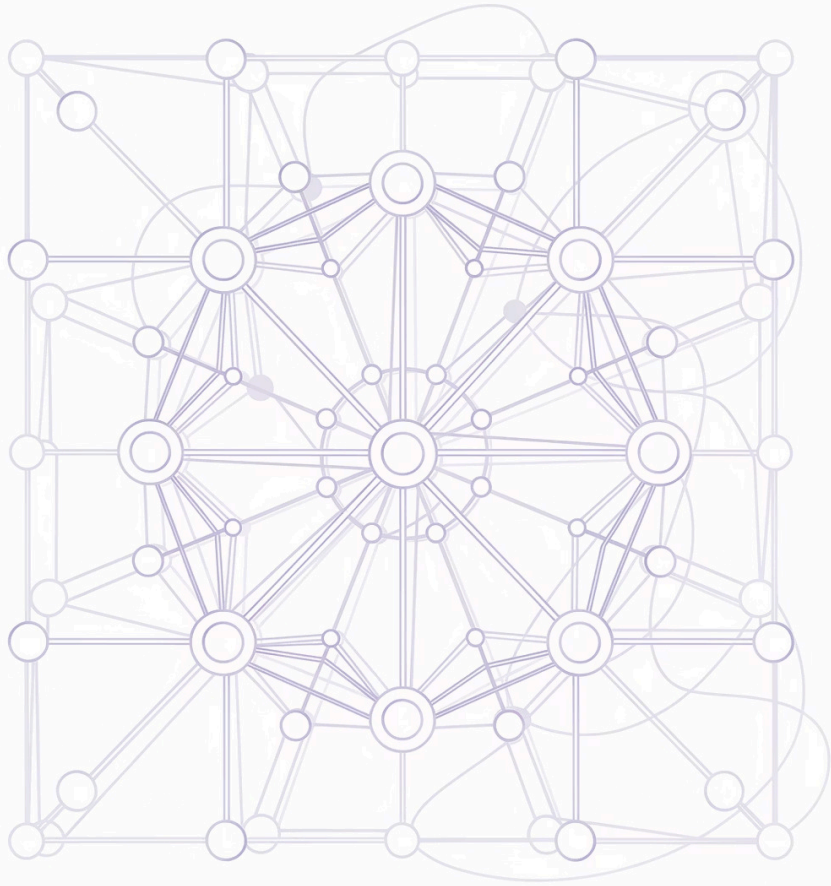
Datos Sintéticos en Aprendizaje Automático

Una exploración técnica de métodos, aplicaciones y desafíos en la generación de datos artificiales para sistemas de machine learning.

Alvaro Esteban Castro - MSP - MSO - Doctorando Ciencias Aplicadas 2025

Castro, A. E. (2025). Datos Sintéticos en Aprendizaje Automático



Contenido del Seminario

01

Fundamentos y Ventajas

Conceptos básicos, beneficios clave y casos de uso principales de datos sintéticos

02

Métodos de Generación

GANs, VAEs, CTGANs y técnicas basadas en escenarios para producción de datos

03

Aplicaciones Prácticas

Detección de anomalías, privacidad, mejora del rendimiento de modelos

04

Desafíos y Ética

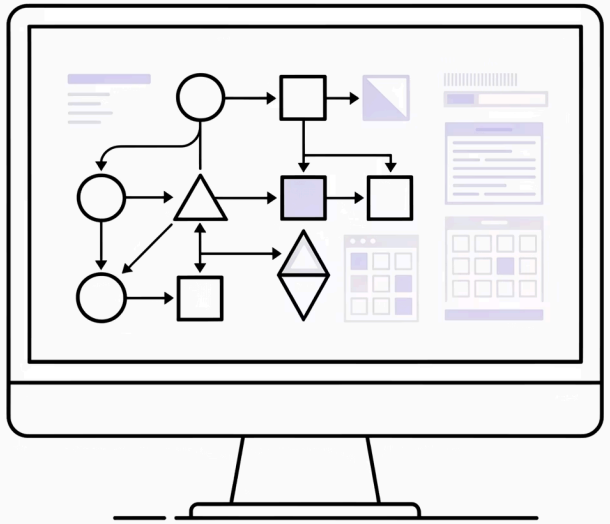
Confiabilidad, validación, implicaciones éticas y control de calidad

¿Qué Son los Datos Sintéticos?

