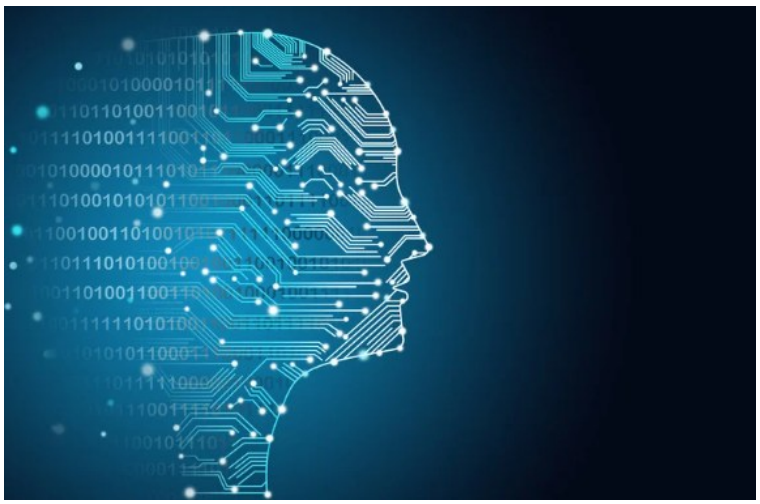
# 4. Comprensión de los algoritmos no supervisados.

No siempre disponemos de datos etiquetados. En muchos casos reales, tenemos grandes volúmenes de información, pero no sabemos de antemano qué categorías o grupos existen en ella, o simplemente no tenemos el tiempo ni los recursos para etiquetar los datos manualmente. En estos escenarios, el aprendizaje no supervisado nos permite descubrir la estructura subyacente de los datos sin ninguna señal de supervisión externa.

## 4.1. Qué son los algoritmos no supervisados: principales técnicas y aplicaciones.

En el aprendizaje no supervisado, el algoritmo recibe únicamente los datos de entrada sin ningún tipo de etiqueta o respuesta correcta. Su objetivo es encontrar patrones, agrupaciones o representaciones comprimidas de los datos que revelen su estructura interna. Aunque puede parecer menos directamente útil que el aprendizaje supervisado —al fin y al cabo, no estamos prediciendo nada concreto—, el aprendizaje no supervisado tiene aplicaciones enormemente valiosas en la práctica.



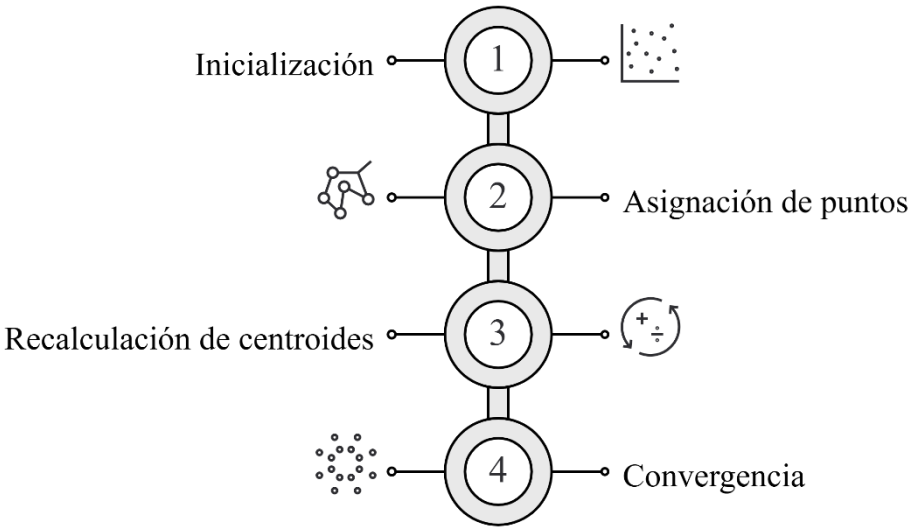
La segmentación de clientes es quizás el caso de uso más común: dado el comportamiento de compra de millones de clientes, el algoritmo identifica automáticamente grupos de clientes con comportamientos similares —los compradores impulsivos de fin de semana, los compradores de oferta planificadores, los clientes de alto valor y baja frecuencia— sin que nadie haya definido previamente cuáles son esos grupos. El resultado es una segmentación basada en los propios datos, mucho más rica y matizada que las segmentaciones demográficas tradicionales.

Otras aplicaciones relevantes incluyen la detección de anomalías —identificar transacciones, lecturas de sensores o comportamientos de red que se desvían significativamente de lo habitual, sin necesidad de ejemplos etiquetados de anomalías previas—, la reducción de dimensionalidad —comprimir la información de cientos de variables en unas pocas que capturen la mayor parte de la varianza—, y el aprendizaje de representaciones —generar representaciones vectoriales compactas de objetos complejos como palabras, imágenes o grafos que pueden usarse como entrada para otros modelos—.

## 4.2. K-means, análisis de componentes principales (PCA) y otros métodos de clustering.

K-means es el algoritmo de clustering más conocido y utilizado. Su funcionamiento es intuitivo: dado un número k de grupos deseados, el algoritmo asigna iterativamente cada punto de datos al centroide más cercano y luego recalcula los centroides como la media de los puntos asignados a cada grupo, repitiendo el proceso hasta que las asignaciones convergen. Es computacionalmente eficiente y escala bien a grandes conjuntos de datos, pero tiene dos limitaciones importantes: hay que especificar el número de clusters k de antemano —lo que requiere conocimiento del dominio o exploración previa—, y solo puede encontrar clusters de forma convexa y tamaño similar.

