

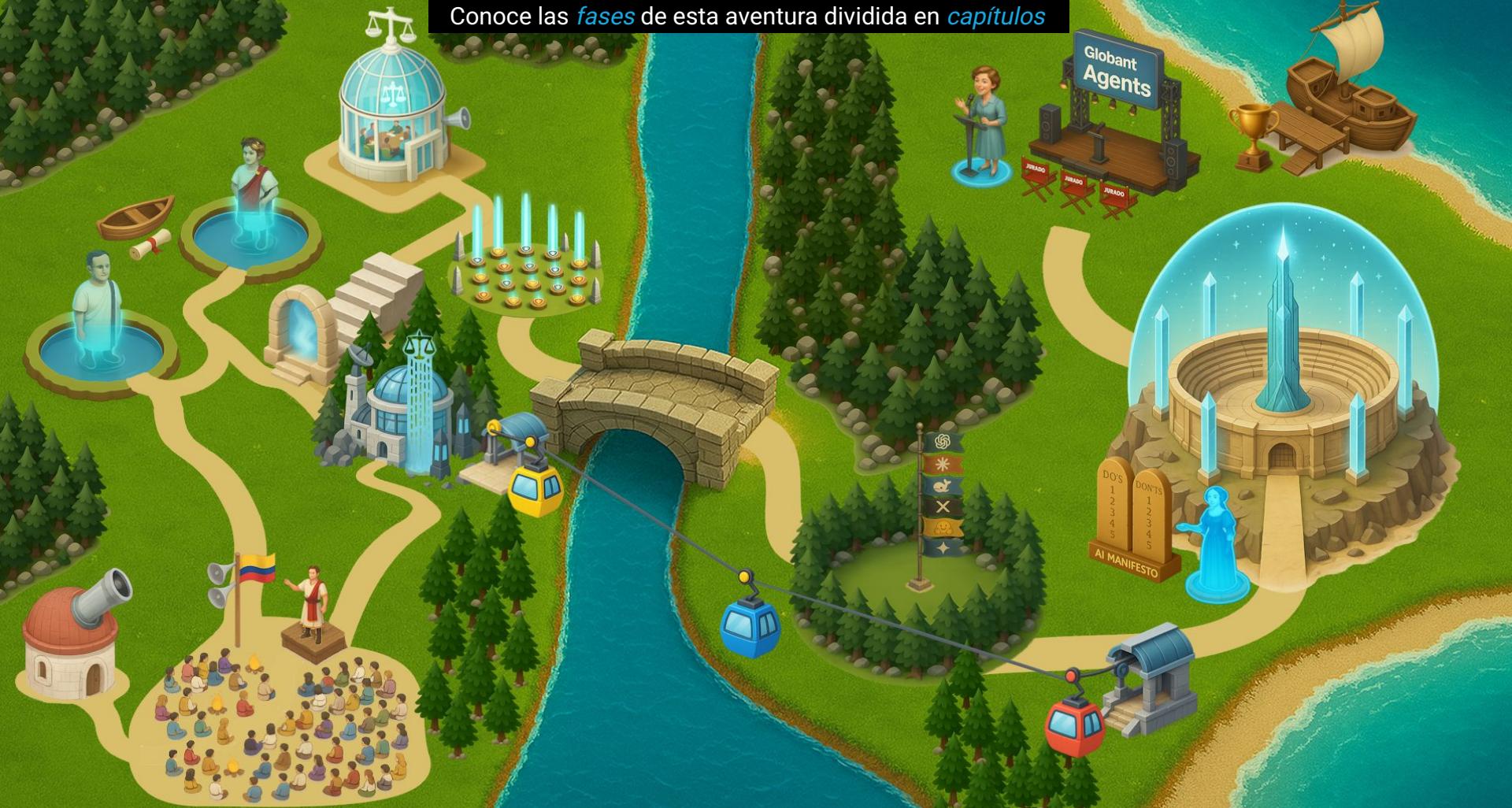


SENA Soft 2025 Synthetic Edition

Construyendo el futuro del trabajo:
humanos y máquinas inteligentes



Conoce las *fases* de esta aventura dividida en *capítulos*



Categoría de Desarrollo Integral

Conoce las *fases* de esta aventura dividida en *capítulos*



Categoría de Desarrollo Integral

Ruta Habilitadora



Datos Sintéticos y Calidad de Datos

Agenda

01

Fundamentos de
calidad de datos

02

Dimensiones de
calidad

03

Datos
sintéticos

Fundamentos de calidad de datos

Conceptos Básicos



¿Por qué los datos son tan **importantes**?

