidentificación de datos sintéticos tabulares a través de un enfoque de cinco pasos.

Para cualquier otro conjunto de datos sintéticos complejos que no estén estructurados, se recomienda a las organizaciones que consideren la posibilidad de contratar a expertos en datos sintéticos, científicos de datos o evaluadores de riesgos independientes para evaluar y mitigar los riesgos de los datos sintéticos generados.

Descripción general del enfoque de cinco pasos para generar datos sintéticos



Paso 1: Conozca sus datos

Antes de embarcarse en cualquier proyecto de datos sintéticos, es necesario tener una comprensión clara del propósito y los casos de uso de los datos sintéticos y los datos de origen que los datos sintéticos deben imitar. Esto ayudará a determinar si el uso de datos sintéticos puede ser relevante e identificar los posibles riesgos de usar los datos sintéticos. Algunas de las consideraciones pueden incluir:

- Cuando las tendencias/percepciones generales de los datos de origen son sensibles, la organización debe tener en cuenta que el uso de datos sintéticos no ofrecerá ninguna protección a las tendencias/percepciones, ya que se replicarán en los datos sintéticos.
- Cuando los datos sintéticos están destinados a ser divulgados públicamente, las organizaciones pueden tener que priorizar la protección de datos sobre la utilidad de los datos en tales circunstancias.
- Cuando proceda, las organizaciones también deben establecer obligaciones contractuales adecuadas para los destinatarios de datos sintéticos cuando sea necesario para evitar ataques de reidentificación de los datos.

Con este conocimiento, el responsable, con la ayuda de las partes interesadas pertinentes, como el equipo de análisis de datos, deben establecer objetivos antes de la generación de datos sintéticos para determinar un umbral de riesgo aceptable[11] de los

[11] El umbral de riesgo de reidentificación representa el nivel de riesgo de reidentificación que es aceptable para un conjunto de datos sintéticos determinado. Actualmente no existe un valor

datos sintéticos generados y la utilidad esperada de los datos. Esto ayudará a proporcionar a las organizaciones los puntos de referencia adecuados para evaluar cualquier compensación entre los riesgos de protección de datos y la utilidad de los datos.

Estos índices de referencia pueden ajustarse adecuadamente para cumplir los objetivos empresariales, teniendo en cuenta las compensaciones entre los riesgos de utilidad de datos y protección de datos después del proceso de generación de datos sintéticos, así como las salvaguardias y controles para mitigar o reducir los riesgos residuales que plantean los datos sintéticos generados. Los criterios de aceptación deben incorporarse a las evaluaciones de riesgos de la organización (por ejemplo, marco de gestión de riesgos institucionales[12] si corresponde) o a una evaluación de impacto para la protección de datos ("EIPD")[13].

Paso 2: Preparar los datos

Al preparar los datos de origen[14] para generar datos sintéticos, es importante tener en cuenta lo siguiente:

- ¿Cuáles son las ideas clave que deben conservarse en los datos sintéticos?
- ¿Cuáles son los atributos de datos necesarios para que los datos sintéticos cumplan con los objetivos de negocio?

Comprender las ideas clave que se deben preservar

Para garantizar que los datos sintéticos puedan cumplir con los objetivos comerciales, las organizaciones deben comprender e identificar las tendencias, las propiedades estadísticas clave y las relaciones de atributos en los datos de origen que deben preservarse para el análisis (por ejemplo, identificar las relaciones entre las características demográficas de la población y sus condiciones de salud).

Las organizaciones deben considerar, en este punto, si las tendencias y los conocimientos atípicos son necesarios para preservar los objetivos comerciales. Las consideraciones clave podrían incluir lo siguiente:

- Si los valores atípicos no son necesarios para cumplir con los objetivos comerciales y el riesgo de reidentificación es alto, las organizaciones deben considerar eliminar los valores atípicos. Esto se puede hacer antes de la generación de datos sintéticos o en etapas posteriores de la generación de datos sintéticos.

---

numérico universalmente aceptado para el umbral de riesgo. Para obtener más detalles, consulte el **Paso 4** (Evaluar los riesgos de reidentificación).

[12] Las organizaciones pueden consultar ISO27001 para obtener más información sobre el desarrollo de un marco de gestión de riesgos empresariales.

[13] Un ejemplo de ello es la Guía de PDPC para las Evaluaciones de Impacto para la Protección de Datos. Una EIPD es aplicable en el caso de que se trate de datos personales. Es posible que la EIPD no sea pertinente en situaciones en las que la generación de datos sintéticos no implique el tratamiento de datos personales. N.T.: La AEPD tiene publicada sus guías y herramientas para la gestión del riesgo en https://www.aepd.es/areas-de-actuacion/innovacion-y-tecnologia#Riesgo

[14] Este paso supone que los datos de origen se han limpiado correctamente (por ejemplo, corrigiendo o eliminando datos incorrectos, dañados, con formato incorrecto, duplicados o incompletos) y que son de calidad aceptable para la generación de datos sintéticos.

- Si el objetivo es imitar las características de los datos de origen lo más fielmente posible, incluidos los valores atípicos, es posible que la organización tenga que preservar la tendencia/información atípica para cumplir con los objetivos empresariales. En tal caso, la organización debe tener en cuenta que los riesgos de reidentificación de las personas en los datos atípicos pueden ser altos y, por lo tanto, debe implementar medidas de mitigación de riesgos.
- Si el objetivo empresarial es equilibrar el número de datos concretos en diferentes categorías de datos, el propio proceso de generación de datos sintéticos puede ayudar a mitigar el problema de los valores atípicos simplemente generando más valores atípicos. Por ejemplo, en un conjunto de datos, el número de datos atípicos que comprenden individuos masculinos se puede equilibrar con puntos de datos atípicos que comprenden individuos femeninos.

Selección de atributos de datos

En función de la información clave necesaria, las organizaciones deben aplicar la minimización de datos para extraer solo los atributos de datos relevantes de los datos de origen. A partir de entonces, elimine o seudonimice todos los identificadores directos[15][14] de los datos extraídos.

Cuando no sea necesaria información detallada, las organizaciones pueden generalizar o añadir más ruido a los datos en este punto o en un paso posterior para reducir el riesgo de reidentificación. Por ejemplo, las organizaciones pueden generalizar la información exacta de altura y peso en bandas de altura y peso para reducir la posibilidad de que se utilicen combinaciones de altura y peso para identificar valores atípicos.

Las organizaciones también deben estandarizar y documentar los detalles de cada atributo de datos (como definiciones de datos, estándares, métricas, etc.) en un diccionario de datos. Esto permite a la organización validar posteriormente la integridad de los datos sintéticos generados para detectar anomalías y corregir cualquier inconsistencia en los datos. Consulte la siguiente lista de verificación en la **Tabla 3** para conocer las consideraciones clave.

Tabla 3: Lista de verificación para la preparación de datos

Lista de verificación de preparación de datos	
Comprender la información clave	
i.	Identificar las tendencias y las relaciones entre entidades que se conservarán para la generación de datos sintéticos.
ii.	Eliminar los valores atípicos si dichas tendencias/perspectivas no son necesarias. Esto se puede realizar después de la generación.
Selección de atributos de datos	
iii.	Aplicar la minimización de datos para seleccionar solo los atributos de datos que son necesarios para satisfacer las necesidades empresariales.
iv.	Eliminar o seudonimizar los identificadores directos (por ejemplo, nombre, números de identificación nacional).

[15] Consulte la Guía de anonimización básica de PDPC sobre cómo identificar identificadores directos en un conjunto de datos.

v.	Generalizar los datos granulares o añadir ruido (por ejemplo, utilizando la privacidad diferencial[16]) a los datos/modelo si dicha información detallada no es necesaria. Esto también se puede realizar después de la generación.
vi.	Estandarizar y documentar el formato, las restricciones y las categorías de los datos de origen en el diccionario de datos (consulte el **Anexo B** para obtener una plantilla de referencia): <u>Formato</u> • Estandarizar las cadenas a mayúsculas y minúsculas • Tipos de datos, nombres de columnas, estructuras, relaciones • Frecuencia de registro de datos Restricciones • Restricciones de valores para cada tipo de datos, por ejemplo, valores mínimos-máximos, valores no negativos, valores no nulos <u>Categoría</u> • Tipos de categorías de datos • Valores esperados o válidos para los atributos de datos dentro de cada categoría de datos. Un ejemplo de una categoría de datos es "país".