**Proceso iterativo del algoritmo K-means**




El Análisis de Componentes Principales —PCA, del inglés Principal Component Analysis— es la técnica de reducción de dimensionalidad más clásica. Dado un conjunto de datos con muchas variables correlacionadas, PCA encuentra las direcciones del espacio —los componentes principales— en las que la varianza de los datos es máxima, y proyecta los datos sobre un número reducido de estas direcciones. El resultado es un conjunto de datos con muchas menos dimensiones que captura la mayor parte de la información relevante. PCA es especialmente útil como paso previo a otros algoritmos de machine learning para reducir el coste computacional y mejorar la señal al eliminar el ruido de las dimensiones menos informativas.

Otros algoritmos no supervisados importantes son DBSCAN —que encuentra clusters de forma arbitraria y puede identificar puntos de ruido sin asignarlos a ningún grupo—, los modelos de mezcla gaussiana —que asumen que los datos provienen de una mezcla de distribuciones gaussianas y estiman sus parámetros—, t-SNE y UMAP —técnicas de reducción de dimensionalidad no lineal especialmente útiles para visualizar datos de alta dimensión en dos o tres dimensiones—, y los autoencoders —redes neuronales que aprenden una representación comprimida de los datos comprimiéndolos y reconstruyéndolos—.

ACTIVIDAD 13

Une cada concepto de la columna A con su definición de la columna B escribiendo la letra correspondiente.

Columna A — Conceptos

1. K-means.
2. PCA.
3. DBSCAN.
4. Detección de anomalías.

Columna B — Definiciones

A. Técnica de reducción de dimensionalidad que proyecta los datos sobre las direcciones de máxima varianza para comprimir la información.

B. Identificación de observaciones que se desvían significativamente del comportamiento habitual sin necesidad de ejemplos etiquetados previos.

C. Algoritmo de clustering que asigna iterativamente cada punto al centroide más cercano; requiere especificar el número de grupos de antemano.

D. Algoritmo de clustering basado en densidad que puede encontrar clusters de forma arbitraria e identificar puntos de ruido.

EJEMPLO PRÁCTICO · Segmentación de clientes en Zara: la IA detrás del escaparate

Inditex, la empresa propietaria de Zara, Pull&Bear, Massimo Dutti y otras marcas, es uno de los mayores grupos de moda del mundo, con más de 5 900 tiendas en 96 países y un volumen de datos de clientes extraordinario: cada compra en tienda y online, cada interacción con la app, cada devolución, cada búsqueda en el catálogo digital.

El equipo de analítica de Inditex aplica algoritmos de clustering no supervisado —principalmente variantes de K-means y modelos de mezcla gaussiana— sobre el comportamiento de compra de sus clientes para identificar segmentos con características homogéneas. A diferencia de una segmentación demográfica clásica —mujeres de 25-35 años de renta media—, la segmentación basada en datos de comportamiento produce grupos mucho más accionables: por ejemplo, "clientes que compran exclusivamente en rebajas, con cestas de alto valor y alta frecuencia en enero y julio", o "clientes que compran regularmente prendas de nueva colección en las primeras dos semanas, sensibles a la exclusividad y poco a los descuentos".

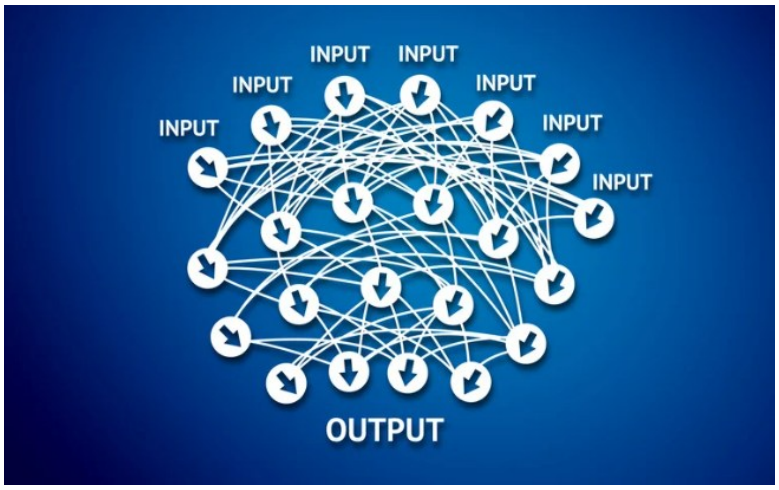
Esta segmentación alimenta directamente las decisiones de producción —cuántas unidades fabricar de cada prenda—, las estrategias de comunicación —qué tipo de mensaje y promoción enviar a cada segmento— y el diseño de la experiencia en tienda —cómo organizar el espacio para cada perfil de cliente—. El resultado es una cadena de moda que produce menos excedente de stock que sus competidores y que personaliza la experiencia de compra a una escala que habría sido imposible sin el análisis de datos a gran escala.

# 5. Asimilación del funcionamiento del Deep Learning y Aprendizaje por Refuerzo.

Si el machine learning clásico —regresión, SVM, árboles de decisión, clustering— ha transformado sectores enteros en las últimas dos décadas, el Deep Learning ha representado un salto cualitativo adicional que ha hecho posibles aplicaciones que hasta hace poco parecían ciencia ficción: sistemas que reconocen la voz con precisión sobrehumana, que generan imágenes y textos indistinguibles de los creados por personas, que juegan al ajedrez o al Go mejor que cualquier ser humano. Comprender sus principios fundamentales es esencial para cualquier profesional del dato en 2026.

## 5.1. Introducción al Deep Learning: Redes neuronales profundas y su aplicación.

El Deep Learning es una rama del machine learning que se apoya en redes neuronales con muchas capas ocultas —de ahí el adjetivo "profundo"—. Lo que hace especial al Deep Learning va más allá de la mayor profundidad de la red: destaca por la capacidad de estas arquitecturas para aprender automáticamente representaciones jerárquicas de los datos. Las primeras capas aprenden características simples y genéricas —bordes en una imagen, fonemas en audio—, las capas intermedias combinan esas características en patrones más complejos —texturas, sílabas—, y las capas finales aprenden representaciones de alto nivel directamente relevantes para la tarea —caras de personas, palabras—.