Los datos sintéticos son información generada artificialmente mediante algoritmos computacionales que imitan las propiedades estadísticas y estructurales de los datos reales sin contener registros reales de individuos o eventos (Wang et al., 2024).

A diferencia de los datos tradicionales recopilados del mundo real, los datos sintéticos se crean con el propósito específico de entrenar, validar y probar modelos de aprendizaje automático. Su capacidad para preservar patrones estadísticos mientras elimina información identificable los convierte en una herramienta fundamental en dominios regulados.

Representan una solución tecnológica a los desafíos contemporáneos de disponibilidad, privacidad y diversidad de datos en sistemas de inteligencia artificial.



Contexto: La Necesidad de Datos Sintéticos

Escasez de Datos

En sectores como salud y finanzas, la recopilación de grandes volúmenes de datos reales es compleja, costosa y limitada por restricciones operativas

Regulaciones de Privacidad

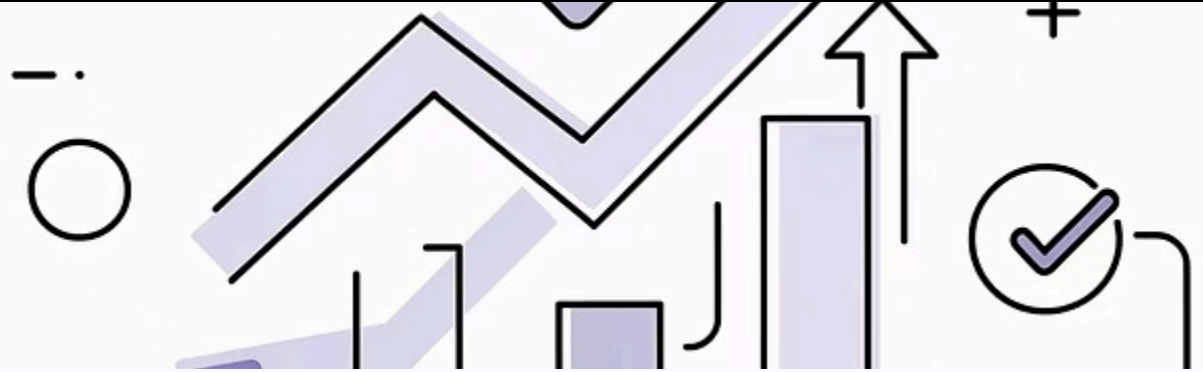
Normativas como GDPR y HIPAA imponen restricciones severas sobre el uso y compartición de datos personales sensibles

Desequilibrio de Clases

Muchos conjuntos de datos reales presentan clases subrepresentadas que afectan negativamente al rendimiento de los modelos predictivos

Costes de Etiquetado

El proceso de anotación manual de datos reales requiere inversiones significativas de tiempo y recursos humanos especializados



Ventajas Principales de los Datos Sintéticos

Aumento y Diversidad

Expansión de conjuntos de datos limitados y mejora de la representación de clases minoritarias, especialmente crítico en dominios con restricciones de recopilación (Wang et al., 2024)

Privacidad y Seguridad

Eliminación de información de identificación personal mientras se preservan propiedades estadísticas, facilitando el cumplimiento normativo (Lu et al., 2023)

Mejora del Rendimiento

Incremento significativo en métricas de precisión, recall y F1-score cuando se combinan datos originales y sintéticos (Sangve et al., 2025)

Rentabilidad

Reducción drástica de costes asociados con la adquisición, almacenamiento y etiquetado de datos del mundo real

Aumento de Datos: Abordando la Escasez

Desafío Crítico en ML

Los algoritmos de aprendizaje automático modernos, especialmente las arquitecturas de deep learning, requieren grandes volúmenes de datos etiquetados para alcanzar niveles óptimos de generalización. Sin embargo, en sectores como la medicina y las finanzas, los conjuntos de datos disponibles suelen ser limitados.

Los datos sintéticos proporcionan una solución escalable al generar muestras adicionales que mantienen las características estadísticas del conjunto original, permitiendo entrenar modelos más robustos sin depender exclusivamente de datos reales costosos o difíciles de obtener (Goyal & Mahmoud, 2024).

Aplicación en Salud: La generación de registros médicos sintéticos permite entrenar sistemas de diagnóstico asistido sin comprometer la privacidad del paciente, abordando simultáneamente la escasez de datos para enfermedades raras.

