1.6 Fuentes de datos

Los datos son la base de la IA. Es fundamental usar **datasets de calidad** y bien documentados. Fuentes típicas incluyen repositorios académicos (Imagenet, COCO, UCI ML), colecciones científicas (GenBank para biología, EarthExplorer para geología, etc.) y plataformas como Hugging Face Datasets o Kaggle Datasets, que agrupan datos populares de múltiples dominios. En educación de ciencias, es interesante explorar datasets abiertos de áreas específicas (genómica, geolocalización, fórmulas matemáticas, espectros de química, etc.).

Hay que considerar la **ética y gobernanza**: respetar licencias y privacidad de datos (p.ej. evitar datos personales sin consentimiento), balancear representatividad (reducir sesgos), y documentar las fuentes. Muchos proyectos de IA académicos ahora incluyen tarjetas de datos (“datasheets”) que explican qué contiene y qué sesgos potenciales hay. Asimismo, organismos como la UE promueven iniciativas para gobernar la IA responsablemente, algo a destacar en clases de ética de la tecnología.

Conjuntos de datos

Los modelos de inteligencia artificial se entrenan a partir de grandes colecciones de datos que contienen ejemplos del tipo de información que el sistema debe aprender a procesar. Estos **datasets** constituyen uno de los elementos fundamentales del desarrollo de la IA, ya que la calidad, diversidad y tamaño de los datos influyen directamente en el rendimiento del modelo.

Datos de texto

En el caso de los modelos de lenguaje, los conjuntos de datos suelen estar formados por **grandes colecciones de textos** procedentes de libros, artículos científicos, páginas web, código fuente o conversaciones. Estos datos permiten a los modelos aprender patrones del lenguaje, estructuras gramaticales y relaciones entre conceptos.

Datos de imagen

Los modelos de visión por computador se entrenan con **bases de datos de imágenes etiquetadas** que indican qué aparece en cada fotografía. Estos datasets permiten que los modelos aprendan a reconocer objetos, personas o escenas. Algunos conjuntos de datos muy conocidos incluyen millones de imágenes clasificadas en distintas categorías.

Datos de audio y voz

Copyright 2026

Los sistemas de reconocimiento o generación de voz utilizan **grabaciones de audio acompañadas de transcripciones**. Con estos datos los modelos aprenden a relacionar sonidos con palabras o a generar voz sintética con características naturales.

Datos multimodales

En los últimos años se han desarrollado datasets que **combinan diferentes tipos de información**, como imágenes con descripciones en texto, vídeos con subtítulos o audio con anotaciones. Estos conjuntos de datos permiten entrenar modelos capaces de comprender varias modalidades de información al mismo tiempo.

Datos especializados o de dominio

Además de los grandes datasets generales, también existen **conjuntos de datos específicos para un ámbito concreto**, como medicina, derecho, finanzas o educación. Estos datasets se utilizan para especializar modelos mediante técnicas como el fine-tuning o para construir sistemas de consulta basados en RAG.

Calidad y preparación de los datos

Antes de utilizar un dataset en un modelo de IA, los datos suelen pasar por procesos de **limpieza, filtrado y anotación** para eliminar errores, duplicados o información irrelevante. También es importante considerar aspectos como el sesgo de los datos, la privacidad y los derechos de uso, ya que estos factores pueden influir en el comportamiento y la fiabilidad de los sistemas de inteligencia artificial.

Fuentes de datasets

Existen diversas plataformas donde investigadores y desarrolladores pueden encontrar conjuntos de datos para proyectos de inteligencia artificial. Algunas de las más utilizadas son **Kaggle**, que ofrece miles de datasets y competiciones de ciencia de datos; **Google Dataset Search**, un buscador especializado en localizar conjuntos de datos públicos en internet; y repositorios como **Hugging Face**, que además de modelos incluye colecciones de datasets preparados para entrenamiento y experimentación.

Revision #3
Created 2026-03-04 17:48:15 CET by Luis Hueso
Updated 2026-03-16 18:20:43 CET by Luis Hueso

Copyright 2026