La imagen muestra una representación conceptual de una red neuronal profunda. La información entra por distintos nodos de entrada, atraviesa varias conexiones internas donde se transforma y finalmente genera una salida. Esta estructura permite al Deep Learning aprender patrones complejos a partir de grandes volúmenes de datos.

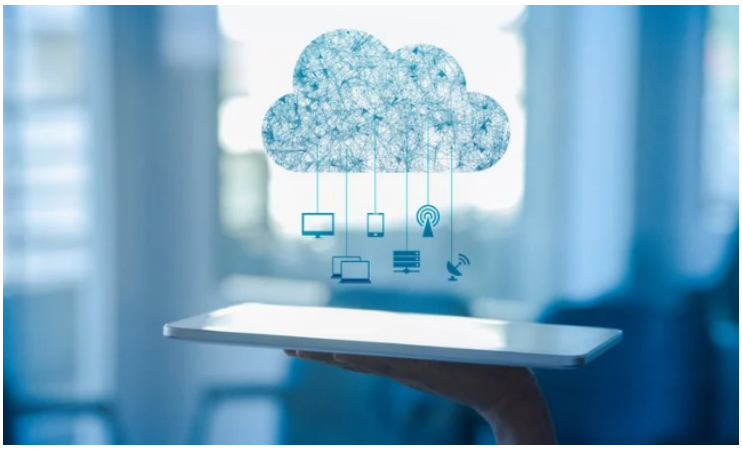
El renacimiento del Deep Learning a partir de 2012 —cuando AlexNet ganó de forma contundente el concurso de reconocimiento de imágenes ImageNet, reduciendo el error a la mitad respecto a los métodos anteriores— fue posible gracias a tres factores que convergieron en el momento adecuado: la disponibilidad de grandes volúmenes de datos etiquetados gracias al Big Data, el aumento masivo de la capacidad de cómputo mediante GPUs —unidades de procesamiento gráfico originalmente diseñadas para videojuegos, pero reutilizadas para el entrenamiento paralelo de redes neuronales—, y el desarrollo de nuevas arquitecturas y técnicas de regularización que hicieron posible entrenar redes más profundas sin que sufrieran el problema del desvanecimiento del gradiente.

Las arquitecturas de Deep Learning más relevantes en la actualidad son las Redes Neuronales Convolucionales —CNN—, especialmente eficaces para el procesamiento de imágenes, las Redes

Neuronales Recurrentes —RNN— y sus variantes como los LSTM —Long Short-Term Memory—, diseñadas para procesar secuencias de datos como texto o series temporales, y los Transformers, la arquitectura que en 2017 revolucionó el procesamiento del lenguaje natural y que es la base de modelos como GPT-4, Gemini, Claude y todos los grandes modelos de lenguaje actuales. En 2026, los Transformers han expandido su dominio más allá del texto y se aplican con éxito a imágenes, audio, vídeo y datos multimodales.

# 5.2. Aprendizaje por Refuerzo: qué es y cómo se utiliza para la toma de decisiones en IA.

El aprendizaje por refuerzo —Reinforcement Learning o RL— es el tercer paradigma principal del machine learning, junto con el supervisado y el no supervisado. A diferencia de los otros dos, no aprende de un conjunto de datos estático, sino de la interacción con un entorno: el agente realiza acciones, el entorno cambia de estado como consecuencia de esas acciones y el agente recibe una recompensa positiva o negativa según el resultado. El objetivo del algoritmo de RL es aprender una política —una función que mapea estados a acciones— que maximice la recompensa acumulada a lo largo del tiempo.



La analogía más intuitiva es la del aprendizaje por ensayo y error: un niño aprende a caminar cayéndose, levantándose y ajustando gradualmente su equilibrio hasta conseguir moverse sin caerse. El aprendizaje por refuerzo formaliza matemáticamente este proceso mediante el marco de los procesos de decisión de Markov —MDP— y lo hace escalable mediante técnicas como Q-learning y sus variantes con redes neuronales profundas, conocidas como Deep Reinforcement Learning.

Las aplicaciones del aprendizaje por refuerzo son especialmente notables en dominios donde hay que tomar secuencias de decisiones en entornos complejos y dinámicos. Los ejemplos más conocidos son los sistemas de juego —AlphaGo y AlphaZero de DeepMind, que superaron a los campeones del mundo en Go, ajedrez y shogi aprendiendo únicamente jugando contra sí mismos—, pero las aplicaciones prácticas son mucho más amplias: gestión de carteras de inversión, optimización de rutas de robots en almacenes, control de sistemas de climatización en centros de datos para minimizar el consumo energético, y personalización de contenidos en plataformas digitales.

ACTIVIDAD 14

Une cada concepto de la columna A con su definición de la columna B escribiendo la letra correspondiente.

Columna A — Conceptos

1. Red Neuronal Convolucional (CNN).
2. Transformer.
3. Aprendizaje por Refuerzo.
4. GPU.

Columna B — Definiciones

A. Unidad de procesamiento gráfico reutilizada para el entrenamiento paralelo de redes neuronales profundas, acelerando el Deep Learning.

B. Arquitectura de Deep Learning especialmente eficaz para el procesamiento de imágenes, que aprende filtros espaciales jerárquicos.

C. Paradigma de aprendizaje en el que un agente aprende a tomar decisiones interactuando con un entorno y recibiendo recompensas o penalizaciones.

D. Arquitectura de redes neuronales basada en mecanismos de atención, base de los grandes modelos de lenguaje actuales como GPT, Gemini y Claude.