Detección de Anomalías con Datos Sintéticos

Datos Normales Limitados

Conjuntos de entrenamiento con abundancia de comportamientos normales pero escasez de anomalías reales

Generación Sintética

Creación controlada de anomalías sintéticas que representan diversos tipos de ataques o comportamientos anómalos

Entrenamiento Balanceado

Modelos entrenados con distribución equilibrada aprenden a distinguir patrones normales de anómalos efectivamente

Detección Robusta

Sistemas de seguridad capaces de identificar amenazas previamente no observadas en datos reales

En el ámbito de la ciberseguridad, los datos sintéticos se utilizan para generar anomalías específicas que entrenan modelos de detección de intrusiones, demostrando eficacia significativa en el aumento del tamaño muestral y la diversidad de ataques simulados (Gurianov, 2024).

Privacidad y Cumplimiento Normativo

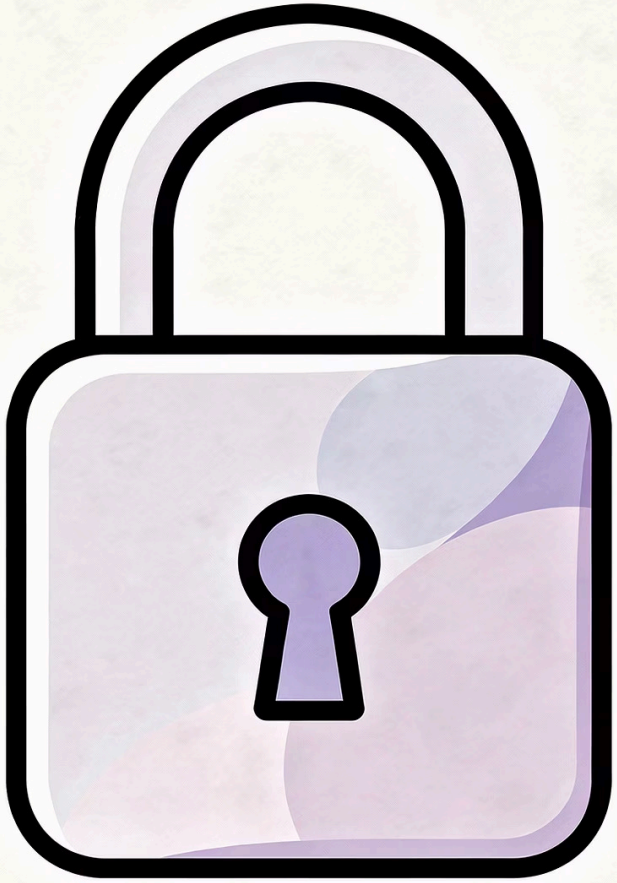
El Dilema de la Privacidad

Las organizaciones enfrentan un dilema fundamental: necesitan grandes volúmenes de datos para desarrollar modelos precisos, pero las regulaciones de privacidad (GDPR, HIPAA, CCPA) restringen severamente el uso y compartición de información personal.

Los datos sintéticos ofrecen una solución elegante al permitir la generación de conjuntos de datos que preservan las propiedades estadísticas útiles para el aprendizaje automático mientras eliminan cualquier vínculo con individuos reales (Lu et al., 2023).