Paso 3: Generar datos sintéticos

Hay muchos métodos diferentes [17] para generar datos sintéticos, por ejemplo, sintetizadores secuenciales basados en árboles, cópulas [18] y modelos generativos profundos (DGM, por sus siglas en inglés). Las organizaciones deben considerar qué métodos son los más adecuados, en función de sus casos de uso, objetivos de datos y tipos de datos. Consulte el **Anexo C** para obtener más información sobre estos métodos de generación de datos sintéticos. A partir de entonces, las organizaciones pueden considerar dividir los datos de origen en dos conjuntos separados, por ejemplo, 80% como conjunto de datos de entrenamiento y 20% como conjunto de datos de control[19] para evaluar los riesgos de reidentificación de los datos sintéticos.

Después de generar datos sintéticos, es una buena práctica que las organizaciones realicen las siguientes comprobaciones sobre la calidad de los datos sintéticos generados:

- Integridad de los datos
- Fidelidad de los datos
- Utilidad de los datos

<u>Integridad de los datos</u>

La integridad de los datos garantiza la exactitud, integridad, coherencia y validez de los datos sintéticos en comparación con los datos de origen. Las organizaciones pueden

[16] El uso de la privacidad diferencial para añadir ruido a los datos sintéticos es ampliamente discutido como un mecanismo para reducir los riesgos de reidentificación. Sin embargo, actualmente no existe un estándar universal sobre cómo implementar la privacidad diferencial. Además, el ruido añadido también puede reducir la utilidad de los datos sintéticos, haciéndolos menos precisos o útiles para ciertos tipos de análisis.

[17] Es posible que esta guía no sea exhaustiva para cubrir todos los demás métodos de generación de datos sintéticos, como el modelo bayesiano y los autocodificadores variacionales (VAE, por sus siglas en inglés).

[18] N.T.: funciones que describen las dependencias estructurales en los datos reales.

[19] Véase el Enfoque 2 del Anexo E para obtener más detalles sobre la evaluación y el marco de evaluación para cuantificar el riesgo de reidentificación.

validar la integridad de los datos sintéticos generados con el diccionario de los datos de origen.

Fidelidad de los datos

La fidelidad de los datos examina si los datos sintéticos siguen de cerca las características y los atributos estadísticos de los datos de origen. Hay algunas métricas para medir la fidelidad de los datos y, por lo general, se realizan comparando estadísticamente los datos sintéticos generados directamente con los datos de origen. Las organizaciones deben utilizar las métricas de rendimiento para la fidelidad de los datos[20] (véase la **Tabla 4**) que mejor se ajusten a sus objetivos de datos.

Tabla 4: Métricas de rendimiento para la fidelidad de los datos

Métricas de rendimiento generalmente utilizadas para evaluar la fidelidad de los datos	
Similitud basada en histogramas	Mide la similitud entre las distribuciones de los datos de origen y sintéticos a través de una comparación de histogramas de cada característica. Esto garantiza que los datos sintéticos conserven propiedades estadísticas importantes, como la tendencia central (media, mediana), la dispersión (varianza, rango) y la forma de distribución (asimetría, curtosis)
Similitud correlacional	Mide la conservación de las relaciones entre los atributos de los datasets de origen y sintéticos. Por ejemplo, si la educación superior se vincula con mayores ingresos en los datos de origen, este patrón también debería ser evidente en los datos sintéticos.

Utilidad de los datos

La utilidad de datos se refiere a la forma en que los datos sintéticos pueden reemplazar o agregar a los datos de origen para el objetivo de datos específico de la organización.

Existen diferentes enfoques para evaluar la utilidad de los datos sintéticos. La verdadera prueba de utilidad es cómo se desempeña en las tareas del mundo real. Un enfoque común para comprobarlo es entrenar modelos de IA/ML idénticos con datos sintéticos y de entrenamiento. Los rendimientos de los dos modelos se comparan con el conjunto de datos de control, simulando pruebas en el entorno de producción, para evaluar la utilidad de los datos sintéticos. Ejemplos de métricas de rendimiento generalmente utilizadas incluyen "exactitud", "precisión", "recuperación", "Puntuación F1" o "Área bajo la curva ROC (AUC-ROC)" para tareas de clasificación, y "Error absoluto medio (MAE)" o "Error cuadrático medio (MSE)" para tareas de regresión[21] (consulte la definición en la **Tabla 5** a continuación). Si sus puntuaciones comparadas son cercanas, indica que los datos sintéticos tienen una alta utilidad. En términos simples, una puntuación de utilidad alta

[20] Hay otras métricas genéricas descritas aquí, además de las enumeradas en la Tabla 4. Véase Khaled El Emam et al., "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study", JMIR Medical Informatics 10, n.º 4 (2022).

[21] Existe otra métrica de rendimiento adecuada para las tareas de regresión, es decir, la replicabilidad, que se utiliza para evaluar la utilidad de los datos y se describe aquí además de las enumeradas en la Tabla 5. Véase Khaled El Emam et al., "An Evaluation of the Replicability of Analyses Using Synthetic Health Data", Scientific Reports 14 (2024), https://www.nature.com/articles/s41598-024-57207-7

significa que las máquinas entrenadas con datos sintéticos funcionan de manera similar a las entrenadas con datos de entrenamiento.

Cuando se trata de maximizar la utilidad de los datos, a menudo existe una compensación inherente entre la utilidad de los datos y la protección de datos. Por lo tanto, es necesario lograr un delicado equilibrio entre la utilidad de los datos y la protección de los datos a través de un proceso iterativo (Pasos 3 y 4) para sintetizar los datos hasta un nivel aceptable para los riesgos de reidentificación, al tiempo que se encuentra el equilibrio adecuado de la utilidad de los datos

Tabla 5: Métricas de rendimiento para la utilidad de los datos

Métricas de rendimiento generalmente utilizadas para evaluar la utilidad de los datos	
Exactitud	Mide la exactitud general del modelo. Se calcula como la relación entre las predicciones correctas (verdaderos positivos y verdaderos negativos) y el total de observaciones. *Por ejemplo, si de los datos de salud de 100 individuos, el modelo predice correctamente el estado de salud de 90 de estos individuos, la precisión es del 90%.*
Precisión	Mide la capacidad del modelo para identificar solo instancias relevantes. Se calcula como la relación entre las predicciones positivas correctas (verdaderos positivos) y todas las predicciones positivas (verdaderos positivos y falsos positivos). *Por ejemplo, si de 100 individuos que el modelo predice como enfermos, 80 de estos individuos se identifican correctamente como enfermos, la precisión es del 80%*
Recall	Mide la capacidad del modelo para encontrar todos los casos relevantes. Se calcula como la relación entre las predicciones positivas correctas (verdaderos positivos) y todos los positivos reales (verdaderos positivos y falsos negativos). *Por ejemplo, si de 100 individuos enfermos, el modelo predice que 90 de estos individuos están enfermos, el recuerdo es del 90%.*
Puntuación F1	Equilibra la precisión y la recuperación en una sola métrica. (matemáticamente, es la media armónica[22] de precisión y recuperación).
Área bajo la curva ROC (AUC-ROC)	Mide la capacidad del modelo para distinguir entre clases. Está representado por el área bajo la curva de características operativas del receptor (AUC-ROC), comparando la tasa de verdaderos positivos con la tasa de falsos positivos en varios umbrales de clasificación. *Por ejemplo, si la puntuación AUC-ROC es 0,9, significa que hay un 90% de posibilidades de que el modelo distinga correctamente entre una instancia positiva elegida al azar y una instancia negativa elegida al azar.*
Error absoluto medio (MAE)	Mide los errores del modelo en las predicciones promediando las diferencias absolutas entre los valores predichos y reales, proporcionando una medida directa de la magnitud media del error sin tener en cuenta la dirección del error. Se calcula como la media de las diferencias absolutas entre los valores reales y los previstos.

[22] Un tipo de promedio que da más peso a los valores más bajos de las puntuaciones de precisión y recuperación.

Error cuadrático medio (MSE)	Mide los errores de predicción del modelo promediando los cuadrados de los errores entre los valores predichos y reales. MSE penaliza en gran medida los errores más grandes que los más pequeños, debido a la elevación al cuadrado de los valores de error. Esto lo hace más sensible a los valores atípicos y a los errores grandes. Se calcula como la media de las diferencias al cuadrado entre los valores reales y los previstos

Utilice la siguiente lista de verificación de la **Tabla 6** como guía de referencia cuando corresponda.