EJEMPLO PRÁCTICO - AlphaGo Zero y el aprendizaje por refuerzo: de cero al nivel de dios en 40 días

El juego del Go es considerado uno de los mayores retos para la inteligencia artificial debido a su complejidad combinatoria: el número de posiciones posibles en un tablero de Go supera el número de átomos del universo observable, lo que hace imposible explorar el espacio de jugadas por fuerza bruta.

En 2016, AlphaGo de DeepMind derrotó al campeón mundial de Go Lee Sedol por 4-1, un hito histórico. Pero ese sistema había aprendido de millones de partidas humanas. En 2017, DeepMind presentó AlphaGo Zero, una versión que aprendió a jugar al Go partiendo absolutamente de cero: sin ninguna partida humana, sin ninguna regla estratégica introducida por expertos. Solo le dieron las reglas del juego y lo dejaron jugar contra sí mismo.

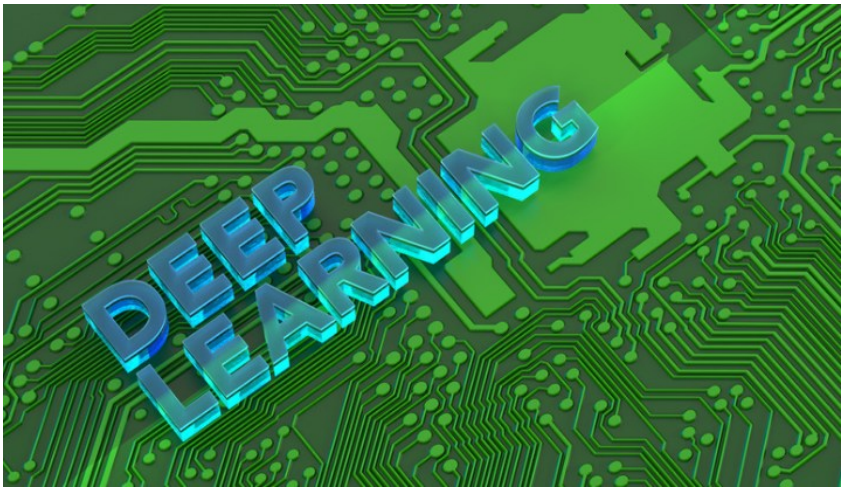
En 72 horas de entrenamiento, AlphaGo Zero superó a AlphaGo original. En 40 días, había alcanzado un nivel de juego que ningún ser humano ha conseguido nunca. El sistema descubrió de forma autónoma estrategias que los maestros de Go habían tardado siglos en desarrollar, e inventó movimientos completamente nuevos que desconcertaron a los expertos humanos.

Lo que hace especialmente relevante este ejemplo para el mundo empresarial es que la misma arquitectura —Deep Reinforcement Learning con búsqueda de árbol de Monte Carlo— se ha aplicado con éxito a problemas como el diseño de chips de silicio, la optimización de la refrigeración de centros de datos de Google y la predicción de estructuras de proteínas. El mismo algoritmo que aprendió a jugar al Go está optimizando las infraestructuras digitales que usamos cada día.

# 6. Comprensión del procesamiento de información no estructurada.

Más del 80 % de los datos que genera el mundo son no estructurados: imágenes, vídeos, grabaciones de audio, documentos de texto, publicaciones en redes sociales. Durante décadas, esta inmensa cantidad de información fue prácticamente inaccesible para el análisis automatizado porque los métodos estadísticos clásicos requieren datos estructurados en tablas con variables numéricas bien definidas.

El Deep Learning ha cambiado la situación de forma radical: hoy existen técnicas maduras y herramientas de código abierto que permiten extraer información de cualquier tipo de dato no estructurado con una precisión que en muchos casos iguala o supera la del ojo o el oído humano.




## 6.1. Imágenes y textos: técnicas utilizadas en el procesamiento de datos no estructurados como imágenes, texto y audio.

En el procesamiento de imágenes, las Redes Neuronales Convolucionales son la herramienta dominante. Una CNN aprende a detectar patrones visuales de forma jerárquica:

- Las primeras capas detectan bordes y gradientes de color;
- Las capas intermedias combinan esos bordes en texturas y formas;
- Las capas finales identifican objetos complejos como caras, coches o tumores.

Las principales tareas de visión artificial son la clasificación de imágenes —decidir qué objeto aparece en la imagen—, la detección de objetos —localizar y clasificar múltiples objetos en la misma imagen—, la segmentación semántica —asignar una categoría a cada píxel de la imagen— y la generación de imágenes, que ha experimentado un auge extraordinario desde la aparición de los modelos de difusión como Stable Diffusion y DALL-E.

En el procesamiento del lenguaje natural —NLP, Natural Language Processing—, la revolución ha venido de la mano de los Transformers y los grandes modelos de lenguaje —LLMs, Large Language Models—. Antes de 2017, el procesamiento de texto se apoyaba en técnicas como la bolsa de palabras —que representa un documento como un vector de frecuencias de palabras ignorando el orden— o los embeddings de palabras como Word2Vec y GloVe —que representan cada palabra como un vector denso en un espacio de alta dimensión donde palabras semánticamente

similares están cerca—. Estas técnicas eran útiles pero limitadas: no capturaban bien el contexto ni la ambigüedad del lenguaje.