Ventajas Regulatorias

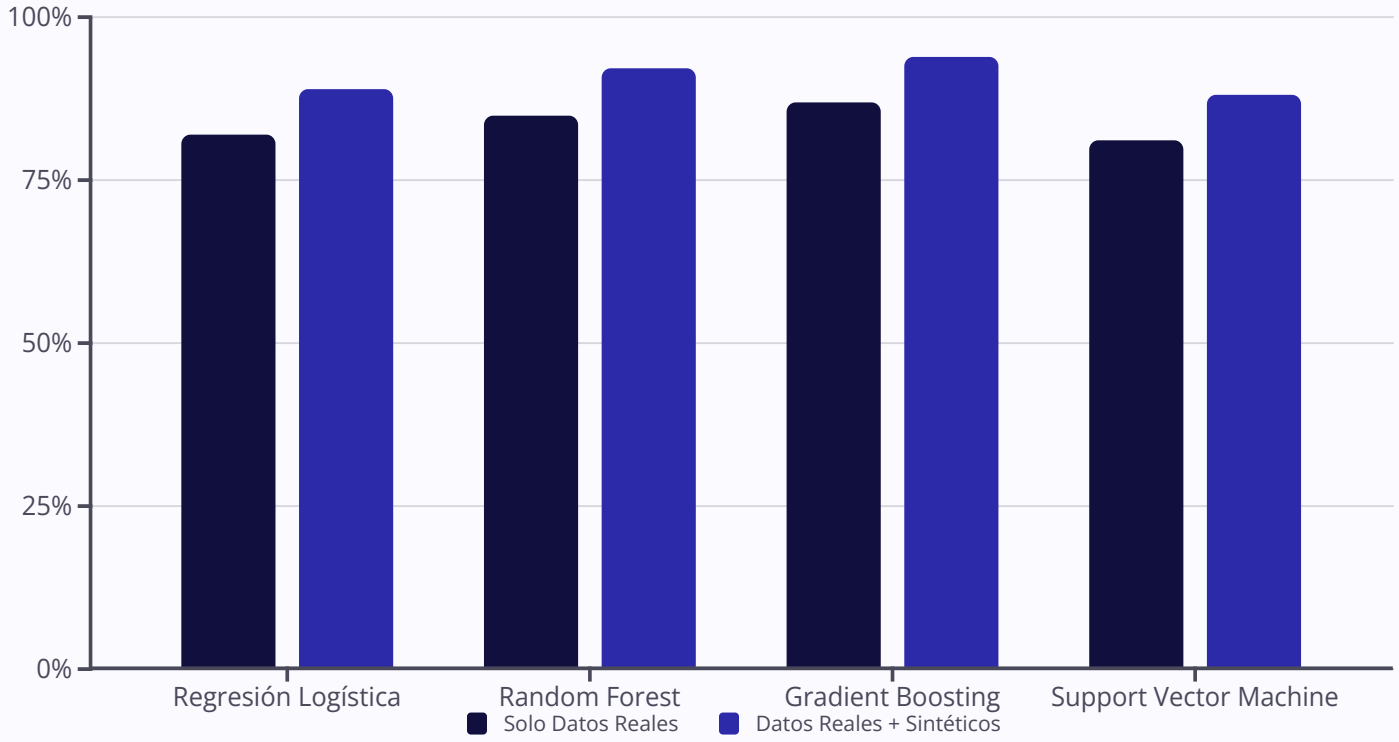
- Compartición sin restricciones entre equipos y organizaciones
- Eliminación de riesgos de brechas de datos personales
- Cumplimiento automático con regulaciones de anonimización
- Reducción de costes de cumplimiento normativo
- Facilitación de colaboraciones internacionales en investigación



Impacto en el Rendimiento de Modelos

Mejoras Cuantificables en Métricas de Evaluación

La investigación empírica demuestra que la combinación estratégica de datos originales y sintéticos produce mejoras significativas en el rendimiento de diversos algoritmos de aprendizaje automático. Los estudios reportan incrementos consistentes en precisión, recall y F1-score cuando se incorporan datos sintéticos al proceso de entrenamiento (Sangve et al., 2025).



Estos resultados son particularmente pronunciados en escenarios con conjuntos de datos pequeños o desequilibrados, donde los datos sintéticos proporcionan la diversidad adicional necesaria para mejorar la capacidad de generalización del modelo.

