DATOS SINTÉTICOS: ¿QUÉ SON Y PARA QUÉ SE USAN?

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Oficina del Dato
Iniciativa aporta
datos.gob.es

Contenido

Contenido elaborado por Alejandro Alija,

Experto en transformación digital y datos abiertos.

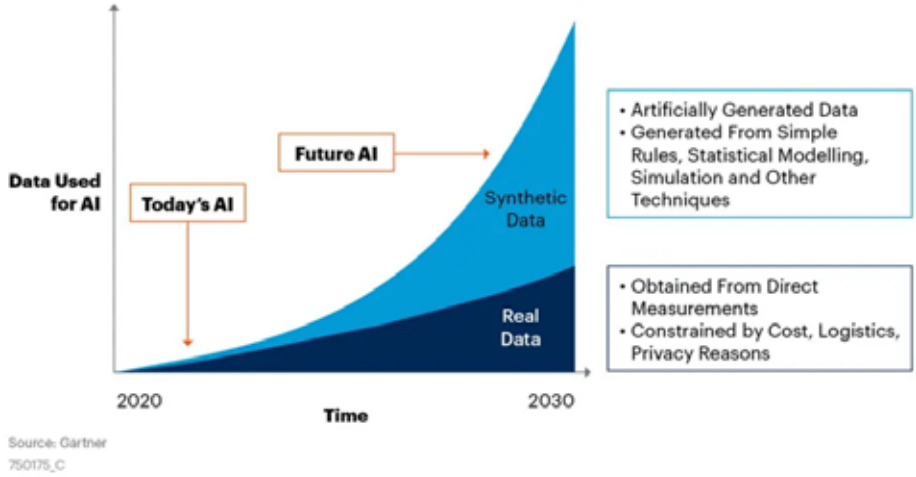
INTRODUCCIÓN

Pese a vivir en la era de los datos, en contra de lo que pueda parecer, **uno de los principales problemas al que nos enfrentamos para construir nuevos productos y servicios digitales es la dificultad para acceder a** de datos de valor en el contexto de ese nuevo producto o servicio. Sí, por extraño que parezca, son muchas las ocasiones en las que no es fácil acceder a datos de calidad para entender un proceso o un sistema desde la óptica de los datos. En ocasiones, los datos existen pero se encuentran fragmentados hasta tal punto que no es posible obtener una vista clara de lo que ocurre en una organización, un producto o un proceso. En otros casos, sabemos que **los datos existen**, pero son prácticamente **inaccesibles por razones de seguridad o privacidad**. Esto es lo que ocurre habitualmente con **los datos personales más sensibles como los datos médicos o bancarios**. Desde otro ángulo, a veces, es fácil encontrar e incluso acceder a los datos reales en sus respectivas fuentes, pero podría llegar a ser económicamente inviable o muy caro utilizarlos. Por ejemplo, podemos tener acceso a grandes bases de datos de recursos gráficos como imágenes o vídeos con los que entrenar algoritmos de Inteligencia artificial (IA), sin embargo, es necesario pagar costes de acceso y disponer de recursos de computación costosos para importarlos a nuestro sistema o entorno de entrenamiento (extraerlos de sus fuentes, almacenarlos, catalogarlos, etc.) . Por estas y otras muchas razones, **existen desde hace mucho tiempo los denominados datos sintéticos**. Tanto es así que Gartner, en un artículo publicado en el Wall Street Journal en 2021, aseguraba que en 2024 el 60% de los datos utilizados para el desarrollo de proyectos de inteligencia artificial y análisis se generarían sintéticamente.

En este informe, tratamos de **profundizar en el campo de los datos sintéticos, explicando qué son y en qué tipos de situaciones o casos de uso son útiles y necesarios**. Analizaremos **sus beneficios** con respecto a los "datos reales" y **explicaremos mediante un ejemplo práctico cómo se generan y un posible uso de estos**. ¡Comenzamos!

1. ¿QUÉ SON LOS DATOS SINTÉTICOS? Y ¿PARA QUÉ SON ÚTILES?

En la era digital en la que vivimos, la generación y el análisis de datos se han convertido en elementos fundamentales para el **desarrollo de nuevas tecnologías, productos y servicios**. Uno de los factores decisivos por los que la vacuna del COVID-19 se pudo conseguir en tan corto espacio de tiempo, tuvo que ver con al análisis de datos masivos procedentes de ensayos clínicos, datos epidemiológicos, modelos matemáticos de simulación, etc. Ello es algo que no había sido posible en ningún otro momento anterior de la historia. Este es solo un ejemplo, porque, hoy en día, cualquier tipo de decisión importante, en casi todas las organizaciones, se toma en base a los datos disponibles. Sin embargo, como comentamos en la introducción, en muchos casos, **acceder a datos reales puede ser un desafío debido a las restricciones de privacidad, confidencialidad o simplemente a la falta de disponibilidad de información completa y actualizada.**



Fuente original: Gartner. Extraída de "Synthetic Data Is About To Transform Artificial Intelligence" de Forbes Gartner