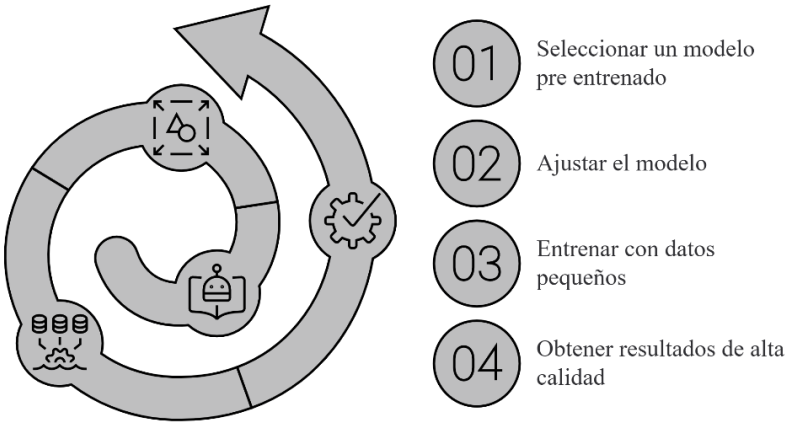
Los Transformers resolvieron este problema mediante el mecanismo de atención: en lugar de procesar el texto de izquierda a derecha de forma secuencial, el modelo puede atender a cualquier parte del texto de entrada al procesar cada palabra, lo que le permite capturar dependencias de largo alcance y resolver ambigüedades contextuales. Sobre esta arquitectura se construyeron los grandes modelos de lenguaje preentrenados —BERT, GPT, T5— que primero se entrenan sobre enormes corpus de texto —centenares de gigabytes de texto de internet, libros y artículos científicos— y luego se afinan para tareas específicas como clasificación de sentimientos, extracción de información, respuesta a preguntas o generación de texto.

## 6.2. Métodos de análisis y extracción de patrones.

Más allá de las arquitecturas de red, existen una serie de técnicas transversales que se aplican frecuentemente en el procesamiento de información no estructurada y que es útil conocer para entender cómo funcionan los sistemas de IA modernos.

El transfer learning —aprendizaje por transferencia— es quizás la técnica más impactante desde el punto de vista práctico. Consiste en tomar un modelo preentrenado en una tarea general —como un modelo de visión entrenado sobre millones de imágenes de ImageNet, o un modelo de lenguaje entrenado sobre todo internet— y ajustarlo para una tarea específica con un conjunto de datos mucho más pequeño. El modelo ya ha aprendido representaciones generales y útiles del dominio; el fine-tuning solo necesita adaptar las capas finales a las particularidades de la nueva tarea. Esto ha democratizado enormemente el uso del Deep Learning: una empresa que quiere detectar defectos en sus piezas de fabricación no necesita millones de imágenes etiquetadas de sus propias piezas; con unos pocos cientos de ejemplos y un modelo preentrenado, puede obtener resultados de alta calidad.

**Proceso de Transfer Learning**




El procesamiento multimodal —combinar diferentes tipos de datos en un mismo modelo— es otra frontera activa. Los modelos modernos como GPT-4 Visión, Gemini y Claude son capaces de

procesar conjuntamente texto e imágenes, lo que abre posibilidades como describir el contenido de una imagen, responder preguntas sobre un gráfico, o extraer información estructurada de facturas o formularios escaneados. En 2026, los modelos multimodales han empezado a incorporar también audio y vídeo, abriendo la puerta a sistemas que entienden el mundo a través de múltiples sentidos de forma simultánea.

ACTIVIDAD 15

Une cada concepto de la columna A con su definición de la columna B escribiendo la letra correspondiente.

Columna A — Conceptos

1. CNN.
2. Transfer Learning.
3. LLM.
4. Mecanismo de atención.

Columna B — Definiciones

A. Componente clave de la arquitectura Transformer que permite al modelo ponderar la importancia de cada parte del texto de entrada al procesar cada palabra.

B. Red neuronal que aprende representaciones jerárquicas de imágenes, desde bordes simples hasta objetos complejos.

C. Técnica que consiste en adaptar un modelo preentrenado en una tarea general para una tarea específica con pocos datos de entrenamiento.

D. Modelo de lenguaje de gran tamaño entrenado sobre enormes corpus de texto, capaz de generar texto, responder preguntas y realizar tareas lingüísticas complejas.

EJEMPLO PRÁCTICO - IA en radiología: cuando el algoritmo detecta cáncer antes que el radiólogo

El cáncer de mama es el cáncer más frecuente en mujeres a nivel mundial. Su detección precoz mediante mamografía es fundamental para mejorar el pronóstico, pero la interpretación de las mamografías es una tarea compleja y exigente que requiere años de formación y que incluso entre radiólogos expertos tiene tasas de error no despreciables: entre el 10 % y el 20 % de los cánceres son perdidos en una primera lectura.

Google Health publicó en 2020 un estudio en la revista Nature en el que presentaba un sistema de IA basado en redes neuronales convolucionales, entrenado sobre más de 90 000 mamografías del sistema sanitario británico NHS y de Kaiser Permanente en Estados Unidos. El sistema superó al diagnóstico promedio de los radiólogos en ambas cohortes, reduciendo los falsos negativos en un 9,4 % y los falsos positivos en un 5,7 % en el conjunto de datos del NHS.

El sistema fue entrenado usando transfer learning a partir de arquitecturas de visión como EfficientNet, previamente entrenadas sobre ImageNet, y ajustadas sobre el conjunto específico de mamografías. El punto clave es que el modelo no sustituye al radiólogo: está diseñado para trabajar como una segunda opinión automática que alerta al especialista cuando detecta una región sospechosa, reduciendo la carga de trabajo y minimizando los errores por fatiga o falta de atención.

En 2026, sistemas similares están aprobados para uso clínico en varios países europeos y en Estados Unidos, y su integración en los flujos de trabajo de radiología es ya una realidad en cientos de hospitales.