Tabla 6: Lista de comprobación para la comprobación de los datos sintéticos generados

Lista de verificación posterior a la generación	
i.	Eliminar los valores atípicos si dichas tendencias/conocimientos no son necesarios para satisfacer las necesidades empresariales.
ii.	Generalizar los datos granulares o añadir ruido a los datos/modelo si dicha información detallada no es necesaria.
iii.	Realizar comprobaciones de integridad de datos en datos sintéticos validando el formato de los datos, las estructuras, etc. con el diccionario de datos documentado anterior.
iv.	Seleccionar métricas relevantes que cumplan con los objetivos de datos para medir la fidelidad de los datos.
v.	Seleccionar métricas de rendimiento relevantes que cumplan con los objetivos de datos para medir la utilidad de los datos.

Paso 4: Evalúe el riesgo de reidentificación[23]

Una vez que se generan los datos sintéticos y se evalúa que la medición de la utilidad es aceptable, las organizaciones deben evaluar y realizar la evaluación de riesgos de reidentificación en función de sus criterios de aceptación internos. En el **Anexo D** se examinan los riesgos ampliamente conocidos de reidentificación de los datos sintéticos. Dado que los datos sintéticos generalmente no replican sus datos concretos de entrenamiento, el riesgo de reidentificación no se puede deducir directamente del escrutinio de si los datos sintéticos generados contienen datos personales.

Por lo general, la evaluación del riesgo de reidentificación (o privacidad) de los datos sintéticos es una evaluación basada en ataques. Evalúa el éxito con el que un adversario, que lleva a cabo ataques de redentificación mediante ataques de singularización, ataques de vinculabilidad y ataques de inferencia (como se describe en el **Anexo D**) en conjuntos de datos sintéticos, puede determinar si un individuo pertenece al conjunto de datos de origen (es decir, inferencia de pertenencia) y/o derivar detalles de un individuo del conjunto de datos de origen que de otro modo no se revelarían (es decir, inferencia de atributos). El objetivo de las organizaciones es garantizar que los niveles de riesgo de reidentificación para los tres ataques de reidentificación clave sean aceptables. Si el nivel de riesgo de reidentificación es inaceptable, repita el paso 3 para volver a generar datos sintéticos y cumplir con el nivel de riesgo aceptable. Esto puede lograrse aplicando más controles de protección de datos sobre los datos de origen, por ejemplo, generalizando los datos o

[23] N.T.: https://www.aepd.es/prensa-y-comunicacion/blog/anonimizacion-iii-el-riesgo-de-la-reidentificacion

añadiendo ruido (véase "Lista de comprobación para la preparación de datos" en la **Tabla 3**).

Se han propuesto varios enfoques para determinar y cuantificar los riesgos de reidentificación. En el **Anexo E** figuran ejemplos de estos enfoques. Es posible que las organizaciones deban contratar al proveedor de soluciones de datos sintéticos para realizar las evaluaciones de riesgos de reidentificación.

Si bien no existe un valor umbral numérico universalmente aceptado para el nivel de riesgo, algunas organizaciones[24] han optado por alinear su nivel de riesgo de reidentificación con las directrices y recomendaciones existentes de la industria para datos anónimos/anonimizados (véase la **Tabla 7**). Sin embargo, las organizaciones deben tener en cuenta que el método de cálculo para el umbral de reidentificación en un conjunto de datos anonimizado/anonimizado es muy diferente al de un conjunto de datos sintético. Sin embargo, la base fundamental para ambos es que la evaluación del riesgo de reidentificación/privacidad es una medición probabilística.

Los valores del umbral de riesgo de reidentificación de la **Tabla 7** resumen el umbral de riesgo aceptable precedente utilizado por algunas organizaciones para evaluar la anonimización/anonimización

Tabla 7: Directrices existentes sobre el umbral de riesgo para la desanonimización/anonimización

Umbral de riesgo para la desanonimización/anonimización	
Agencia Europea de Medicamentos (EMA, por sus siglas en inglés)	La Agencia Europea de Medicamentos (EMA) estableció una política sobre la publicación de datos clínicos de medicamentos. Las directrices que acompañan a la política recomiendan un umbral de riesgo máximo de 0,09.[25]
Ministerio de Salud de Canadá	Health Canada implementó el mismo umbral que la EMA, 0,09, para el intercambio de datos clínicos.[26]
ISO/IEC 27559 *Privacy enhancing data de-identification framework*.	La norma ISO/IEC 27559 resume una lista de umbrales de ejemplo que proporciona un rango de valores aceptables que abarca 0,09.

Paso 5: Gestionar los riesgos residuales

En este último paso, las organizaciones deben identificar todos los posibles riesgos residuales e implementar controles de mitigación adecuados (técnicos, de gobernanza y contractuales) para minimizar los riesgos identificados. Estos riesgos y controles deben

[24] Samer El Kababji et al., "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets," *JCO Clinical Cancer Informatics* 7 (2023), https://ascopubs.org/doi/full/10.1200/CCI.23.00116 .

[25] European Medicines Agency, "European Medicines Agency Policy on Publication of Clinical Data for Medicinal Products for Human Use," 2019, https://www.ema.europa.eu/en/documents/other/policy-70-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use_en.pdf

[26] Health Canada, "Guidance Document on Public Release of Clinical Information: Profile Page," 2019, https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html

ser documentados y aprobados por la dirección y las partes interesadas clave como parte del marco de riesgos empresariales de la organización.

Las organizaciones pueden tener en cuenta los siguientes riesgos como parte de la evaluación de riesgos.

Nuevos conocimientos derivados de los datos sintéticos

Se pueden obtener nuevos conocimientos sobre los datos de origen mediante el análisis del conjunto de datos sintéticos solo o en combinación con otros conjuntos de datos disponibles. Las organizaciones deben evaluar si estos conocimientos pueden ser sensibles o pueden desinformar.

Impacto potencial en grupos de individuos debido a la divulgación de la pertenencia

La divulgación de la pertenencia es cuando un adversario, utilizando la información de datos sintéticos, determina que un grupo objetivo de individuos fue incluido en el conjunto de datos de origen. Las posibles divulgaciones o inferencias de atributos del conjunto de datos sintéticos relacionados con grupos de personas pueden considerarse de naturaleza confidencial, por ejemplo, puede haber un impacto de estigma social en las personas de un grupo de asesoramiento si se revelara su membresía.

Al determinar el conjunto de datos de origen para los datos sintéticos, es importante tener en cuenta también la fracción de muestreo del conjunto de datos de población, que es la relación entre el tamaño de la muestra dentro de los datos de origen en comparación con el tamaño de la población. Por ejemplo, un adversario tendrá menos posibilidades de predecir si un grupo objetivo de individuos de una población está incluido en un conjunto de datos sintético que se entrena a partir de un conjunto de datos de origen muestreado del 20% de la población, en comparación con un conjunto de datos de origen muestreado del 90% de la población.

Partes que reciben datos sintéticos

Las partes receptoras de los datos sintéticos, incluidos los intermediarios de datos, pueden plantear riesgos de cumplimiento de brechas de datos al manejar datos sintéticos. Las organizaciones deben evaluar la capacidad y la motivación del destinatario de los datos para volver a identificar a las personas a partir del conjunto de datos. Un destinatario de datos que posea habilidades o tecnologías especializadas puede ser capaz de combinar conocimientos especiales u obtener conocimiento público para volver a identificar a cualquier individuo del conjunto de datos. Dichos riesgos deben tenerse en cuenta en el ejercicio de evaluación de riesgos.

Entorno cambiante

La probabilidad de que se vuelvan a identificar los riesgos de cualquier conjunto de datos sintéticos aumenta con el tiempo, debido al aumento de la potencia informática y a la mejora de las técnicas de enlace de datos.

Fuga de modelos

Un modelo que ha sido entrenado con datos de origen para generar los datos sintéticos puede ser susceptible de un ataque malintencionado por parte del adversario para reconstruir (partes de) los datos de origen.

Salvaguardas y mejores prácticas

En la siguiente **Tabla 8** se enumeran ejemplos de las mejores prácticas que las organizaciones pueden considerar implementar para gestionar los riesgos residuales que plantea el uso de datos sintéticos.