En este contexto, los datos sintéticos han surgido como una solución prometedora. **Los datos sintéticos, a diferencia de los datos reales, son información fabricada artificialmente, en lugar de aquella generada por eventos del mundo real**. Los datos sintéticos, entre otros usos, **se diseñan para imitar las características y distribuciones de los datos reales, sin contener información personal o sensible** que pueda identificar a individuos o comprometer su privacidad. Estos datos **se crean mediante algoritmos y técnicas de generación** que **preservan la estructura, las relaciones y las propiedades estadísticas** de los datos originales, **brindando una alternativa segura y confiable** para el **análisis**, la **experimentación** y el **entrenamiento de modelos** de inteligencia artificial. Los datos sintéticos no son una idea nueva . La novedad es que ahora se acercan a un punto de inflexión crítico en términos de impacto en el mundo real. **Está a punto de cambiar toda la cadena de valor**, y el conjunto de tecnologías de inteligencia artificial, todo un proceso que tendrá inmensas implicaciones económicas. No hay más que prestar atención al torrente de acontecimientos relacionados con las tecnologías de IA generativas de los últimos meses.

Los datos sintéticos son información fabricada artificialmente en lugar de aquella generada por eventos del mundo real. Los datos sintéticos **se diseñan para imitar las características y distribuciones de los datos reales, sin contener información personal o sensible** que pueda identificar a individuos o comprometer su privacidad. Estos datos **se crean mediante algoritmos y técnicas de generación** que **preservan la estructura, las relaciones y las propiedades estadísticas** de los datos originales, **brindando una alternativa segura y confiable** para el análisis, la experimentación y el entrenamiento de modelos de Inteligencia Artificial.

[1] Muy recomendable esta lectura de un artículo de Forbes en el que se explica cómo el desarrollo de los vehículos autónomos contribuyó en gran medida al desarrollo de los datos sintéticos.

2. ¿PARA QUÉ SON ÚTILES LOS DATOS SINTÉTICOS?

Los datos sintéticos tienen múltiples aplicaciones y resultan especialmente útiles en situaciones en las que la disponibilidad de datos reales es limitada; su uso requiere **proteger la privacidad** de las personas involucradas. A continuación, ilustramos con tres ejemplos o casos de uso concretos, **las potenciales aplicaciones** de los datos sintéticos:

INVESTIGACIÓN CIENTÍFICA

Los datos sintéticos permiten a los investigadores explorar y desarrollar nuevos enfoques, modelos y algoritmos sin la necesidad de acceder a datos reales sensibles. Esto **acelera la investigación y fomenta la colaboración**, al tiempo que **mantiene la integridad y privacidad** de los participantes en los estudios. A modo de ejemplo, **los datos genómicos** son uno de los tipos de datos más complejos, multidimensionales y ricos en información del mundo. Con poco más de 3 mil millones de pares de bases de longitud, la secuencia de ADN única de cada ser humano define en gran medida quiénes somos, desde nuestra altura hasta el color de nuestros ojos, pasando por el riesgo que tenemos de contraer una enfermedad cardíaca o de abusar de ciertas sustancias. Si bien no es un lenguaje natural, **las secuencias genómicas son datos textuales**; la secuencia de ADN de cada individuo se puede codificar a través de un simple "alfabeto" de 4 letras. El análisis del genoma humano con IA de vanguardia permite a los investigadores desarrollar una comprensión más profunda de las enfermedades, la salud y cómo funciona la vida misma. Pero esta investigación se ha visto limitada por la **disponibilidad muy escasa de datos genómicos**. Las estrictas regulaciones de privacidad y las restricciones de intercambio de datos en torno a los datos genéticos humanos impiden la capacidad de los investigadores para trabajar con conjuntos de datos genómicos a escala. Los datos sintéticos ofrecen una solución potencialmente revolucionaria: pueden **replicar las características y los patrones de los conjuntos de datos genómicos reales** mientras eluden las preocupaciones sobre la privacidad de los datos, ya que se generan artificialmente y no corresponden a ningún individuo en particular en el mundo real.

PRUEBAS DE SOFTWARE Y TEST DE SISTEMAS

Los datos sintéticos **se utilizan ampliamente para probar y validar software y sistemas informáticos**. Al generar conjuntos de datos realistas, pero sintéticos, los desarrolladores pueden simular escenarios diversos y evaluar el rendimiento, la escalabilidad y la seguridad de sus aplicaciones sin exponer datos reales ni correr riesgos innecesarios. Por ejemplo, **en el desarrollo y testeo de un sistema de control de calidad por visión artificial** (en una línea de producción de bienes de equipo), es más fácil generar artificialmente 100.000 imágenes de, digamos, teléfonos inteligentes, que tener que recopilar esas imágenes en el mundo real una por una. Para obtener datos reales de esta aplicación, necesitamos tiempo de proceso, sofisticados sistemas de visión artificial y dedicar una cantidad nada despreciable de horas de trabajo para poner a punto el sistema. Con los datos sintéticos **podemos generar imágenes artificialmente** que nos permitan reproducir características o defectos en los bienes producidos.