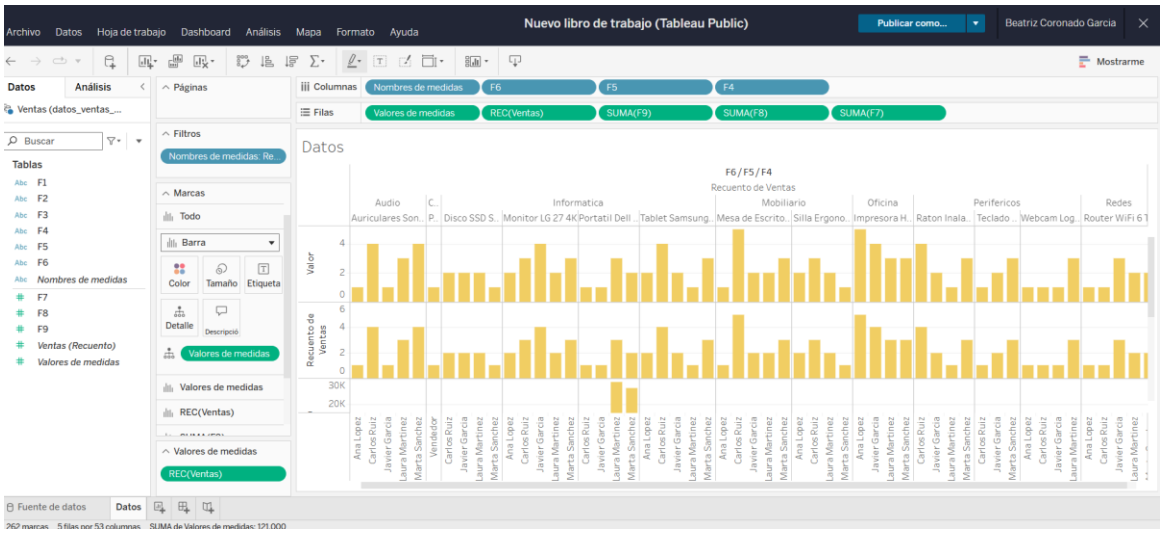
# 7. Conocimiento de técnicas para la visualización de datos.

De nada sirve construir el modelo de machine learning más preciso del mundo si los resultados no se pueden comunicar de forma comprensible a las personas que tienen que tomar decisiones basándose en ellos. La visualización de datos es la disciplina que se ocupa exactamente de este puente: transformar números, predicciones y patrones en representaciones visuales que faciliten la comprensión, revelen tendencias y apoyen la toma de decisiones informadas. En el contexto del Big Data, donde los volúmenes de datos son inmensos y las relaciones entre variables son complejas, la visualización es más necesaria que nunca.

## 7.1. Creación de visualizaciones interactivas: herramientas como Tableau y Power BI para representar datos y resultados.

Las herramientas de visualización de datos para entornos empresariales han evolucionado enormemente en la última década. Dos plataformas dominan el mercado corporativo de inteligencia de negocio con capacidades de visualización avanzada: Tableau y Microsoft Power BI.

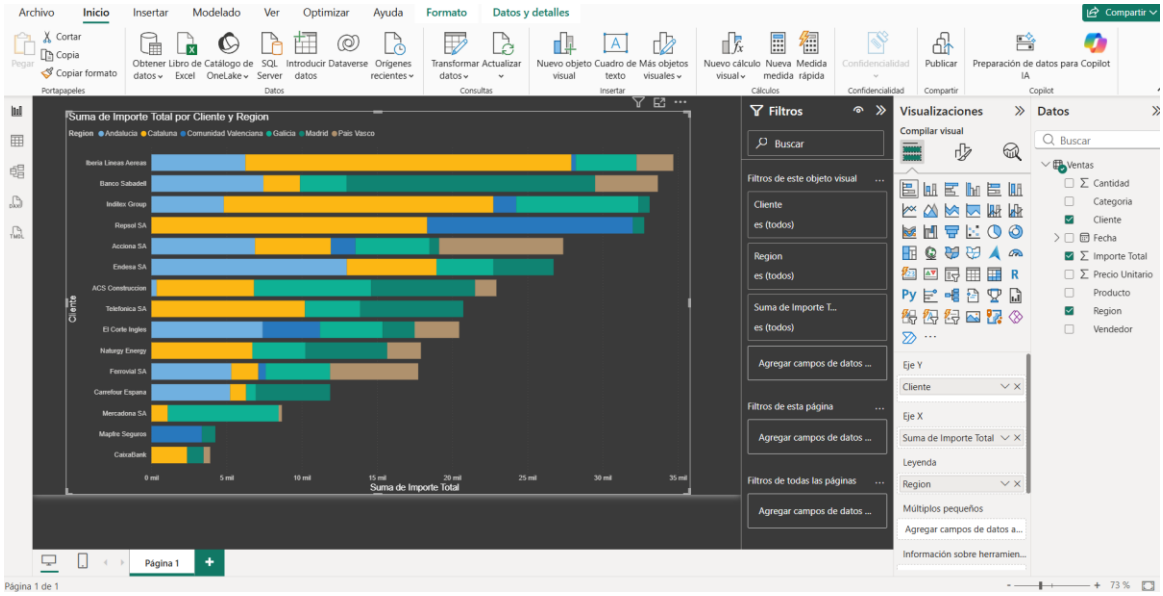
Tableau es una herramienta de visualización interactiva que permite a los usuarios crear gráficos, mapas y dashboards complejos mediante una interfaz de arrastrar y soltar, sin necesidad de escribir código. Su punto fuerte es la flexibilidad y la calidad estética de las visualizaciones, y su capacidad de conectarse a prácticamente cualquier fuente de datos: archivos CSV o Excel, bases de datos relacionales, almacenes de datos en la nube como Snowflake o BigQuery, APIs REST y sistemas de Big Data como Hadoop y Spark. Tableau fue adquirida por Salesforce en 2019 y en 2026 se integra estrechamente con el ecosistema de CRM y datos de clientes de esa compañía. Su punto débil es el precio: las licencias de Tableau son significativamente más caras que las de sus competidores, lo que limita su adopción a grandes organizaciones.