Tabla 8. Mejores prácticas y controles de seguridad para implementar y gestionar riesgos

Gobernanza	Controles de acceso	Implemente el control de acceso para los datos de origen y el modelo de generador de datos sintéticos. Aplique el control de acceso a los datos sintéticos en los que la reidentificación o el riesgo residual sean altos, especialmente si los datos contienen información o conocimientos muy confidenciales.
	Gestión de activos	Etiquete correctamente los datos sintéticos para evitar errores humanos al administrar tanto los datos de origen como los datos sintéticos.
	Gestión de riesgos	Llevar a cabo periódicamente revisiones de riesgo de reidentificación de conjuntos de datos sintéticos, especialmente si estos se dan a conocer públicamente.
	Controles legales	Contar con acuerdos contractuales para detallar las responsabilidades de los terceros destinatarios de los datos sintéticos y/o modelos, así como de cualquier proveedor de soluciones de terceros que proporcione las herramientas de generación de datos sintéticos. Esto incluye salvaguardar los datos/modelo y prohibir los intentos de volver a identificar a las personas. En situaciones en las que la organización puede necesitar depender del proveedor de soluciones de datos sintéticos para realizar la evaluación de riesgos y las mitigaciones, es posible que se requiera que el proveedor de soluciones proporcione garantías de que se han implementado los controles adecuados.
Controles TIC	Seguridad de la base de datos	Almacenamiento segregado para datos sintéticos y datos de origen.
Operaciones de TI	Registro y supervisión	Registre y supervise correctamente el uso de los datos de origen y sintéticos, así como el acceso al modelo de generador de datos sintéticos.
Gestión de riesgos	Mantenimiento de la información	Elimine de forma segura los datos de origen, los datos sintéticos y el modelo generador de datos sintéticos cuando ya no sean necesarios o hayan llegado al final del período de retención.

Gestión de incidentes

Las organizaciones deben identificar los riesgos de brechas de datos que involucran datos sintéticos, modelo generador de datos sintéticos y parámetros del modelo, e incorporar escenarios relevantes en sus planes de gestión de incidentes. Las siguientes consideraciones pueden ser relevantes para las investigaciones internas de las organizaciones[27].

Pérdida de datos totalmente sintéticos (para datos sintéticos que no están destinados a ser divulgados públicamente)

Los datos totalmente sintéticos que tienen incorporadas las mejores prácticas de protección de datos en su proceso de generación y que se ha evaluado como de bajo riesgo de reidentificación generalmente no se consideran datos personales. Sin embargo, las organizaciones deben proceder a investigar el incidente para comprender la causa raíz y mejorar sus salvaguardas internas contra tales sucesos en el futuro. Las organizaciones también deben supervisar si hay alguna evidencia de reidentificación real y evaluar si se trata de una brecha de datos notificable a la PDPC.

Pérdida del modelo generador de datos sintéticos, parámetros y/o datos sintéticos

Tanto el modelo generador de datos sintéticos como sus parámetros pueden proporcionar información útil a un adversario para realizar un ataque de inversión del modelo. Con el acceso a los datos sintéticos generados, puede mejorar aún más la capacidad del adversario para recuperar los datos de origen. Las organizaciones deben proceder a investigar el incidente para comprender la causa raíz y mejorar sus salvaguardas internas. También debe vigilar un posible ataque exitoso de inversión del modelo que pueda dar lugar a la reconstrucción y divulgación de los datos de origen. Cuando se detecte dicha reconstrucción y divulgación de los datos de origen, las organizaciones tendrán que evaluar si dicha brecha fuera notificable.

[27] Para la notificación de brechas de datos a la PDPC, las organizaciones tendrán que evaluar si se trata de una brecha notificable en función de la obligación de notificación de brechas de datos de la PDPA.

Anexo B: Formato de diccionario de datos

A continuación, se muestra un ejemplo de formato de diccionario de datos:

COLUMNA	DESCRIPCIÓN	POSIBLES VALORES	OBSERVACIONES
NOMBRE	Nombre de la columna	Sexo, fecha de nacimiento	
DESCRIPCIÓN	Breve descripción de la variable		
TIPO	Tipo de datos generales. En concreto, cómo aparece superficialmente	Numérico, cadena, fecha	Los datos a menudo pueden aparecer en varios formatos. Por ejemplo, los datos categóricos a menudo se pueden guardar como números enteros ordinales, booleanos (que parecen numéricos), o como texto, por ejemplo, 'SÍ' o 'NO', o simplemente como una secuencia única de números, por ejemplo, '3828' y '4271' (que parece numérico, pero en realidad es una cadena). Otro ejemplo es cuando las fechas se guardan como cadenas, o como formatos de fecha usando Excel, o su equivalente numérico. Estas dos columnas ayudan a los usuarios de datos a navegar por esta confusión y facilitan el desarrollo de scripts automatizados. TIPO indica cómo aparecen los datos superficialmente, si deben procesarse como un número, una cadena o una fecha al cargar o guardar los datos. TIPO DE DATOS ESPECÍFICO indica cómo se deben procesar los datos en la situación ideal durante los análisis, por ejemplo. Lo ideal es que un tipo de datos, como Education Level, sea ordinal, aunque se puede tratar como categórico, ordinal o intervalo.
TIPO DE DATOS ESPECÍFICO	La forma ideal de tratar la variable	Ej. 1: Si es numérico, especifique como 'float', 'int', 'boolean', 'ordinal', 'categórico', 'date', etc. Ej. 2: Si es una cadena, especifique como 'categórico', 'ordinal', 'texto libre', etc. Ej. 3: Si la fecha, especifique como 'fecha'	
CÓDIGOS		Ej. 1: Si TIPO es “fecha” utilice la convención de Excel para indicar el formato de fecha, por	Preste especial atención a las letras mayúsculas o minúsculas para evitar confusiones

		ejemplo, dd/mm/aaaa, mm-dd-aaaa, etc. Ej. 2: Si TIPO DE DATOS ESPECÍFICO es 'booleano', 'ordinal', 'categórico', especifique exhaustivamente todas las entradas posibles, delimitadas por '; ', por ejemplo, SÍ; NO; N.A. o Hombre; Mujer, OR 1; 2; 3; 4; 5 Ej. 3: Si TIPO es 'numérico', especifique el rango. Por ejemplo, [0,100] O (3,4).	
FRECUENCIA	Para datos longitudinales. Se utiliza para indicar si la variable se recopila durante un tipo de visita determinado	Ej. 1: LÍNEA BASE; 6 SEMANAS; 6 MES Ej. 2: VISITA 1; VISITA 2	Deje en blanco si no son datos longitudinales
CATEGORÍA	Se utiliza para agrupar la variable en una categoría específica	Ej. 1: DEMOGRAFÍA Ej. 2: ECO Ej. 3: VARIABLE DE ESTILO DE VIDA	
SECUNDARIA	Si la variable se puede derivar de otras variables presentes en el conjunto de datos.	Ej. 1: En caso afirmativo, 'Y' Ej. 2: Si no, 'N' o déjelo en blanco.	Por ejemplo, el IMC es una variable secundaria si se calcula a partir de la "altura" y el "peso", y las dos variables también se incluyen en el conjunto de datos. Otros ejemplos son 'década', donde los sujetos de entre 30 y 40 años se agrupan como '30-40' para reducir la granularidad de la variable, o 'demencia', un diagnóstico derivado de las respuestas a preguntas también presentes en el conjunto de datos. En caso afirmativo, explique cómo se calculó la variable a partir de otras variables, como la fórmula del IMC, el estándar/criterio de

			diagnóstico, etc., ya sea en la columna RESTRICCIONES o OBSERVACIONES.
RESTRICCIONES	Cómo la variable depende de otras variables	Ej. 1: 'Head_circ' (circunferencia de la cabeza) es una variable recogida para la 'edad' <= 6. Dejar vacío si la "edad" > 6. Ej. 2: Se recogieron solo para los datos de la cohorte '<NOMBRE DE LA COHORTE>' o del hospital '<HOSPITAL A>'. Ej. 3: 'Ever_pregnant' solo se recolecta para mujeres mayores de 12 años. Si es "hombre" o "mujer" menor de 12 años, registrado como "N.A." Si es "mujer" mayor de 12 años, "SÍ", "NO" o "DESCONOCIDO". Ej. 4: 'IMC' solo computable si también se recogen 'altura' y 'peso'. Déjelo en blanco si alguno de los valores está en blanco	Esta información ayudará a los usuarios de datos a decidir si "falta valor"/"se desconoce" (se debe recopilar pero no ha recopilado) o si no es aplicable (no recogida debido al procedimiento). Tenga en cuenta que el valor de una variable puede depender (o ser condicional) de otras variables, pero no necesariamente se deriva de otras variables; RESTRICCIONES y SECUNDARIA son complementarios, pero el primero no implica el segundo.
COMENTARIOS	Comentarios adicionales, como la forma en que se codifican los datos y/o preocupaciones relacionadas con la variable	Ej. 1: Cómo se codifican las variables categóricas como números enteros: 1 = NO, 0 = SÍ, - 1 = N.A. Ej. 2: Variable sensible OR autoinformada, etc. Ej. 3: Unidad métrica utilizada para la colección, 'cm', 'm', 'pulgadas', etc.	A menudo es necesario dejar una nota para recordar a los responsables/usuarios de los datos las dificultades encontradas durante la recopilación de datos, las acciones tomadas y las preocupaciones asociadas. Algunas de estas observaciones pueden incluirse en la descripción de la variable, o aquí, si se consideran misceláneas.