1. Son la bases de la diseño de **estrategias y operaciones** de negocio
2. Incrementan la **satisfacción y lealtad** de los clientes.
3. Procesos optimizados **reducen los costos** y aumentan la eficiencia.
4. Permiten crear **experiencias de cliente** personalizadas e innovadoras.
5. Datos confiables respaldan la **toma de decisiones** informadas en todos los niveles.

**Son el insumo fundamental para el
entrenamiento de modelos de IA**



¿Qué es la **calidad** de los datos?

La calidad de los datos es una **medida de qué tan bien** un determinado conjunto de datos satisface necesidades específicas de los usuarios.

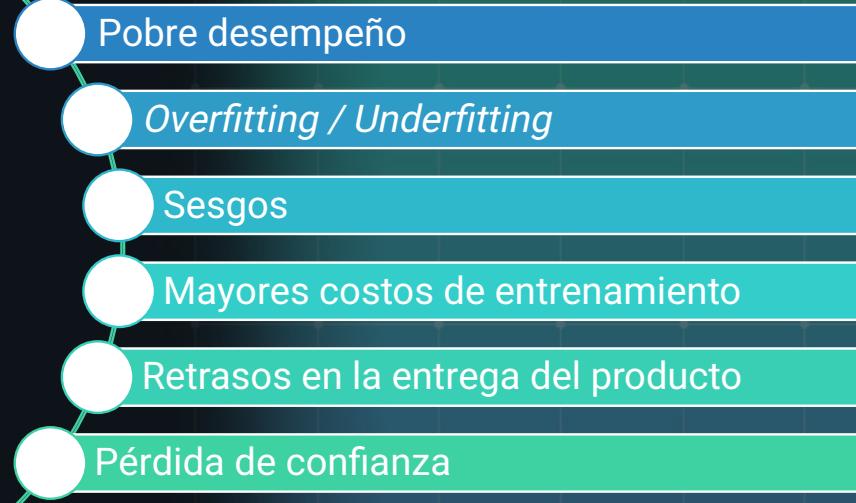
Junto con la **gestión de datos**, garantiza que los datos sean fiables (**verificación**) y adecuados para su finalidad prevista (**validación**)



Piensa en los costos



¿Cuáles son los costos de tener datos de mala calidad desde la perspectiva de **Machine Learning**?



¿Cómo analizar y comprender los datos?

Data Profiling

DIMENSIONES DE CALIDAD

Aplicación de las dimensiones de la calidad como marco de evaluación.

VISUALIZACIÓN

Uso de gráficos, tablas, diagramas, mapas y otros elementos visuales para determinar patrones, tendencias, relaciones y anomalías en los datos.

ANÁLISIS ESTADÍSTICO

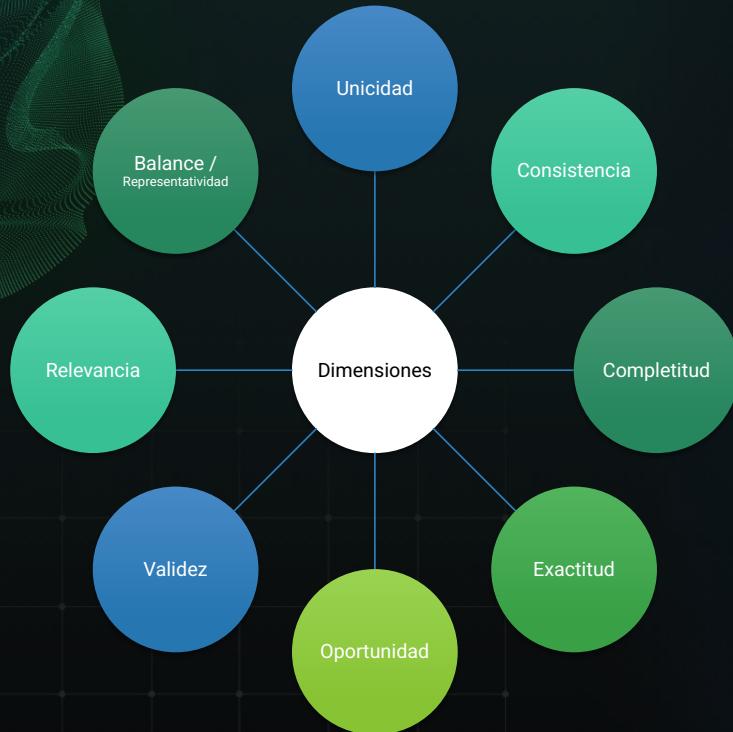
Cálculo de estadísticas básicas (promedio, mediana, desviación estándar, varianza) para obtener información sobre la distribución, el rango y la variabilidad de los datos.