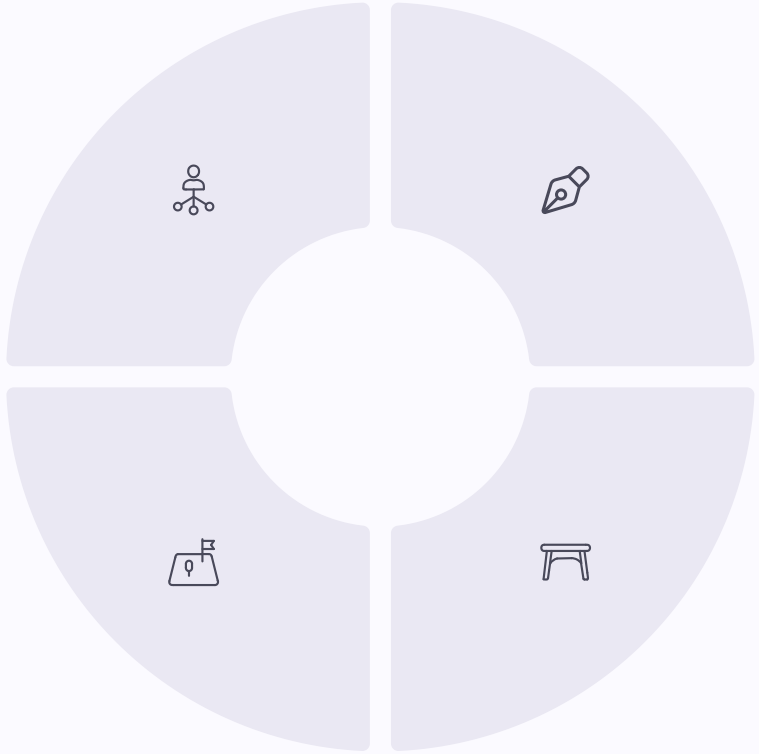
Métodos de Generación: Panorama General

GANs

Redes Generativas Adversarias: arquitectura dual generador-discriminador para datos realistas

VAEs

Autoencoders Variacionales: codificación probabilística en espacio latente continuo



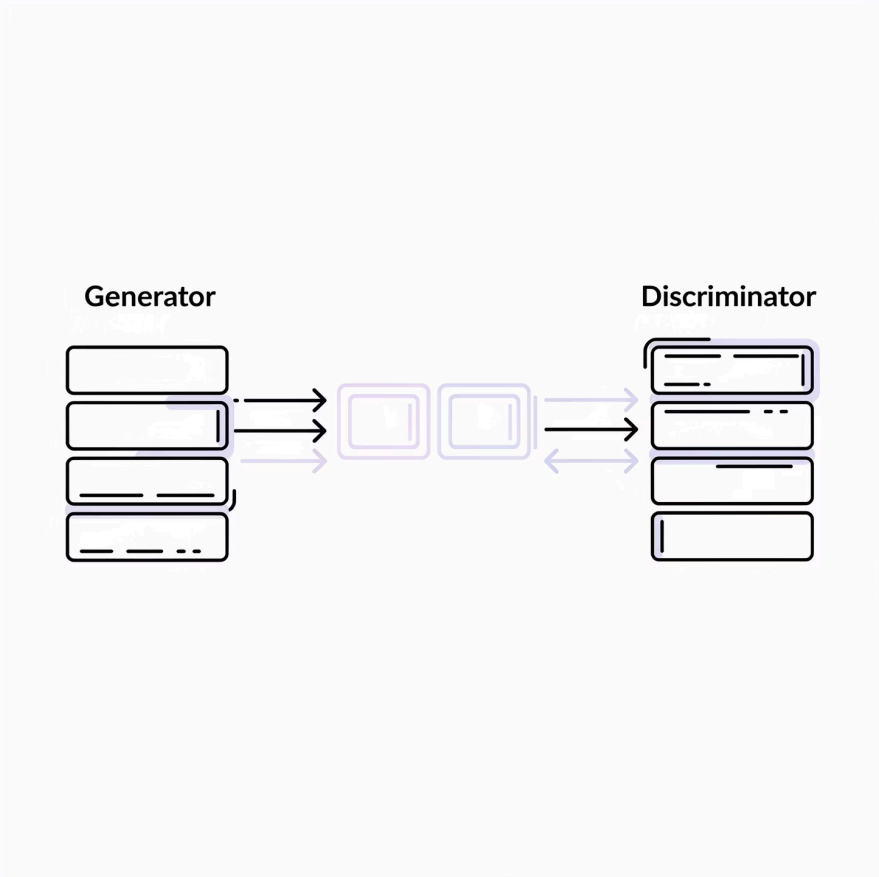
Métodos Basados en Escenarios

Generación orientada a cobertura del dominio operativo completo

CTGANs

GANs Tabulares Condicionales: especializados en preservar propiedades de datos estructurados

Redes Generativas Adversarias (GANs)



Arquitectura Adversaria

Las GANs, introducidas por Goodfellow et al., emplean una arquitectura innovadora basada en competición: un generador crea muestras sintéticas mientras un discriminador intenta distinguirlas de los datos reales. Este proceso adversario conduce a la generación de datos sintéticos cada vez más realistas.

Componentes Clave

- **Generador:** Red neuronal que transforma ruido aleatorio en datos sintéticos
- **Discriminador:** Red clasificadora que evalúa autenticidad de las muestras
- **Función de pérdida adversaria:** Guía el entrenamiento conjunto de ambas redes

Las GANs se han convertido en el método más utilizado para generar datos sintéticos realistas en dominios que van desde imágenes hasta datos tabulares complejos (et al., 2025).