Anexo C: Ejemplos de Métodos de Generación de Datos Sintéticos

Métodos Estadísticos

(A) Redes Bayesianas

Contribución de Betterdata.ai

Las redes bayesianas (BN) son modelos probabilísticos que utilizan un grafo acíclico dirigido (DAG) para representar dependencias condicionales entre variables, lo que permite la generación de datos sintéticos estadísticamente similares a los datos originales. Los BN son útiles en sectores como la sanidad y las finanzas, donde las relaciones con los datos precisos son esenciales. Normalmente, las BN requieren una gran experiencia en el dominio para una modelización precisa a través de un enfoque dirigido por expertos [28] . Alternativamente, también se pueden estructurar a través de métodos basados en datos, aunque estos comprometen la precisión debido a inferencias menos confiables sobre las relaciones de datos subyacentes. PrivBayes[29] es un ejemplo de BN que aborda datos de dimensiones moderadas al tiempo que preserva la privacidad. Construye una red bayesiana para modelar las relaciones entre los atributos de los datos y aproxima la distribución de los datos utilizando márgenes de baja dimensión. Al inyectar ruido en estos márgenes para garantizar la privacidad, PrivBayes genera un conjunto de datos sintético que refleja fielmente el original al tiempo que logra un equilibrio eficiente entre la utilidad de los datos y la privacidad.

Sin embargo, la escalabilidad [30] de las BN es limitada, ya que su complejidad computacional puede variar de polinómica a exponencial dependiendo del número de características y algoritmos de aprendizaje utilizados.

La complejidad polinómica se puede lograr con estructuras definidas por expertos o mediante la implementación de restricciones de precisión, como un número limitado de padres por nodo. Sin estas restricciones, el aprendizaje se convierte en un problema NP-difícil, lo que hace que la complejidad aumente exponencialmente con el número de características. Si bien los algoritmos de aproximación pueden ayudar a gestionar las demandas computacionales, pueden reducir la precisión. Por lo tanto, los BN se ven favorecidos para escenarios que requieren interpretabilidad, pero menos para conjuntos de datos de alta dimensión

---

[28] Anthony Costa Constantinou, Norman Fenton, and Martin Neil, "Integrating Expert Knowledge with Data in Bayesian Networks: Preserving Data-Driven Expectations When the Expert Variables Remain Unobserved," *Expert Systems with Applications* 56 (2016): 197–208, https://www.sciencedirect.com/journal/expert-systems-with-applications/vol/56/suppl/C

[29] Ergute Bao et al., "Synthetic Data Generation with Differential Privacy via Bayesian Networks," *Journal of Privacy and Confidentiality* 11, no. 3 (2021), https://dr.ntu.edu.sg/handle/10356/164213

[30] Ole J. Mengshoel, "Understanding the Scalability of Bayesian Network Inference Using Clique Tree Growth Curves," *Artificial Intelligence* 174, no. 12–13 (2010): 987–1006, https://ntrs.nasa.gov/api/citations/20090033938/downloads/20090033938.pdf

donde el aprendizaje profundo ofrece una solución más práctica debido a su capacidad para manejar datos a gran escala de manera eficiente.

(B) Cópulas condicionales

Contribución de la Agencia para la Ciencia, la Tecnología y la Investigación (A*STAR)

Las cópulas condicionales son las más adecuadas para la generación de datos sintéticos cuando los conjuntos de datos de entrenamiento tienen un tamaño moderado, lo que a menudo genera una replicación robusta y eficiente en el tiempo de las distribuciones conjuntas de datos requeridas. En comparación con los métodos de aprendizaje automático relativamente costosos, que como proceso basado en datos depende en gran medida de la cardinalidad y el tamaño de los datos de entrenamiento disponibles, las cópulas proporcionan una alternativa rentable que equilibra la disponibilidad de datos con el conocimiento experto previo, generando diversos conjuntos de muestras basados en condiciones predeterminadas para pruebas metodológicas y entrenamiento de algoritmos.

El marco centrado en cópula elíptica para la generación de datos sintéticos es un proceso simple de dos pasos. En el primer paso, se estiman las distribuciones marginales de las variables de entrada, seguidas de sus parámetros de correlación por pares, y luego se combinan ambos para reproducir una estimación estadística de la distribución conjunta del conjunto de datos de entrenamiento. El segundo paso es relativamente sencillo; Uno simplemente toma muestras de la distribución conjunta aprendida para producir cualquier número de puntos de muestra sintéticos que requiera, bastante seguro de sus propiedades estadísticas con referencia a lo que se ha aprendido. El marco de la cópula condicional mejora aún más el primero al afinar el proceso de aprendizaje; Utilizando la distribución conjunta aprendida como línea de base, se divide el conjunto de datos de entrenamiento en subconjuntos significativos en función de las condiciones identificadas, como grupos de edad, sexo, razas, etc., a través de la reiteración del proceso de aprendizaje, esencialmente se vuelven a muestrear los puntos de datos generados con distribuciones condicionales renovadas (y condiciones). Esta mejora adicional mejora la flexibilidad del método centrado en cópula y lo adapta a conjuntos de datos de entrenamiento complejos con distribuciones multimodales o incluso relaciones no monótonas y no lineales.

La implementación detallada y el rendimiento de este método están disponibles en https://github.com/BiomedDAR/copula-tabular

(C) Síntesis de datos Marginal-Based

Aportado por el profesor Xiao Xiaokui, Escuela de Computación, Universidad Nacional de Singapur

La síntesis de datos *Marginal-Based* es un enfoque ampliamente utilizado para sintetizar datos tabulares. Este enfoque implica seleccionar un conjunto de marginales de una tabla de entrada *T*, cada uno de los cuales es un proyecto de *T* en un subconjunto de sus atributos. Por ejemplo, considere la siguiente tabla *T* con 4 atributos: Edad, Género, Educación, Ocupación e Ingresos.

Edad	Género	Educación	Ocupación	Ingresos
…	…	…	…	…

Tabla *T*

A continuación, se presentan algunos ejemplos de posibles marginales de *T*:

Edad	Género	Educación
…	…	…

Marginal de *T* en {edad, género, educación}

Edad	Ocupación	Ingresos
…	…	…

Marginal de *T* en {edad, género, ingresos}

Género	Ocupación
…	…

Marginal de *T* en {género, ocupación}

Después de elegir un conjunto de marginales, el enfoque construye un modelo estadístico (por ejemplo, una red bayesiana[31]) para capturar las correlaciones entre los atributos dentro de los marginales. A continuación, este modelo se utiliza para generar datos sintéticos que conservan las correlaciones de atributos.

La síntesis de datos *marginal-based* tiene tres ventajas:

- Simplicidad: el concepto es simple y fácil de entender.
- Efectividad: cuando los márgenes elegidos cubren todas las correlaciones de atributos importantes, los datos sintéticos podrían preservar las propiedades estadísticas de los datos originales.
- Privacidad: el proceso de síntesis de datos podría ofrecer una fuerte protección de la privacidad, si se introduce cuidadosamente el ruido durante la selección y construcción de los márgenes y el entrenamiento del modelo estadístico.

La síntesis de datos marginal-based ha ganado una adopción generalizada en aplicaciones prácticas. Los métodos representativos incluyen PrivBayes[32], MST[33]y PrivMRF[34]. En particular, PrivBayes y MST estuvieron entre los ganadores del NIST Differential Privacy Synthetic Data Challenge[35]de 2018, mientras que PrivMRF ganó

[31] "Bayesian Network," Wikipedia, 2024, https://en.wikipedia.org/wiki/Bayesian_network

[32] Jun Zhang et al., "PrivBayes: Private Data Release via Bayesian Networks," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014, 1423–34, https://dl.acm.org/doi/10.1145/2588555.2588573

[33] Ryan McKenna, Gerome Miklau, and Daniel Sheldon, "Winning the NIST Contest: A Scalable and General Approach to Differentially Private Synthetic Data," *Journal of Privacy and Confidentiality* 11, no. 3 (2021), https://journalprivacyconfidentiality.org/index.php/jpc/article/view/778

[34] Kuntai Cai et al., "Data Synthesis via Differentially Private Markov Random Fields," Github, n.d., https://github.com/caicre/PrivMRF

[35] National Institute of Standards and Technology, "Disassociability Tools," NIST, 2023, https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools#dpchallenge

el primer lugar en la edición de 2020 del desafío[36]. Además, PrivBayes se ha implementado en Data Intelligence Cloud de SAP[37], así como en numerosas herramientas de síntesis de datos de código abierto[38].