ENTRENAMIENTO DE MODELOS DE INTELIGENCIA ARTIFICIAL

Los datos sintéticos son esenciales en el entrenamiento y la mejora de modelos de aprendizaje automático (machine learning). Por ejemplo, **recopilar datos de conducción del mundo real para cada escenario que un vehículo autónomo pudiera encontrar en la carretera sería, simplemente, imposible**. Dado lo impredecible e ilimitado que es el mundo, se necesitan, literalmente, cientos de años de conducción en el mundo real para recopilar todos los datos necesarios para construir un vehículo autónomo verdaderamente seguro. En este contexto, las compañías dedicadas al desarrollo de sistemas de conducción autónomos, crearon motores de simulación sofisticados para generar sintéticamente el volumen de datos requerido y exponer de manera eficiente sus sistemas de IA a los diferentes escenarios de conducción. Un ejemplo para ilustrar este uso puede ser la actividad empresarial que lleva a cabo Waabi, una compañía dedicada a crear datos sintéticos de simulación sobre escenarios de conducción autónoma.

En esta sección hemos destacado sólo tres ejemplos concretos y representativos para ilustrar la importancia de los datos sintéticos. Sin embargo, existen multitud de casos de aplicación donde son esenciales los datos sintéticos, como la detección de fraudes, medicina preventiva, la evaluación de riesgos crediticios, entrenamiento de modelos de lenguaje, gestión de siniestros en el campo de los seguros, etc.

3. FORMAS DE GENERAR DATOS SINTÉTICOS

Los datos sintéticos se generan utilizando técnicas y algoritmos diseñados para **imitar las características y distribuciones de los datos reales**, sin contener información personal o sensible. A continuación, se presentan algunos enfoques comunes utilizados para generar datos sintéticos.

¿Cómo se generan datos sintéticos?

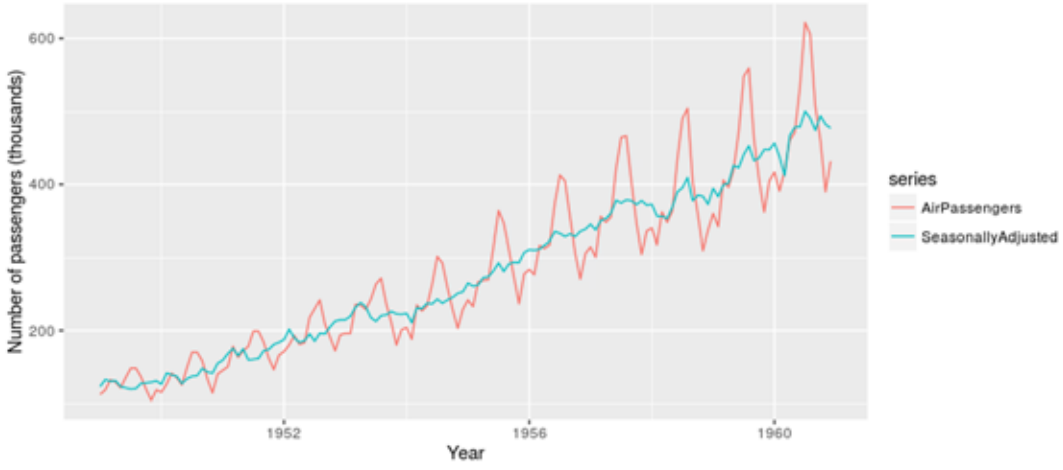
Técnicas de remuestreo: Reorganizar conjuntos de datos existentes para obtener nuevos.

Modelado probabilístico y generativo: Crear conjuntos de datos a partir de probabilidades obtenidas de conjuntos reales.

Métodos de perturbación y enmascaramiento: Modificar conjuntos de datos reales para proteger la privacidad conservando algunas características.

TÉCNICAS DE REMUESTREO

Mediante esta técnica tratamos de extraer parte de los datos originales y reorganizarlos aleatoriamente para **crear nuevos conjuntos de datos**. Esta selección parcial y aleatoria de números de una distribución es un método común para crear datos sintéticos. Sin embargo, aunque este método no captura toda la información de los datos del mundo real, puede producir una distribución de datos que se parece mucho a los datos reales. Una muestra de esta técnica, aunque con un propósito diferente, es la desestacionalización de datos. Por ejemplo, para obtener datos de desempleo que se encuentran altamente correlacionados con las épocas particulares del año, el remuestreo se emplea para extraer la tendencia subyacente (línea verde) a los datos estacionales (línea roja).



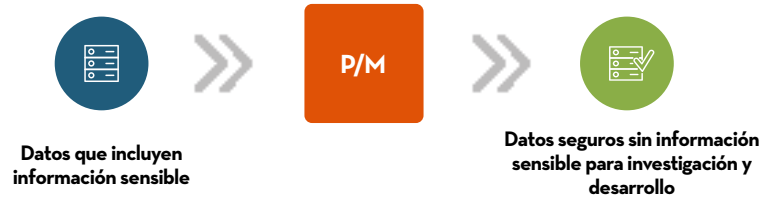
Fuente: StackOverflow

MODELADO PROBABILÍSTICO Y GENERATIVO

En este enfoque se construyen modelos probabilísticos basados en las distribuciones y relaciones observadas en los datos reales. Estos modelos pueden ser estadísticos o de aprendizaje automático. Los modelos se entrenan utilizando datos reales, y luego se utilizan para generar nuevos datos sintéticos. Algunos ejemplos de modelos generativos incluyen redes antagónicas generativas (GANs) y los denominados autoencoders.