Dimensiones de calidad

Evaluación Objetiva de los Datos



Dimensiones de Calidad de Datos



Conceptos clave

Las dimensiones de calidad representan atributos de los datos y miden hasta qué punto cumplen con estándares y requisitos específicos.

Estos aspectos se expresan como reglas que definen los criterios o condiciones utilizados para evaluar su calidad.

Dimensiones de Calidad de Datos

Unicidad	Un solo elemento representa la misma información
Consistencia	No hay contradicciones en los datos
Compleitud	Todos los datos necesarios están disponibles
Exactitud	Los datos corresponden con la realidad
Oportunidad	Los datos representan la realidad dentro del periodo de tiempo esperado
Validez	Los datos se ciñen a un dominio, estructura o formato específico
Relevancia	Los datos son significativos para la tarea a predecir
Balance y Representatividad	Las clases están representadas proporcionalmente

Datos Sintéticos

Creación de Datos Personalizados



¿Qué son datos sintéticos?

Datos generados artificialmente con el objeto de imitar datos reales en cuanto a su estructura y distribuciones, sin incluirlos en el dataset final.

Se crean a través de algoritmos y modelos. El resultado de su generación puede ser de tres tipos:

- Completamente sintéticos
- Parcialmente sintéticos
- Híbridos o aumentados





Uso de Datos Sintéticos

SUSTITUCIÓN DE DATOS SENSIBLES

Reemplazar datos que se encuentran protegidos por regulaciones o son de carácter privado.

GENERACIÓN DE VOLUMEN

Crear datos cuando estos son insuficientes y su costo o tiempo de obtención es elevado.

BALANCE Y COBERTURA

- Ajustar distribución de las clases para lograr balance o representatividad.
- Adicionar escenarios límite o extremos.

Métodos de generación de Datos Sintéticos

MÉTODOS ESTADÍSTICOS

Análisis de los datos reales para entender sus propiedades estadísticas – distribución, media y otras medidas de tendencia central – para aplicar un modelo estadístico en la generación de nuevas muestras de datos.

MODELOS BASADOS EN AGENTES (ABM)

Creación de agentes autónomos en un ambiente específico cuyas interacciones bajo determinadas reglas derivan en la generación de nuevos de datos.