(D) Sintetizadores secuenciales basados en árboles (SEQ)

Contribución del Dr. Khaled El Emam, Universidad de Ottawa

Una forma de generar datos sintéticos es aplicar un árbol de decisión secuencial basado en algoritmos de regresión y árboles de clasificación ("CART") de uso común, aunque también se pueden usar variantes (por ejemplo, árboles potenciados) de estos. El principio es sintetizar secuencialmente las variables utilizando modelos de clasificación y regresión.[39]

Se puede pensar que el proceso se ajusta inicialmente a una serie de modelos. Estos modelos componen el generador. Luego, estos modelos se pueden utilizar para generar datos. Cuando se utiliza un modelo para generar datos, tomamos muestras del nodo terminal predicho para obtener los valores sintéticos. La distribución de en el nodo se puede suavizar antes del muestreo.

Véase la nota a pie de página para ver los documentos pertinentes sobre este método.[40]

Modelos Generativos Profundos

(E) Redes generativas adversarias (GAN), aportado por Betterdata.ai

Las redes generativas adversarias (GAN, por sus siglas en inglés) son modelos generativos profundos que sobresalen en la síntesis de conjuntos de datos complejos y de alta dimensión. A través de un proceso antagónico, el generador crea datos sintéticos que un discriminador evalúa para determinar su realismo, lo que provoca una mejora continua en la salida sintética. Este refinamiento iterativo permite a las GAN producir datos sintéticos que se parecen mucho a los originales, superando a las técnicas de aprendizaje no profundo en conjuntos de datos complejos del mundo real.

---

[36] National Institute of Standards and Technology, "2020 Differential Privacy Temporal Map Challenge," NIST, 2022, https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2020-differential-privacy-temporal

[37] SAP Community, "SAP Data Intelligence: Data Synthesizer for Machine Learning Operator", Technology Blogs by SAP, 2021, https://community.sap.com/t5/technology-blogs-by-sap/sap-dataintelligence-data-synthesizer-for-machine-learning-operator/ba-p/13501498

[38] "Reprosyn: Synthesising Tabular Data," Github, 2022, https://github.com/alan-turing-institute/reprosyn; "Synthcity," Github, 2024, https://github.com/vanderschaarlab/synthcity; "DataSynthesizer," Github, 2023, https://github.com/DataResponsibly/DataSynthesizer DataCebo, "SDGym," Github, 2024, https://github.com/sdv-dev/SDGym; "DPART | Differentially Private Auto-Regressive Tabular," Github, 2024, https://github.com/hazy/dpart

[39] Para obtener más información, consulte Khaled El Emam, Lucy Mosquera, and Richard Hoptroff, "Evaluating Synthetic Data Utility," in *Practical Synthetic Data Generation Balancing: Privacy and the Broad Availability of Data* (O'Reilly Media, Inc, 2020)

[40] Khaled El Emam, Lucy Mosquera, and Chaoyi Zheng, "Optimizing the Synthesis of Clinical Trial Data Using Sequential Trees," *Journal of the American Medical Informatics Association* 28, no. 1 (2020): 3–13.

Las GAN también demuestran la capacidad de manejar diferentes estructuras de datos que se encuentran comúnmente en los entornos empresariales. El desarrollo de modelos especializados como CTGAN y CTABGAN+ para datos tabulares estáticos, TimeGAN para datos de series temporales e IRG para datos relacionales pone de manifiesto la adaptabilidad de las GAN en diversos entornos de datos.

(F) Modelos lingüísticos

Desarrollados originalmente para tareas de tratamiento del lenguaje natural, los Transformers y los grandes modelos de lenguaje (LLM) también han demostrado ser eficaces en la síntesis de datos tabulares. Estos modelos utilizan el mecanismo de la atención para comprender relaciones complejas dentro de los datos, lo que los hace ideales para crear conjuntos de datos sintéticos que reflejen la complejidad del mundo real.

Los LLM también sobresalen cuando los datos originales son limitados. Aprovechar un amplio conocimiento preentrenado para llenar los vacíos en los datos originales dispersos y generar datos enriquecidos en entornos de datos escasos. Sin embargo, aunque los LLM ofrecen una capacidad notable en la síntesis de datos tabulares, requieren una potencia computacional sustancial y tiempo para entrenarse, lo que presenta una compensación. Otro beneficio significativo de los modelos generativos profundos es la capacidad de integrar el descenso de gradiente estocástico de privacidad diferencial (DP-SGD), que introduce ruido durante el entrenamiento para garantizar la privacidad de los datos con garantías demostrables. Sin embargo, es importante tener en cuenta que, si bien DP-SGD mejora la privacidad, también puede limitar la utilidad de los datos generados, lo que presenta un compromiso entre la protección de la privacidad y la utilidad de los datos.

Anexo D: Riesgos de reidentificación

Dado que los datos sintéticos generalmente intentan retener las propiedades estadísticas y las características de sus datos de origen, los adversarios pueden intentar volver a identificar o extraer información confidencial sobre un individuo a partir de los datos sintéticos. A continuación, se describen los diferentes tipos de ataques de reidentificación (comúnmente denominados ataques de privacidad) en conjuntos de datos sintéticos.

(A) Ataque singularización

El ataque de singularización generalmente se lleva a cabo para valores atípicos, por ejemplo, atributos únicos, atributos de datos raros o combinación única de atributos. A medida que los puntos de datos sintéticos generados intentan reflejar o capturar la presencia y las características de dichos valores atípicos, ofrecen una mayor posibilidad de seleccionar registros de datos únicos, y los valores atípicos son especialmente susceptibles. Si bien es posible que la selección no represente un riesgo de reidentificación por sí misma, puede permitir que el adversario obtenga información sobre el registro de datos mediante el uso de conjuntos de datos relacionados u otra información de fondo (consulte el ejemplo en ataque de vinculabilidad).

(B) Ataque de vinculabilidad

Para que ocurra un ataque de vinculabilidad, se supone que el adversario tiene acceso a dos conjuntos de datos, es decir, (i) datos sintéticos y (ii) otros datos disponibles públicamente o conjuntos de datos privados donde el adversario tiene acceso privilegiado. En un ataque de vinculación, el adversario intenta determinar si algún punto de datos de los dos conjuntos de datos pertenece al mismo individuo o grupo de individuos.

Por ejemplo, un adversario podría concluir que en el conjunto de datos sintéticos de pacientes en un hospital comunitario (mediante la señalización) hay una alta posibilidad de que exactamente un individuo que sea hombre, mayor de 80 años, tenga diabetes y gane un ingreso anual de $ 100,000 a $ 200,000. Un ataque exitoso ocurre cuando el adversario adivina correctamente que los datos sintéticos se entrenan a partir de un conjunto de datos que contiene el registro de datos de un paciente masculino con diabetes de 86 años, de la cuenta de redes sociales del paciente vinculado al hospital comunitario y el conocimiento privado de que no hay otro paciente diabético masculino mayor de 80 años en ese hospital comunitario. El adversario ahora tiene conocimiento adicional sobre este paciente, es decir, tiene un ingreso anual de $100,000 a $200,000.

La evaluación del ataque de vinculabilidad examina si la disponibilidad adicional de datos sintéticos mejora la capacidad de un adversario para formar vínculos entre conjuntos de datos diferentes. Intuitivamente, es probable que las posibilidades de éxito del adversario sean un ataque que mejore cuando aumente la utilidad de los datos sintéticos generados, es decir, cuanto más se parezcan a las características estadísticas de los datos de origen, mayores serán las posibilidades de que un ataque tenga éxito.

Por lo tanto, es importante que se incorporen prácticas de protección de datos durante el proceso de preparación de datos para la generación de datos sintéticos.

También es imperativo que cualquier evaluación de los riesgos de reidentificación diferencie entre las mejoras de la utilidad de los datos que son deseables, es decir, que se asemejen a las tendencias generales de la población que no traicionan la participación de un individuo en un conjunto de datos de origen, o indeseables, es decir, que dan lugar a un aumento de los riesgos de reidentificación de algunos individuos en el conjunto de datos de origen.