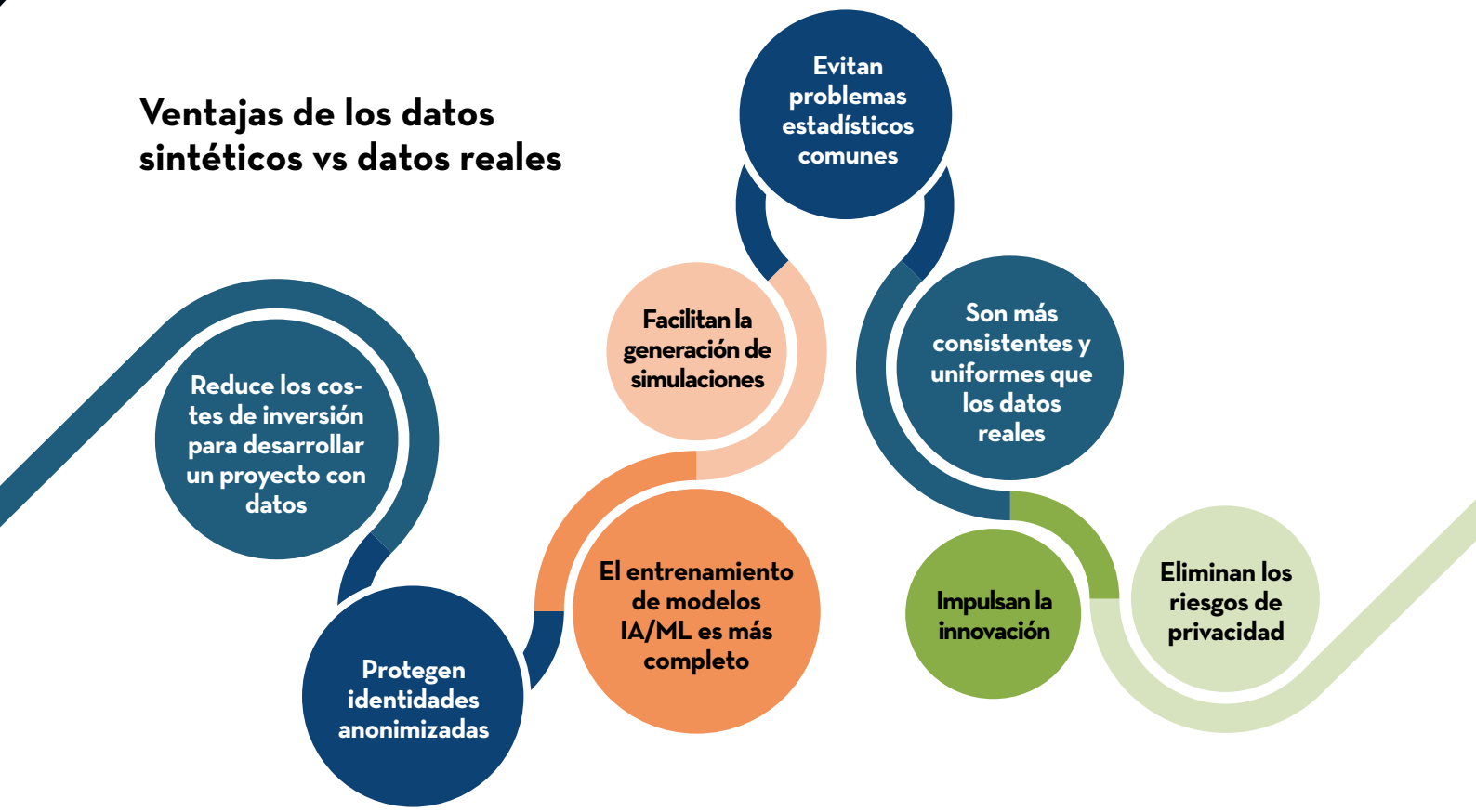
MÉTODOS DE PERTURBACIÓN Y ENMASCARAMIENTO

En este enfoque los datos reales se modifican y perturban de manera controlada para proteger la privacidad mientras se conservan ciertas características importantes. Tal y como vemos en el esquema, esto puede implicar enmascarar información sensible, como reemplazar nombres reales por nombres ficticios o distorsionar valores numéricos para evitar la identificación directa.



4. BENEFICIOS DE USAR DATOS SINTÉTICOS

Ventajas de los datos sintéticos vs datos reales



A lo largo del informe, hemos apuntado ya algunas de las principales motivaciones para usar datos sintéticos. Profundicemos ahora en los principales beneficios de los datos sintéticos frente a los reales:

1 **Superar restricciones regulatorias**: los datos sintéticos evitan las restricciones regulatorias de los datos reales. Pueden replicar todas las características importantes de los datos reales sin exponer la información real, lo que elimina las obligaciones sobre las regulaciones de privacidad.

