

DATOS SINTÉTICOS

¿QUÉ SON?



Es información fabricada artificialmente, diseñada para imitar las características y distribuciones de los datos reales, sin contener referencias personales o sensibles.

¿POR QUÉ SON NECESARIOS?

Aunque existan cada vez más datos, en ocasiones:

- Es difícil acceder a datos de valor
- Están demasiado fragmentados
- Contienen información sensible
- Son insuficientes para las necesidades del proyecto



Los datos sintéticos son una alternativa.



¿PARA QUÉ SE USAN?

Se utilizan principalmente para el análisis, la experimentación y el entrenamiento de modelos de inteligencia artificial.

ALGUNOS EJEMPLOS DE CASOS DE USO DE DATOS SINTÉTICOS:



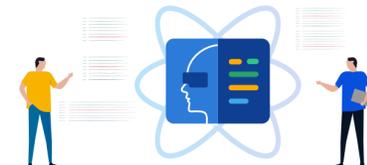
INVESTIGACIÓN CIENTÍFICA

Ej. Aceleración en la investigación del genoma humano, gracias a la generación de datos genómicos sintéticos.



PRUEBAS DE SOFTWARE Y TEST DE SISTEMAS

Ej. Avances relevantes en el desarrollo y testeo de sistemas de control de calidad por visión artificial, debido a la facilidad en la generación de imágenes.



ENTRENAMIENTO DE MODELOS DE IA

Ej. Progresos significativos en la construcción de vehículos autónomos, al hacer posible la creación de diversidad de escenarios para evaluar el rendimiento del software.

¿CÓMO SE CREAN?

Mediante algoritmos y técnicas de generación que preservan la estructura, relaciones y propiedades estadísticas de los datos originales.

Ventajas de los datos sintéticos vs datos reales

Técnicas más comunes para generar datos sintéticos:



TÉCNICA DE REMUESTREO:

Se extraen parte de los datos originales y se reorganizan aleatoriamente para crear nuevos conjuntos de datos.



MODELO PROBABILÍSTICO Y GENERATIVO (IA GENERATIVA):

Se construyen modelos probabilísticos basados en las distribuciones y relaciones observadas en los datos reales.



MÉTODOS DE PERTURBACIÓN Y ENMASCARAMIENTO:

Los datos reales se modifican de manera controlada para proteger la privacidad mientras se conservan características importantes.

Reduce los costes de inversión para desarrollar un proyecto con datos

Protegen identidades anonimizadas

El entrenamiento de modelos IA/ML es más completo

Facilitan la generación de simulaciones

Son más consistentes y uniformes que los datos reales

Impulsan la innovación

Eliminan los riesgos de privacidad



Consulta un ejemplo práctico de generación de datos sintéticos con métodos generativos [aquí](#)

EL PRESENTE Y EL FUTURO

Los datos sintéticos ya son una realidad.

En 2024 alrededor del 60% de los datos utilizados para proyectos de IA y análisis se generarán sintéticamente*, aunque estas previsiones podrían superarse.



- Datos generados artificialmente
- Generados a partir de reglas simples, modelos estadísticos, simulación y otras técnicas

- Obtenidos a partir de mediciones directas
- Limitados por razones de coste, logística y privacidad

*Fuente: Gartner. Extraída de Synthetic Data is About To Transform Artificial Intelligence de Forbes.

Puedes saber más sobre los datos sintéticos en el informe **'Datos sintéticos: ¿Qué son y para qué se usan?'**, donde además se incluye un caso práctico de cómo se generan y un posible uso.