(C) Ataque de inferencia

Se supone que el adversario tiene acceso a un conjunto de atributos de datos comunes al conjunto de datos de origen y utiliza la información presente en los datos sintéticos para inferir atributos confidenciales (por ejemplo, otras complicaciones médicas) sobre las personas en el conjunto de datos de origen.

Por ejemplo, un ataque exitoso ocurre cuando un adversario puede inferir con alta confianza que un hombre de 86 años con diabetes (a partir del conjunto de datos de origen del hospital comunitario) tiene otras complicaciones médicas como hipertensión.

En un ataque de inferencia, estamos examinando si la disponibilidad adicional de datos sintéticos conduciría a una mayor probabilidad de inferencia exitosa con respecto a los atributos sensibles sobre los individuos en el conjunto de datos de origen. Al igual que antes, se podía esperar que cualquier dato sintético con suficiente utilidad mejorara la tasa de éxito de un adversario.

Es importante destacar que esta observación puede aplicarse a cualquier persona que pertenezca a la misma distribución (por ejemplo, hombres mayores de 80 años con diabetes), incluso cuando sus datos nunca se han utilizado para el entrenamiento.

Por lo tanto, una evaluación de ataques de inferencia debe medir el riesgo de reidentificación y, a continuación, comparar la incidencia de ataques exitosos con alguna línea de base establecida, por ejemplo, personas fuera del conjunto de datos de origen. En tal escenario, se está midiendo si la probabilidad de inferir con éxito los datos de alguien en el conjunto de datos de origen es mayor o menor que la de inferir datos de alguien que no está en el conjunto de datos de origen, a fin de aislar y cuantificar la fuga de privacidad a los individuos además de las tendencias poblacionales identificadas.

Anexo E: Ejemplos de enfoques para evaluar los riesgos de reidentificación

Este anexo presenta diferentes enfoques para evaluar los riesgos de reidentificación / privacidad adoptados por tres miembros de la industria. Estos enfoques se pueden aplicar a los datos sintéticos, independientemente del método de generación utilizado.

(A) Enfoque 1

Contribución del Dr. Khaled El Emam, Universidad de Ottawa

Divulgación de atribución. Se trata de una extensión de la noción tradicional de divulgación de identidad a los datos sintéticos. Considera la similitud de un registro sintético con un registro real, condicionada al riesgo de divulgación de identidad del conjunto de datos original (real).[41] Conceptualmente, evalúa hasta qué punto un adversario aprendería algo nuevo sobre un individuo al encontrar un registro que se parezca a él (es decir, que tenga los mismos valores en los identificadores indirectos) en los datos sintéticos. La divulgación de la atribución puede interpretarse como una probabilidad y a menudo se utiliza un valor aceptable de 0,09 (dentro del rango definido en la norma ISO/IEC 27599).

Para el cálculo de la divulgación de la atribución, el siguiente artículo describe el proceso en profundidad: https://www.jmir.org/2020/11/e23139/

Divulgación de pertenencia. Esto evalúa el grado en que un adversario se enteraría de que un individuo de la misma población que los datos reales fue incluido en el conjunto de datos de entrenamiento para el modelo generativo.[42] El conocimiento de que alguien está en el conjunto de datos de entrenamiento puede revelar algo sobre el individuo objetivo si el conjunto de datos de entrenamiento tiene características definitorias (por ejemplo, todos eran personas con una enfermedad en particular). Esto se puede definir como una puntuación F1 relativa que mide la precisión en la determinación de la pertenencia corregida por una determinación ingenua, con un valor típico de F1 relativo = 0,2 utilizado como umbral.

Para la divulgación de la membresía, el siguiente artículo describe los detalles del cálculo:

https://academic.oup.com/jamiaopen/article/5/4/ooac083/6758492?searchresult=1

Referencias

Emam, Khaled El, Lucy Mosquera, and J. Bass. “Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation.” *Journal of Medical Internet Research* 22, no. 11 (2020): e23139.

[41] Khaled El Emam, Lucy Mosquera, and J. Bass, “Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation,” Journal of Medical Internet Research 22, no. 11 (2020): e23139.

[42] Khaled El Emam, Lucy Mosquera, and Xi Fang, “Validating A Membership Disclosure Metric For Synthetic Health Data,” Journal of the American Medical Informatics Association 5, no. 4 (2022): 00ac083.

Emam, Khaled El, Lucy Mosquera, and Xi Fang. “Validating A Membership Disclosure Metric For Synthetic Health Data.” *Journal of the American Medical Informatics Association* 5, no. 4 (2022): 00ac083.

Emam, Khaled El, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. “Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study.” *JMIR Medical Informatics* 10, no. 4 (2022).

Kababji, Samer El, Nicholas Mitsakakis, Xi Fang, Ana-Alicia Beltran-Bless, Greg Pond, Lisa Vandermeer, Dhenuka Radhakrishnan, and Khaled El Emam. “Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets.” *JCO Clinical Cancer Informatics* 7 (2023). https://ascopubs.org/doi/full/10.1200/CCI.23.00116

Yang, S. “Process Mining the Trauma Resuscitation Patient Cohorts.” In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 29–35, 2018.

(B) Enfoque 2

Contribución de A*Star

En el Enfoque 2, el marco de evaluación y evaluación para cuantificar el riesgo de privacidad en datos sintéticos se detalla en https://github.com/statice/anonymeter, en el que el conjunto de métodos propuesto generará tres puntuaciones de umbral de riesgo de privacidad basadas en evaluaciones basadas en ataques para los tres riesgos principales, principalmente (i) ataque de singularización, (ii) ataque de vinculabilidad y (iii) ataque de inferencia.

En comparación con las alternativas basadas en ML, el marco proporciona un método computacionalmente eficiente y estadísticamente robusto para medir los riesgos de privacidad. Como primer paso, dividimos el conjunto de datos de origen en dos subconjuntos disjuntos, a saber, (i) control (20-30%) y (ii) entrenamiento (70-80%), con el primero desempeñando el papel de un pseudo conjunto de individuos públicos y el segundo como conjunto de datos de entrenamiento. Claramente, el conjunto de datos de control nunca se puede usar para el entrenamiento, sino que sirve como una línea de base de ataque a la privacidad que mide la capacidad del conjunto de datos sintético para "infringir" la privacidad de las personas que nunca ha utilizado. En otras palabras, el conjunto de datos de control representa la población que no ha contribuido al proceso de generación de datos sintéticos, y el conocimiento exitoso obtenido al atacar el conjunto de datos de control es, hasta cierto punto, alguna medida de la "utilidad" del conjunto de datos sintéticos.

Con ese fin, se realizaron dos ataques separados, a saber: (i) ataque de control y (ii) ataque principal. El ataque de control se dirige al conjunto de datos de control y mide patrones comunes a toda la población; El ataque principal se dirige al conjunto de datos de entrenamiento y mide los patrones comunes a toda la población y los posibles sesgos hacia el conjunto de datos de entrenamiento. La asimetría calculada entre los dos ataques proporciona una medida justa de la eficacia de los datos sintéticos para diferenciar a los individuos en el conjunto de datos de entrenamiento de la población más grande, al tiempo que basa la métrica de riesgo de privacidad obtenida con una línea de base razonable a partir de la cual hacer más interpretaciones.

Por último, el marco también mide una línea de base "ingenua" que asume que no hay ningún conocimiento previo del conjunto de datos sintético y, por lo tanto, depende completamente de la suerte. Esto cierra una laguna en la que se podría suponer erróneamente que el conjunto de datos sintéticos generado está libre de riesgos porque tiene una fidelidad/utilidad extremadamente pobre y/o cuando los ataques de inferencia/vinculabilidad diseñados o los datos sintéticos son insensibles en primer lugar. En estos escenarios, el ataque "ingenuo" podría superar a los otros dos ataques, lo que indica que la prueba es defectuosa.

La asimetría calculada entre el ataque principal y el ataque de control se normaliza para obtener una métrica de fuga de riesgo de privacidad, conocida como "R". Este valor está delimitado entre 0 y 1 y aumenta con los riesgos de fuga de privacidad. Es razonable decidir primero un valor umbral aceptable de "R" antes de generar los datos sintéticos; La inversión de este proceso expone a uno a una considerable libertad para justificar su producto. Dicho umbral puede fijarse en función de la política y mitigarse aún más en función de la sensibilidad del conjunto de datos de entrenamiento y la disponibilidad del conjunto de datos sintético generado.