2 **Preservación de la privacidad.** Los datos sintéticos resuelven el dilema entre privacidad y utilidad porque no es necesario proteger los datos sintéticos contra ataques o fugas porque no son datos reales. Por lo tanto, emplear datos sintéticos es una decisión que permite utilizar un conjunto de datos útil sin poner en riesgo información vulnerable y manteniendo la privacidad de los dato.

3 **Resistencia a la reidentificación:** un estudio de 2016 demostró que, después de solo 15 minutos de registrar los patrones de frenado de un conductor, los investigadores pudieron identificar a ese conductor con una precisión del 87%. Resulta que la forma en que se presiona el pedal del freno es casi completamente única de un individuo. Existen técnicas que permiten re-identificar a personas incluso con datos reales anonimizados. Sin embargo, los datos puramente sintéticos no contienen información real y, por lo tanto, no se pueden identificar.

4 **Facilitar la innovación y monetización:** por lo general, los datos sintéticos son rápidos y sencillos de generar (no en todos los casos, pero, por ejemplo, sí en el de los datos tabulares). Como los datos sintéticos no presentan problemas de privacidad, es posible compartir rápidamente estos conjuntos de datos con terceros para investigación e innovación, e incluso utilizarlos como una herramienta de monetización.

5 **Agilizar la simulación:** los datos sintéticos permiten generar datos que simulen condiciones que aún no han ocurrido en la vida real. Cuando los datos reales no están disponibles, los datos sintéticos son la única solución; sirva como ejemplo el caso comentado anteriormente sobre conjuntos de datos que representan escenarios posibles en situaciones de conducción autónoma. En este video se explica a la perfección una situación como ésta.

6 **Evitar problemas estadísticos:** los datos sintéticos son inmunes a algunos problemas estadísticos comunes, como la falta de respuesta a elementos, patrones de salto y otras restricciones lógicas. Al diseñar cuidadosamente las reglas para generar los datos sintéticos, se pueden evitar problemas estadísticos comunes.

7 **Lograr una mayor consistencia:** los datos sintéticos tienden a ser más uniformes y consistentes que los datos reales, lo que los hace más adecuados para realizar análisis precisos. Por el contrario, es verdad que, algunos datos sintéticos son de baja fidelidad, y pueden no contener "outliers" o huecos que tanto caracterizan a los datos procedentes del mundo real.

8 **Facilitar el entrenamiento de modelos y permitir una manipulación sencilla:** los datos sintéticos pueden enriquecer y complementar a datos reales para ayudar a entrenar modelos de IA/ML, especialmente cuando no existen suficientes datos reales por motivos de privacidad, regulación y/o falta de acceso o de tiempo necesario para capturar eventos del mundo real.

9 **Viabilizar proyectos incipientes y aumentar la rentabilidad:** cuando arrancamos un nuevo proyecto basado en datos, no hemos tenido tiempo para capturarlos o incluso podríamos no tener los recursos económicos para comprar conjuntos de datos reales de calidad. Con los datos sintéticos podríamos invertir una pequeña cantidad en extraer un patrón real para luego generar una cantidad mucho mayor de datos mediante sintetizadores de datos. A menores costes de inversión, la rentabilidad aumenta, además de proporcionar una estrategia viable para hacer crecer el proyecto, producto o servicio que estemos construyendo.

UN EJEMPLO PRÁCTICO

En este informe, hemos hablado de varios métodos de generación de datos sintéticos, cómo el remuestreo, los modelos generativos y el enmascaramiento de datos reales. Para ilustrar la generación de datos reales vamos a usar un software open source disponible aquí que procede de un desarrollo llevado a cabo en el entorno académico del MIT (Instituto Tecnológico de Masachussets). El proyecto se denomina SDV (Synthetic Data Vault) y es una librería de Python diseñada para ser una herramienta completa de creación de datos sintéticos tabulares. El SDV utiliza una variedad de algoritmos de aprendizaje automático para aprender patrones de sus datos reales y emularlos en datos sintéticos. Por lo tanto, pertenece a la clase de **métodos generativos para datos sintéticos**. La versión comercial del producto se distribuye a través de una compañía de reciente creación denominada DataCebo. Un detalle muy interesante de este proyecto es que cuenta con varios tutoriales prácticos que se pueden ejecutar en Google Colab.

A continuación, desgranamos las principales claves con un ejemplo práctico de generación de datos sintéticos sobre clientes (ficticios) alojados en un hotel (igualmente ficticio) utilizando Colab. Todos los materiales utilizados en este ejemplo práctico, incluyendo el cuaderno de Jupyter de este ejemplo para ejecutar en Google Colab, están disponibles en el repositorio de código de Github de datos.gob.es.

Si desea replicar el ejercicio, dejamos acceso a:

- Repositorio de GitHub:
  https://github.com/datosgobes/Synthetic-data/tree/main/Google_Colab
- Sección en Google Colab:
  https://colab.research.google.com/drive/1Uo2PbmVPO4_evlbCvqwUM1c7gy360sOv?usp=sharing

Comencemos.