Autoencoders Variacionales (VAEs)

1

Codificación

La red codificadora mapea datos de entrada a una distribución probabilística en el espacio latente

2

Espacio Latente

Representación comprimida de características esenciales con estructura probabilística continua

3

Muestreo

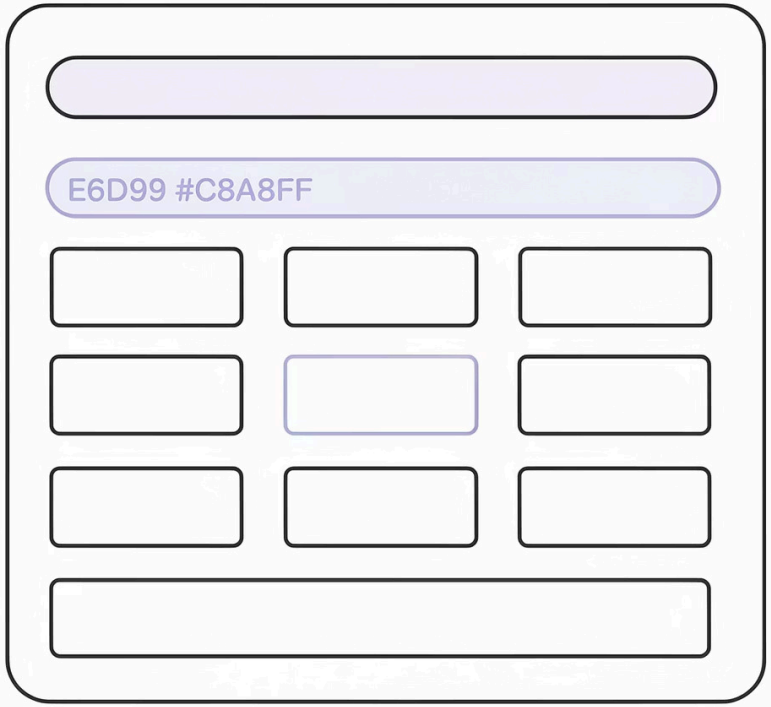
Selección de puntos del espacio latente mediante distribución gaussiana aprendida

4

Decodificación

Reconstrucción de datos sintéticos nuevos a partir de las representaciones latentes muestreadas

Los VAEs ofrecen ventajas sobre las GANs en términos de estabilidad de entrenamiento y capacidad de interpolación suave en el espacio latente. Su enfoque probabilístico permite generar variaciones controladas de datos sintéticos, lo cual es valioso para aplicaciones que requieren diversidad explícita y cuantificable (et al., 2025).



CTGANs para Datos Tabulares

Desafíos Específicos de Datos Tabulares

Los datos tabulares presentan desafíos únicos para la generación sintética: combinan variables continuas y categóricas, exhiben correlaciones complejas entre columnas, y contienen distribuciones asimétricas. Las GANs estándar, diseñadas principalmente para datos de imagen, no capturan adecuadamente estas propiedades.

Innovaciones de CTGAN

- Mode-specific normalization para variables continuas
- Conditional generator para columnas categóricas
- Training-by-sampling para manejar desequilibrios
- Preservación de correlaciones inter-columna

Aplicaciones Empresariales

Los CTGANs han demostrado eficacia particular en tareas de inteligencia empresarial, permitiendo generar datos sintéticos que mantienen las propiedades estadísticas complejas necesarias para análisis financieros, segmentación de clientes y predicción de riesgos (Sangve et al., 2025; Dildabek & Абдиахметова, 2023).

Métodos Basados en Escenarios

01

Definición del Dominio Operativo

Identificación exhaustiva del rango completo de condiciones, parámetros y situaciones relevantes para el sistema

02

Análisis de Cobertura

Evaluación de qué regiones del dominio están representadas insuficientemente en datos reales disponibles

03

Generación Dirigida

Creación sistemática de escenarios sintéticos que cubren gaps identificados en el espacio de diseño

04

Validación de Cobertura

Verificación cuantitativa de que los datos sintéticos proporcionan cobertura adecuada del dominio completo

Este enfoque es particularmente valioso en sistemas autónomos y aplicaciones de seguridad crítica, donde garantizar cobertura exhaustiva del dominio operativo es fundamental para la confiabilidad del sistema (Hirschle et al., 2024).