Es crucial tener en cuenta que los riesgos de privacidad se evalúan con respecto a las personas en la base de datos de capacitación, y no al público en general. Como tal, la privacidad se ve comprometida cuando a un adversario le resulta más fácil (i) determinar si un individuo pertenece a la base de datos de entrenamiento y (ii) derivar detalles de un individuo de la base de datos de entrenamiento que de otro modo no se revelaría.

Referencias

Para obtener más detalles, se puede encontrar una descripción del marco y los algoritmos de ataque en el artículo de M. Giomi et al. "A Unified Framework for Quantifying Privacy Risk in Synthetic Data". *In Proceedings on Privacy Enhancing Technologies* Symposium (PETS 2023), 2023.

(C) Enfoque 3

Contribución de Betterdata.ai

El enfoque 3 audita la integridad de la privacidad de la canalización de generación de datos sintéticos (SDG) de extremo a extremo y no solo de los datos sintéticos generados. Este enfoque se basa en la Privacidad Diferencial (DP) [43] , que proporciona una cuantificación matemática de la privacidad individual. DP cuantifica el riesgo de que alguien deduzca que se incluyeron datos personales específicos en el conjunto de datos de entrenamiento mediante el análisis de los datos sintéticos. La pérdida de privacidad se calcula utilizando dos parámetros $\varepsilon>0$ y $0\leq\delta\leq1$, donde ε representa la pérdida de privacidad máxima permitida y δ representa la tolerancia aceptable de que se exceda esta pérdida de privacidad, que generalmente se mantiene cerca de cero.

[43] Cynthia Dwork et al., “Calibrating Noise to Sensitivity in Private Data Analysis,” in *Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, Vol 3876*, ed. S. Halevi and T. Rabin (Berlin: Springer, 2006); Cynthia Dwork and Aaron Roth, “The Algorithmic Foundations of Differential Privacy,” *Foundations and Trends® in Theoretical Computer Science* 9, no. 3–4 (2014): 211–407

El proceso de auditoría comienza con la identificación de los valores atípicos más confidenciales en el conjunto de datos de origen. Estos valores atípicos tienen un 50% de posibilidades de ser excluidos aleatoriamente de los datos utilizados en la canalización de los ODS. El siguiente paso consiste en realizar ataques de inferencia de pertenencia a los datos sintéticos para intentar identificar estos valores atípicos. El éxito de estos ataques, en particular el hecho de que supere una línea de base del 50% (suposición aleatoria), indica una posible fuga de privacidad. Esto establece el límite inferior para la pérdida de privacidad real. Utilizando el galardonado análisis de auditoría de privacidad desarrollado por Steinke, Nasr y Jagielski [44], convertimos las estadísticas de ataque a la membresía en límites inferiores de alta confianza en el presupuesto de privacidad ε para el δ de tolerancia.

Para obtener metodologías detalladas sobre cómo llevar a cabo ataques de inferencia de membresía, consulte el marco TAPAS[45][42].

Las organizaciones pueden adaptar su presupuesto de privacidad, ε, a sus requisitos específicos, reflejando la sensibilidad de sus datos y alineándose con los puntos de referencia de la industria. A continuación, se muestran ejemplos de presupuestos de privacidad informados públicamente que se utilizan en aplicaciones del mundo real:

Nombre de la organización	Tipo de datos	DP Presupuesto (ε)	Período de recopilación	Propósito de la recopilación de datos
Apple [5,6]	Datos de salud Safari Emoji QuickType	2.0 4.0 4.0 8.0	2017-2024	Análisis
Datos del censo de EE. UU. 2020 [7,8]	Datos de unidades de vivienda Archivo de la persona	2.47 17.14	2020	Decidir la distribución de fondos, ayudar a los estados

Para obtener más detalles, consulte Betterdata.ai URL en How_it_works (betterdata.ai).

Referencias

Abowd, John M. “The US Census Bureau Adopts Differential Privacy.” In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.

[44] Thomas Steinke, Milad Nasr, and Matthew Jagielski, “Privacy Auditing with One (1) Training Run,” in *NIPS ’23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, ed. A. Oh, T. Naumann, and A. Globerson (Curran Associates Inc., 2023), 49268–80, https://dl.acm.org/doi/10.5555/3666122.3668265

[45] TAPAS, “Welcome to TAPAS’s Documentation!,” tapas, 2022, https://tapas-privacy.readthedocs.io/en/latest/index.html

Apple Privacy Team. “Differential Privacy.” Apple.com, n.d. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Dwork, Cynthia, F. McSherry, K. Nissim, and A. Smith. “Calibrating Noise to Sensitivity in Private Data Analysis.” In Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, Vol 3876, edited by S. Halevi and T. Rabin. Berlin: Springer, 2006.

Dwork, Cynthia, and Aaron Roth. “The Algorithmic Foundations of Differential Privacy.” Foundations and Trends® in Theoretical Computer Science 9, no. 3–4 (2014): 211–407.

Houssiau, F, J Jordon, SN Cohen, O Daniel, A Elliot, J Geddes, C Mole, and C Rangel-Smith. “Tapas: A Toolbox for Adversarial Privacy Auditing of Synthetic Data.” arXiv Preprint ArXiv:2211.06550, 2022.

National Conference of State Legislatures. “Differential Privacy for Census Data Explained.” National Conference of State Legislatures, 2021. https://www.ncsl.org/technology-and-communication/differential-privacy-for-census-data-explained

Steinke, Thomas, Milad Nasr, and Matthew Jagielski. “Privacy Auditing with One (1) Training Run.” In NIPS ’23: Proceedings of the 37th International Conference on Neural Information Processing Systems, edited by A. Oh, T. Naumann, and A. Globerson, 49268–80. Curran Associates Inc., 2023. https://dl.acm.org/doi/10.5555/3666122.3668265

Tang, Jun, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. “Privacy Loss in Apple’s Implementation of Differential Privacy on MacOS 10.12.” *ArXiv:1709.02753*, 2017. https://arxiv.org/abs/1709.02753

TAPAS. “Welcome to TAPAS’s Documentation!” tapas, 2022. https://tapas-privacy.readthedocs.io/en/latest/index.html

AGRADECIMIENTOS

La PDPC y la Autoridad de Desarrollo de Medios de Comunicación de Infocomm (IMDA) expresan sinceramente su agradecimiento por las contribuciones editoriales en el desarrollo de esta publicación de las siguientes fuentes:

- Betterdata.ai
- Arquitectura y Repositorio de Datos Biomédicos (DAR), BMRC, A*STAR
- Dr. Khaled El Emam, Universidad de Ottawa
- Escuela de Computación, Universidad Nacional de Singapur (NUS)

PDPC e IMDA también expresan su agradecimiento y reconocimiento por todos los valiosos comentarios recibidos de las siguientes organizaciones:

- Centro de Capacidad de Protección y Privacidad de Datos, Agencia Gubernamental de Tecnología (GovTech)
- Johnson & Johnson (J&J)
- Mastercard
- Ministerio de Salud (MOH)
- Universidad Tecnológica de Nanyang, Singapur (NTU Singapore)

En esta guía se hace referencia a las siguientes guías/artículos:

Agencia Española Protección Datos. "Datos sintéticos y protección de datos". Blog, 2023. https://www.aepd.es/prensa-y-comunicacion/blog/datos-sinteticos-y-proteccion-de-datos

Financial Conduct Authority (U.K.). “Exploring Synthetic Data Validation – Privacy, Utility and Fidelity.” Publications, 2023. https://www.fca.org.uk/publications/research-articles/exploring-synthetic-data-validation-privacy-utility-fidelity

Giomi, Matteo, Franziska Boenisch, Christoph Wehmeyer, and Borbala Tasnadi. “A Unified Framework for Quantifying Privacy Risk in Synthetic Data.” In *Proceedings on Privacy Enhancing Technologies Symposium Issue 2*, 312–28, 2023. https://petsymposium.org/popets/2023/popets-2023-0055.php

Information Commissioner’s Office (U.K.). “Chapter 5: Privacy-Enhancing Technologies (PETs).” ICO call for views: Anonymisation, pseudonymisation and privacy enhancing technologies guidance, 2022.

———. “G7 DPAs’ Emerging Technologies Working Group Use Case Study on Privacy Enhancing Technologies.” UK GDPR guidance and resources, n.d. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/case-studies/g7-dpas-emerging-technologies-working-group-use-case-study-on-privacy-enhancing-technologies/

Jordon, James, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. “Synthetic Data -- What, Why and How?” ArXiv:2205.03257, 2022. https://arxiv.org/abs/2205.03257

FIN DEL DOCUMENTO

ELABORADO CONJUNTAMENTE POR