El flujo de trabajo que vamos a seguir durante el ejemplo es similar al que ilustra la siguiente figura:

1 **Como punto de partida, utilizamos un conjunto de datos reales a partir de los cuales vamos a generar nuestros nuevos datos sintéticos** que mantendrán las propiedades y las distribuciones del conjunto original sin contener datos privados reales. Para generar datos sintéticos, además de un conjunto de datos reales, necesitamos los denominados metadatos. Los metadatos no son otra cosa que la descripción del conjunto de datos originales. Es decir, una caracterización de sus columnas (en este caso por tratarse de datos tabulares), describiendo en cada caso, qué tipo de dato es el que puebla esa columna (o campo). Por ejemplo, en el caso de que los datos originales contengan la edad de los clientes del hotel, este será un campo de tipo numérico, entero y con un rango entre 0-120.

2 Una vez que tengamos los datos reales y sus metadatos asociados, **creamos y entrenamos el sintetizador.**

3 El sintetizador es básicamente **el programa que creará datos ficticios a partir de los reales** utilizando la técnica de modelado probabilístico y generativo.

4 Puesto que los datos reales pueden contener datos sensibles y privados como nombres, direcciones, números de cuenta, etc., el **paquete SDV nos proporciona herramientas de anonimización de datos para detectar aquellos datos sensibles y eliminarlos, enmascararlos, etc.**

5 **Como paso final**, llega el turno de **comprobar la calidad de los datos sintéticos generados** y demostrar que, efectivamente, éstos mantienen las propiedades del conjunto real pero siendo ahora completamente inventados.

Detallando punto por punto, en el paso 1 podemos ver los tipos de datos reales que constituyen nuestro conjunto de datos de partida:

index	guest_email	has_rewards	room_type	amenities_fee	checkin_date
0	michaelsanders@shaw.net	false	BASIC	37.89	27 Dec 2020
1	randy49@brown.biz	false	BASIC	24.37	30 Dec 2020
2	webermelissa@neal.com	true	DELUXE	0.0	17 Sep 2020
3	gsims@terry.com	false	BASIC	NaN	28 Dec 2020
4	misty33@smith.biz	false	BASIC	16.45	05 Apr 2020

Datos reales sobre clientes alojados en un hotel.

Asociados a éstos, tenemos los metadatos correspondientes:


```
"METADATA_SPEC_VERSION": "SINGLE_TABLE_V1",
"primary_key": "guest_email",
"columns": {
        "guest_email": {
                "sdtype": "email",
                "pii": true
        },
        "has_rewards": {
                "sdtype": "boolean"
        },
        "room_type": {
                "sdtype": "categorical"
        },
        "amenities_fee": {
                "sdtype": "numerical",
                "computer_representation": "Float"
        },
        "checkin_date": {
                "sdtype": "datetime",
                "datetime_format": "%d %b %Y"
        },
        "checkout_date": {
                "sdtype": "datetime",
                "datetime_format": "%d %b %Y"
        },
        "room_rate": {
                "sdtype": "numerical",
                "computer_representation": "Float"
        },
        "billing_address": {
                "sdtype": "address",
                "pii": true
        },
        "credit_card_number": {
                "sdtype": "credit_card_number",
                "pii": true
        }
}
```


En el paso número 2 lo que hacemos es proporcionar los datos reales y los metadatos al sintetizador, el programa encargado de crear nuevos datos a partir de un conjunto real. En el notebook de Colab encontramos los fragmentos de código que ejecutarán esta parte:

```
from sdv.lite import SingleTablePreset

Synthesizer = SingleTablePreset (
        metadata,
        name = 'FAST_ML'
)
```

Tras este 'entrenamiento', el sintetizador ya cuenta con referencias de datos reales, y está listo para 'producirlos'; es decir, podemos generar tantos datos sintéticos como deseemos a partir del patrón (datos reales y metadatos) proporcionado.

Por ejemplo, con la ejecución de código (paso 3) que presentamos a continuación, generamos un nuevo conjunto sintético de 500 registros:

```
Synthetic_data = synthesizer.sample (
        num_rows=500
)

Synthetic_data.head()
```

Y la última línea de código nos sirve para comprobar el resultado, al mostrarnos los primeros registros del conjunto de datos sintetizado:

	guest_email	has_rewards	room_type	amenities_fee	checkin_date	checkout_date	room_rate
0	dsullivan@example.net	False	BASIC	9.436498	20 Mar 2020	11 Apr 2020	141.635838
1	steven59@example.org	False	BASIC	20.158516	20 Jun 2020	11 Aug 2020	185.529627
2	brandon15@example.net	False	BASIC	22.907020	16 Apr 2020	11 Apr 2020	145.403493
3	humphreyjennifer@example.net	False	BASIC	25.121149	04 Jun 2020	17 Jun 2020	180.463870
4	joshuabrown@example.net	False	BASIC	21.185741	11 Nov 2019	25 Oct 2019	180.288810