Desafío: Confianza y Confiabilidad

El Problema de la Validación

Uno de los desafíos fundamentales en el uso de datos sintéticos es cuantificar la confiabilidad de los hallazgos y predicciones derivados de ellos. ¿Cómo podemos estar seguros de que un modelo entrenado parcialmente con datos sintéticos se comportará correctamente en el mundo real?

Este interrogante requiere protocolos de validación rigurosos que evalúen no solo la fidelidad estadística de los datos sintéticos, sino también su impacto en las métricas de rendimiento downstream de los modelos.

Estrategias de Validación

- **Métricas de fidelidad:** Comparación de distribuciones estadísticas entre datos reales y sintéticos
- **Pruebas de utilidad:** Evaluación del rendimiento de modelos en tareas específicas
- **Validación cruzada:** Entrenamiento con datos sintéticos, validación con datos reales
- **Análisis de privacidad:** Verificación de ausencia de membresía de datos originales

La comunidad científica enfatiza la necesidad de establecer estándares rigurosos para la validación de datos sintéticos antes de su despliegue en aplicaciones críticas (Beyond Privacy: Navigating the Opportuni..., 2023).

Consideraciones Éticas y Equidad

Perpetuación de Sesgos

Los datos sintéticos generados a partir de datos reales sesgados pueden heredar y amplificar discriminaciones sistemáticas presentes en los datos originales

Equidad Algorítmica

Necesidad de garantizar que los modelos entrenados con datos sintéticos no discriminen contra grupos protegidos o minoritarios

Transparencia

Obligación de divulgar cuándo y cómo se utilizan datos sintéticos en sistemas que afectan decisiones sobre personas

Responsabilidad

Establecimiento de marcos claros de accountability para errores derivados del uso de datos sintéticos

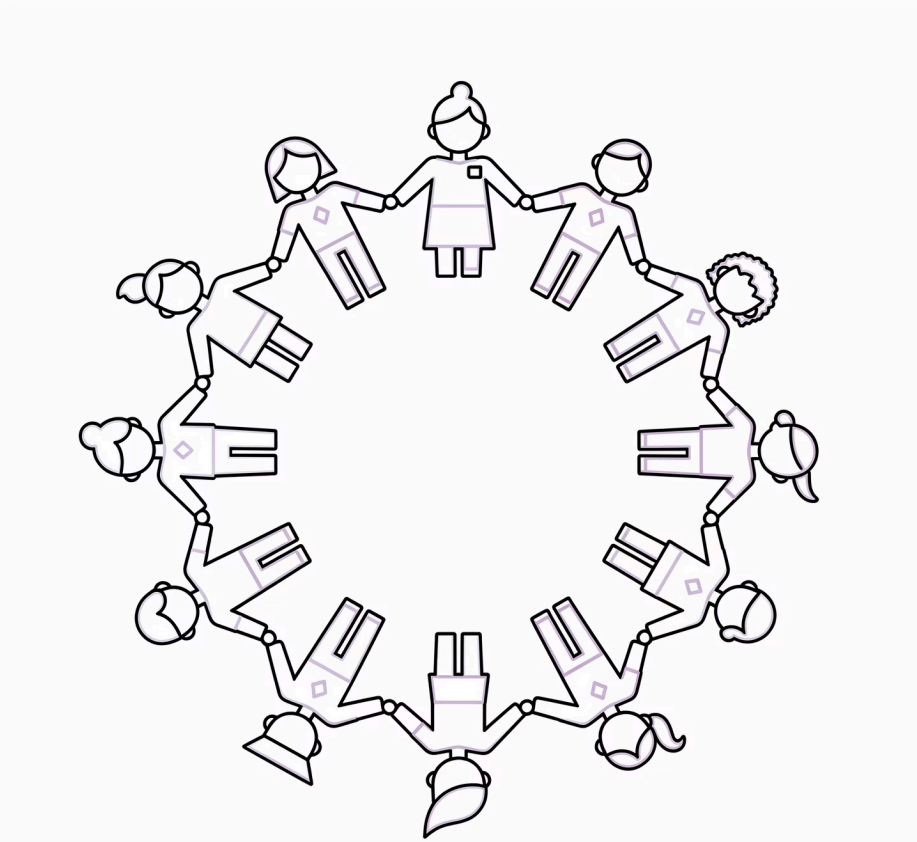
Las implicaciones éticas del uso de datos sintéticos deben abordarse proactivamente para garantizar un desarrollo responsable de la inteligencia artificial. Esto incluye el diseño intencional de métodos que corrijan sesgos y protejan a grupos vulnerables (Lu et al., 2023; Rodriguez & Howe, 2019).