*Interfaz de Tableau*

Microsoft Power BI es la respuesta de Microsoft a Tableau y, gracias a su integración con el ecosistema de Microsoft —Azure, Excel, Teams, SharePoint, Dynamics 365—, ha ganado una

cuota de mercado enorme en empresas que ya usan productos de Microsoft. Power BI Desktop es gratuito para uso individual; las capacidades colaborativas y de publicación en la nube requieren licencia. Su lenguaje de fórmulas DAX —Data Analysis Expressions— permite crear medidas y cálculos complejos directamente en el modelo de datos, y su lenguaje M permite transformar y limpiar datos con gran flexibilidad. En 2026, Power BI ha incorporado capacidades de IA generativa —Copilot para Power BI— que permiten crear visualizaciones y describir datos usando lenguaje natural.



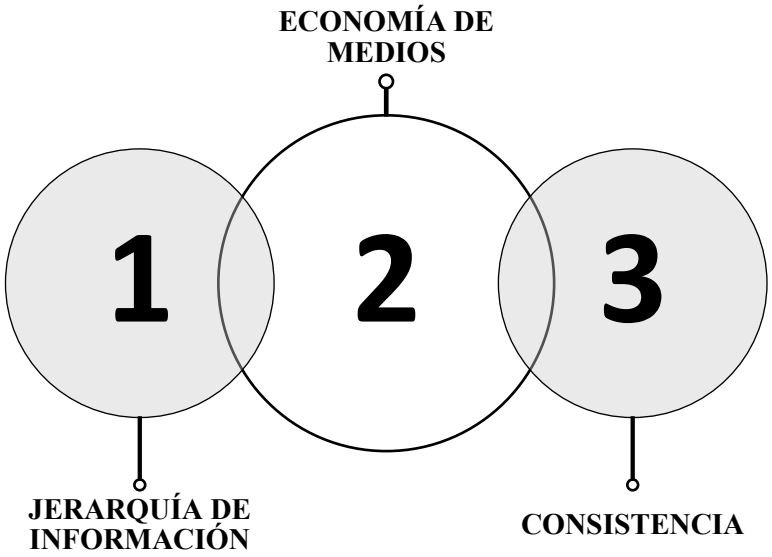

*Interfaz de Power BI*

Más allá de estas dos plataformas comerciales, existen alternativas de código abierto ampliamente utilizadas. Apache Superset es una plataforma de exploración y visualización de datos orientada a entornos de Big Data, con capacidad de conectarse directamente a motores como Presto, Druid o BigQuery. Grafana es la herramienta de referencia para la monitorización de sistemas en tiempo real, especialmente popular para visualizar métricas de infraestructura, logs y series temporales de sensores IoT.

## 7.2. Dashboards: diseño y desarrollo de dashboards para la toma de decisiones.

Un dashboard —cuadro de mando— es una representación visual consolidada de los indicadores clave de rendimiento —KPIs— más relevantes para un responsable de tomar decisiones específico. No es simplemente una colección de gráficos: un buen dashboard está diseñado para responder preguntas concretas de forma inmediata, alertar sobre desviaciones respecto a los objetivos y permitir la exploración interactiva de los datos subyacentes cuando el usuario quiere entender la causa de un problema.

El diseño de un dashboard eficaz responde a principios bien establecidos:




- El primero es la jerarquía de información: los indicadores más críticos deben estar en la parte superior izquierda —donde el ojo humano comienza a leer—, con el contexto y el detalle disponibles en niveles más profundos.
- El segundo es la economía de medios: cada elemento visual debe tener un propósito claro; los elementos decorativos que no aportan información —como los fondos degradados, las sombras innecesarias o los gráficos tridimensionales que distorsionan las proporciones— deben eliminarse.
- El tercero es la consistencia: el mismo tipo de dato siempre se representa de la misma forma y con los mismos colores a lo largo de todo el dashboard.

En la práctica, los dashboards más útiles son los que responden a tres niveles de preguntas: el nivel estratégico —cómo estamos respecto a nuestros objetivos anuales—, el nivel táctico —qué ha pasado esta semana y por qué— y el nivel operativo —qué está pasando ahora mismo y necesita atención inmediata—. Un dashboard bien diseñado permite navegar entre estos tres niveles con clics mínimos y sin perderse en la información.

ACTIVIDAD 16

Une cada concepto de la columna A con su definición de la columna B escribiendo la letra correspondiente.

Columna A — Conceptos

1. Tableau.
2. Power BI.
3. Apache Superset.
4. KPI.

Columna B — Definiciones

A. Indicador clave de rendimiento: métrica crítica que refleja el progreso hacia un objetivo estratégico o operativo.

B. Herramienta de visualización interactiva de alto nivel estético, adquirida por Salesforce en 2019, con conexión a múltiples fuentes de datos.

C. Plataforma de exploración y visualización de datos de código abierto orientada a entornos Big Data, con conexión a motores como Presto y Druid.

D. Herramienta de inteligencia de negocio de Microsoft, integrada con el ecosistema Azure y Office 365, con capacidades de IA generativa en 2026.

EJEMPLO PRÁCTICO - El dashboard de control de pandemia de la Johns Hopkins University

En enero de 2020, cuando el coronavirus empezó a extenderse más allá de China, el Center for Systems Science and Engineering de la Universidad Johns Hopkins lanzó un dashboard interactivo de seguimiento de la COVID-19 que se convirtió en la referencia mundial para gobiernos, medios de comunicación, investigadores y ciudadanos durante los años siguientes.

El dashboard, construido con tecnología ArcGIS de Esri e integrado con fuentes de datos de la OMS, los CDC, las autoridades sanitarias de cada país y repositorios de datos académicos, mostraba en tiempo real el número de casos confirmados, fallecidos y recuperados en cada país, región y ciudad del mundo, actualizado varias veces al día.