En el conjunto de datos original, teníamos algunas columnas con datos, como el correo electrónico, la dirección de facturación y el número de teléfono del cliente (marcados por el metadato *"pii": true*). En los datos sintéticos confidenciales (**paso 4**), estas **columnas están totalmente anonimizadas**[2]: contienen valores completamente falsos que siguen el formato original. Este proceso ha sido posible gracias a las herramientas de anonimización del paquete SDV que permite detectar datos sensibles para eliminarlos o enmascararlos. Podemos ver el resultado con las siguientes líneas de código:

```
sensitive_column_names = [ 'guest_email', 'billing_address', 'credit_card_number']

real_dat[sensitive_column_names].head(3)
```

	guest_email	billing_address	credit_card_number
0	michaelsanders@shaw.net	49380 Rivers Street\nSpencerville, AK 68265	4075084747483975747
1	randy49@brown.biz	88394 Boyle Meadows\nConleyberg, TN 22063	180072822063468
2	webermelissa@neal.com	0323 Lisa Station Apt. 208\nPort Thomas, LA 82585	389834769713 80

Conjunto de datos original (real)

	guest_email	billing_address	credit_card_number
0	dsullivan@example.net	90469 Karla Knolls Apt. 781\nSusanberg, NC 28401	5161033759518983
1	steven59@example.org	1080 Ashley Creek Apt. 622\nWest Amy, NM 25058	4133047413145475690
2	brandon15@example.net	99923 Anderson Trace Suite 861\nNorth Haley, T...	4977328103788

Conjunto de datos sintético (ficticio)

Finalmente, en el **paso número 5** evaluamos la calidad de los datos sintéticos generados. Es importante **verificar que los datos generados mantienen las propiedades del conjunto** original, puesto que de ello va a depender que los sistemas, procesos y algoritmos que entrenemos posteriormente se comporten según lo esperado.

Para esto, el paquete SDV, nos proporciona una serie de **herramientas específicas de control de calidad** que facilitan el trabajo. Por ejemplo, el siguiente código genera un pequeño informe de calidad:

```
from sdv.evaluation.single_table import evaluate_quality

quality_report = evaluate_quality(
real_data,
synthetic_data,
metadata
)
```

Con el siguiente resultado:

```
Creating report: 100%|██████████| 4/4 [00:00<00:00, 27.02it/s]
Overall Quality Score: 89.12%

Properties:
Column Shapes: 90.27%
Column Pair Trends: 87.97%
```

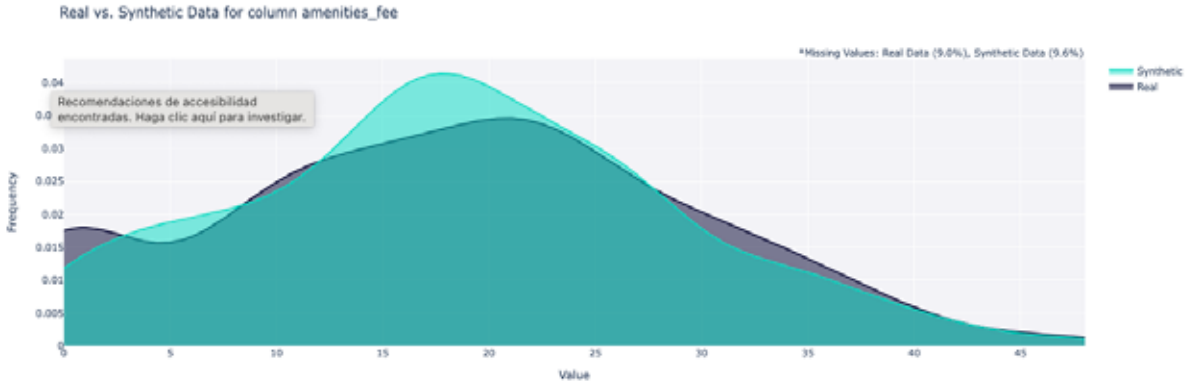
[2] *Véase la Guía de Anonimización de datos: Técnicas y casos prácticos: https://datos.gob.es/es/documentacion/introduccion-la-anonimizacion-de-datos-tecnicas-y-casos-practicos*

El informe de calidad evalúa la capacidad de los datos sintéticos generados para capturar las propiedades matemáticas de los datos reales. Esto también se conoce cómo **informe de fidelidad de datos sintéticos**. El informe ejecuta métricas seleccionadas, como por ejemplo, **la distribución matemática de los valores de una columna** o **la correlación entre una columna de datos sintéticos y su correspondiente columna con datos reales**.

Por ejemplo, la métrica, *Column Shapes*, evalúa la similitud entre la distribución matemática de una columna sintética frente a su correspondiente real. Por ejemplo, imaginemos que estamos generando datos sintéticos correspondientes a un conjunto de notas de unos estudiantes en un examen. Si suponemos que la distribución de notas en un examen sigue una distribución normal o campana de Gauss, una buena columna sintética debería de reproducir la misma distribución. Esto es lo que mide la métrica *Column Shapes*. Cuanto más alto sea el valor de esta métrica, más se parecen las distribuciones de ambas columnas. La métrica general es la media de las métricas individuales de cada par de columnas. En el caso de la métrica *Column Pair Trends* se evalúa la correlación matemática entre una columna sintética y su correspondiente original. Recordar que una correlación máxima de valor 1 significa que los puntos representados de una columna frente a la otra dibujan una línea recta perfecta de pendiente igual a 1. Lo mismo ocurre con una correlación negativa (anti correlación) pero con signo opuesto. Para más información sobre los informes de calidad de la herramienta se puede consultar la siguiente documentación (en inglés) https://docs.sdv.dev/sdmetrics/reports/quality-report/whats-included.