Mitigación de Sesgos mediante Datos Sintéticos

Corrección de Desequilibrios Demográficos

Los datos sintéticos pueden utilizarse estratégicamente para corregir sesgos existentes en conjuntos de datos originales. Al generar muestras adicionales de grupos subrepresentados, es posible mejorar métricas de equidad como la paridad demográfica y la igualdad de oportunidades sin comprometer la precisión predictiva.

Por ejemplo, las GANs se han empleado exitosamente para generar conjuntos de datos sintéticos que mejoran la representación de minorías demográficas en sistemas de reconocimiento facial, abordando así los sesgos documentados en estos sistemas (Fabuyi, 2024).



Principios de Diseño Justo

- Análisis pre-generación de sesgos en datos originales
- Generación dirigida a grupos subrepresentados
- Validación de métricas de equidad post-entrenamiento
- Auditorías independientes de impacto discriminatorio

Calidad y Fidelidad de Datos Sintéticos

Dimensiones de la Calidad

La efectividad de los datos sintéticos depende críticamente de su calidad y fidelidad. Los datos sintéticos de baja calidad no solo fallan en mejorar el rendimiento de los modelos, sino que pueden introducir artefactos y patrones espurios que degradan la capacidad de generalización (Wang et al., 2024).

Fidelidad Estadística

Grado en que las propiedades estadísticas (distribuciones, correlaciones, momentos) de los datos sintéticos coinciden con las de los datos reales

Utilidad Predictiva

Capacidad de los datos sintéticos para entrenar modelos que generalicen efectivamente a datos reales no vistos

Privacidad Preservada

Garantía de que los datos sintéticos no permiten re-identificación o inferencia de información sobre individuos en el conjunto original

Diversidad

Riqueza de variación en los datos sintéticos, evitando mode collapse y asegurando cobertura del espacio de características

Aplicaciones Avanzadas: Gemelos Digitales

Cohortes Virtuales en Medicina Personalizada

Una aplicación particularmente prometedora de los datos sintéticos es la creación de gemelos digitales de pacientes: representaciones virtuales que permiten simular respuestas a tratamientos y predecir resultados clínicos sin exponer a pacientes reales a riesgos experimentales.

Ensayos In Silico

Los datos sintéticos permiten generar cohortes virtuales de pacientes para estimar resultados de ensayos clínicos mediante simulación computacional. Esta aproximación puede:

- Reducir drásticamente el tiempo y coste de desarrollo farmacéutico
- Identificar subpoblaciones que responden diferencialmente a tratamientos
- Evaluar seguridad antes de exposición humana
- Generar datos contrafácticos para análisis causal

Medicina de Precisión

Los gemelos digitales permiten personalizar tratamientos al simular cómo responderá un paciente específico a diferentes intervenciones terapéuticas, considerando su perfil genético, historial clínico y características fenotípicas únicas (Wang et al., 2024).

Esta capacidad representa un cambio paradigmático hacia una medicina verdaderamente personalizada y predictiva.

Conclusiones y Direcciones Futuras

1 Estado Actual

Los datos sintéticos se han establecido como herramienta fundamental en el ecosistema del aprendizaje automático, ofreciendo soluciones viables a desafíos de privacidad, escasez y sesgos

2 Desafíos Pendientes

Persisten interrogantes sobre validación, confiabilidad, implicaciones éticas y establecimiento de estándares de calidad universalmente aceptados

3 Investigación Futura

Desarrollo de métodos más sofisticados, métricas de evaluación robustas, frameworks éticos y aplicaciones en dominios emergentes como medicina personalizada

---

Reflexión Final

El campo de los datos sintéticos está en una fase de maduración acelerada. Su adopción creciente en industria y academia sugiere que se convertirán en componente estándar de las pipelines de aprendizaje automático. Sin embargo, su éxito a largo plazo dependerá de nuestra capacidad colectiva para abordar los desafíos técnicos, éticos y regulatorios de manera rigurosa y responsable, asegurando que esta tecnología sirva al objetivo último de desarrollar sistemas de inteligencia artificial más justos, confiables y beneficiosos para la sociedad.

Referencias clave: Wang et al. (2024), Sangve et al. (2025), Lu et al. (2023), Rodriguez & Howe (2019), Gurianov (2024), Fabuyi (2024), Hirschle et al. (2024), Dildabek & Абдиахметова (2023)