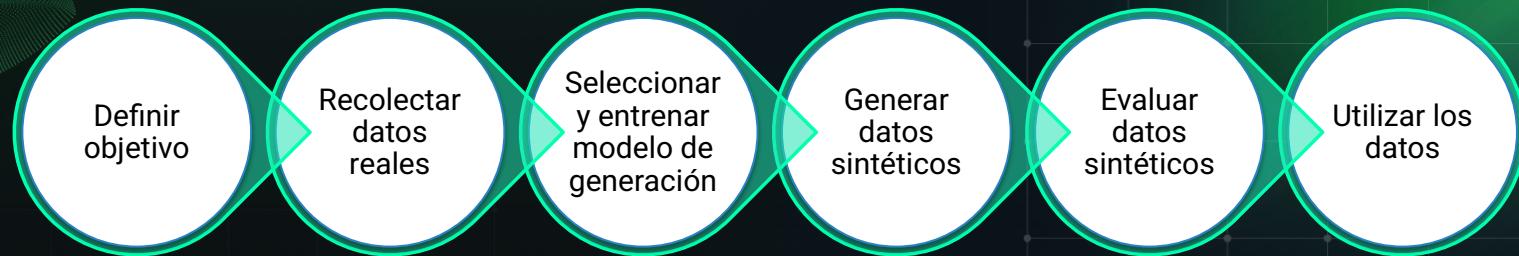
MÉTODOS DEEP LEARNING

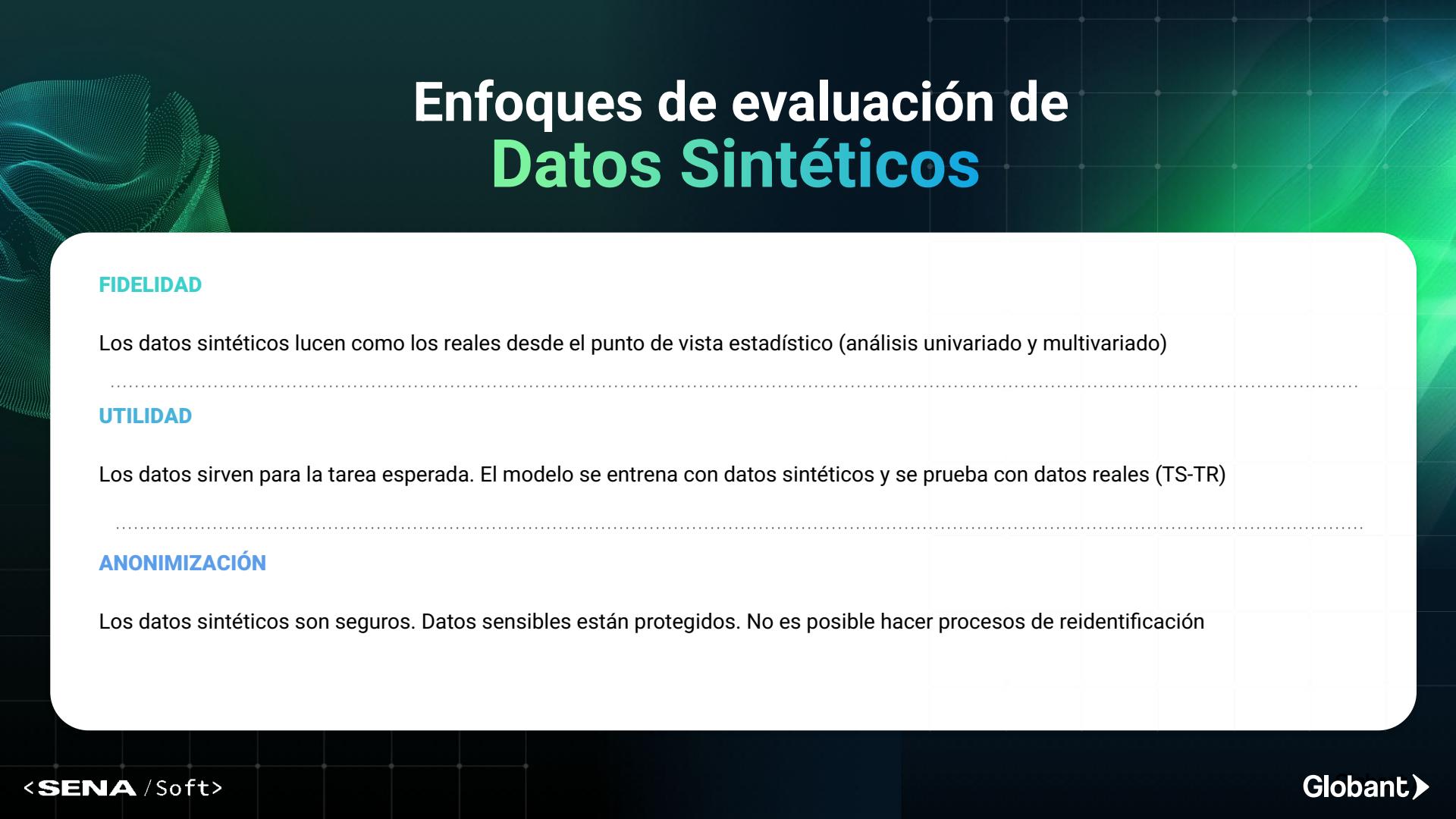
Uso de redes neuronales para abstraer patrones de los datos de entrenamiento para generar nuevos datos a partir de ese aprendizaje:

- Generative Adversarial Networks (GAN)
- Variational Autoencoders (VAEs)
- Modelos difusos

Datos Sintéticos

Resumen del proceso





Enfoques de evaluación de Datos Sintéticos

FIDELIDAD

Los datos sintéticos lucen como los reales desde el punto de vista estadístico (análisis univariado y multivariado)

UTILIDAD

Los datos sirven para la tarea esperada. El modelo se entrena con datos sintéticos y se prueba con datos reales (TS-TR)

ANONIMIZACIÓN

Los datos sintéticos son seguros. Datos sensibles están protegidos. No es posible hacer procesos de reidentificación