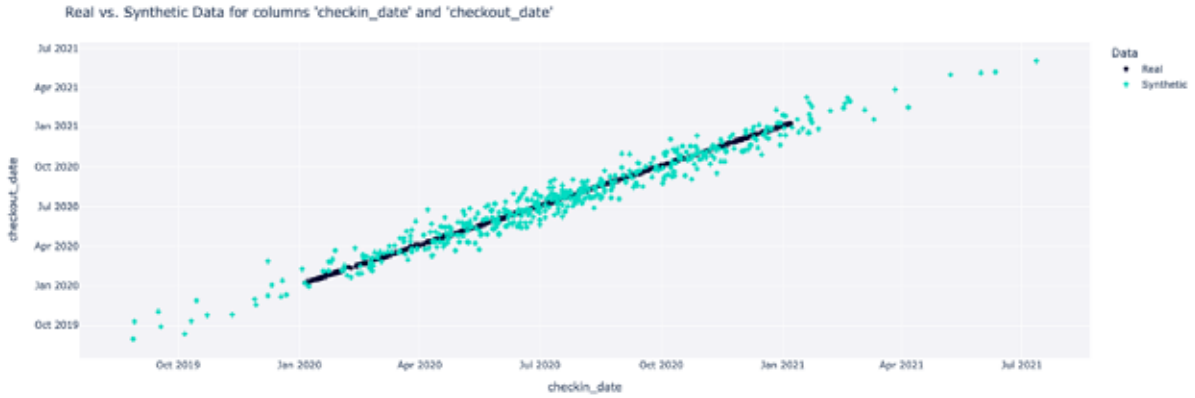
El informe de calidad también se puede consultar de forma más gráfica y visual. El siguiente código genera una visualización que nos permite representar visualmente la métrica Column Shapes (real vs sintético) para comprobar de forma visual que mantienen la misma distribución estadística.

```
fig = get_column_plot(
   real_data=real_data,
   synthetic_data=synthetic_
data,
   column_name='amenities_fee',
   metadata=metadata
)

fig.show()
```



De forma complementaria, podemos ver la correlación (recordar que se mide mediante la métrica *Column Pair Trend*) de determinadas variables, como por ejemplo, las fechas de entrada y salida del hotel de ambos conjuntos:

```
fig = get_column_pair_plot(
   real_data=real_data,
   synthetic_data=synthetic_data,
   column_names=['checkin_date', 'checkout_date'],
   metadata=metadata
)

fig.show()
```



En este ejemplo, hemos visto, probablemente, el ejercicio de generación de datos sintéticos más sencillo de ejecutar. Sin embargo, SDV pone a nuestra disposición otros tutoriales guiados en Google Colab, muy interesantes para seguir profundizando de una forma práctica en la generación de datos sintéticos tabulares.

- Podemos, por ejemplo, explorar la generación de datos sintéticos a partir de nuestros propios datos en formato csv.
- También tenemos la posibilidad de generar datos sintéticos a partir de bases de datos con varias tablas relacionadas.
- También podemos tratar de personalizar el sintetizador de datos con algunos parámetros y reglas de negocio como, por ejemplo, forzar que determinados datos cumplan un rango determinado (fechas en un periodo determinado)

CONCLUSIONES

Los datos sintéticos ofrecen una alternativa segura y asequible cuando los datos reales son escasos, inaccesibles o no se pueden utilizar debido a restricciones de privacidad. Este tipo de datos son generados artificialmente pero mantienen propiedades similares a los datos reales, lo que los hace útiles en determinadas variedades de casos de uso. Sin embargo, es importante **evaluar la calidad de los datos sintéticos** y garantizar que cumplan con los requisitos y objetivos específicos para evitar introducir sesgos o falta de realismo antes de su implementación.

Los datos sintéticos son hoy una realidad muy presente, especialmente desde la explosión a finales del 2022 y comienzos del 2023 de las tecnologías de IA generativa, los modelos del lenguaje masivos y los generadores de imágenes. Es probable que las predicciones de Gartner para 2024 (realizadas en 2021) se vean superadas con creces (60% de los datos serán sintéticos) y que nos enfrentemos a nuevos desafíos relacionados con los datos sintéticos, como por ejemplo, su falta de precisión, su dependencia de los conjuntos reales, el riesgo de no contemplar situaciones anómalas o excepcionales y la falta de transparencia hacia los consumidores por el hecho de no saber sí los datos son reales o ficticios. Es imposible pronosticar qué nos depara el futuro de los datos sintéticos, especialmente con el **desarrollo exponencial de las tecnologías de inteligencia artificial**, pero lo que está claro es que van a jugar un papel crucial para todos nosotros, consumidores, profesionales, académicos, estudiantes, etc. Nos vemos en el próximo informe sobre datos, sintéticos o reales, ¿quién sabe?.

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Oficina del Dato
Iniciativa aporta
datos.gob.es