Lo que hacía especialmente valioso este dashboard no era solo la cantidad de datos que mostraba, sino las decisiones de diseño que facilitaban su interpretación: el mapa interactivo permitía hacer zoom hasta el nivel de condado en Estados Unidos o de provincia en España; las gráficas de evolución temporal mostraban tanto la escala absoluta como la escala logarítmica para facilitar la comparación entre países con distintos tamaños de población; y los colores —de amarillo a rojo oscuro— reflejaban intuitivamente la gravedad de la situación en cada región.

En el punto álgido de la pandemia, el dashboard recibía más de mil millones de visitas al mes. Fue una demostración en tiempo real de que una visualización bien diseñada puede convertirse en una herramienta de salud pública crítica, democratizando el acceso a información epidemiológica que antes estaba reservada a especialistas.

# 8. Resumen.

En este módulo hemos recorrido el corazón intelectual del ecosistema de datos: las disciplinas, los algoritmos y las herramientas que convierten los datos en conocimiento y en inteligencia. Estos son los siete puntos esenciales que debes retener.

1. La Ciencia de Datos combina estadística, programación y conocimiento del dominio para extraer información accionable de los datos. Es complementaria al Big Data: mientras este proporciona la infraestructura, la Ciencia de Datos aporta los métodos. La IA, y especialmente el Machine Learning, es la herramienta que permite a los sistemas aprender de los datos y generalizar a situaciones nuevas.
2. Python y R son los lenguajes dominantes de la Ciencia de Datos. Python destaca por su accesibilidad y la riqueza de su ecosistema —pandas, Scikit-learn, PySpark—; R, por su profundidad estadística y su uso extendido en entornos académicos y regulados. En 2026, Python lidera en adopción profesional, pero ambos lenguajes siguen siendo relevantes y complementarios.
3. El aprendizaje supervisado aprende de datos etiquetados para predecir categorías —clasificación— o valores numéricos —regresión—. Sus algoritmos más importantes incluyen la regresión lineal y logística, las SVM, las redes neuronales y los métodos de ensemble como XGBoost y Random Forests, estos últimos dominantes en datos tabulares en entornos de producción.
4. El aprendizaje no supervisado descubre estructura en datos sin etiquetar. K-means es el algoritmo de clustering más conocido; PCA es la técnica de reducción de dimensionalidad clásica. Sus aplicaciones más valiosas incluyen la segmentación de clientes, la detección de anomalías y el aprendizaje de representaciones.
5. El Deep Learning, basado en redes neuronales profundas, ha revolucionado el procesamiento de imágenes —CNN—, texto —Transformers y LLMs— y secuencias —LSTM—. El Aprendizaje por Refuerzo permite a los agentes aprender a tomar decisiones óptimas mediante la interacción con un entorno, con aplicaciones desde el juego hasta la optimización industrial.
6. El procesamiento de información no estructurada —imágenes, texto, audio— es una de las fronteras más activas de la IA. El transfer learning ha democratizado el Deep Learning permitiendo adaptar modelos preentrenados a tareas específicas con pocos datos. Los modelos multimodales de 2026 integran texto, imagen y audio en un mismo sistema.
7. La visualización de datos es el puente entre los análisis y las decisiones. Tableau y Power BI dominan el mercado empresarial; Apache Superset y Grafana son las alternativas de código abierto más populares. Un dashboard eficaz responde a preguntas concretas, respeta la jerarquía de la información y elimina todo elemento que no aporte valor.

# 9. Autoevaluación.

PARTE A — Test de opción múltiple.

1. La principal diferencia entre el aprendizaje supervisado y el no supervisado es:

   A) El aprendizaje supervisado requiere más datos que el no supervisado.

   B) El aprendizaje supervisado aprende de datos etiquetados con la respuesta correcta; el no supervisado trabaja con datos sin etiquetar.

   C) El aprendizaje supervisado solo sirve para clasificación; el no supervisado solo sirve para regresión.

   D) El aprendizaje no supervisado es siempre más preciso que el supervisado.

2. ¿Cuál de los siguientes algoritmos pertenece al grupo del aprendizaje no supervisado?

   A) Regresión lineal.

   B) XGBoost.

   C) K-means.

   D) SVM.

3. La arquitectura de redes neuronales que es la base de los grandes modelos de lenguaje actuales como GPT, Gemini y Claude es:

   A) Red Neuronal Convolucional (CNN).

   B) Red Neuronal Recurrente (RNN).

   C) Autoencoder.

   D) Transformer.

4. El Transfer Learning consiste en:

   A) Transferir datos de un sistema de almacenamiento a otro para acelerar el procesamiento.

   B) Tomar un modelo preentrenado en una tarea general y ajustarlo para una tarea específica con pocos datos.

   C) Compartir los pesos de un modelo entre múltiples servidores para acelerar el entrenamiento.

   D) Convertir modelos de un lenguaje de programación a otro para mejorar el rendimiento.

5. En el diseno de dashboards, la 'economia de medios' implica:

   A) Usar el mínimo número de fuentes de datos posibles para reducir costes.

   B) Eliminar todos los elementos visuales que no aporten información relevante para la toma de decisiónes.

   C) Reducir el número de KPIs a uno para facilitar la comprension.

   D) Usar únicamente gráficos de barras para garantizar la máxima claridad.

PARTE B — Completa las frases.

Rellena cada hueco con el termino o concepto más adecuado. Las respuestas se encuentran al final del manual.

1. El sistema de IA de DeepMind que en 2020 predijo la estructura tridimensiónal de más de 200 millones de proteinas, recibiendo el Premio Nobel de Química en 2024, se llama _____.

2. La técnica de reducción de dimensiónalidad clásica que proyecta los datos sobre las direcciones de máxima varianza se denomina _____ (siglas en ingles: _____).

3. En el aprendizaje por refuerzo, el componente que realiza acciónes en el entorno y recibe recompensas o penalizaciones como resultado se denomina _____.

4. La arquitectura de redes neuronales especialmente eficaz para el procesamiento de imagenes, que aprende representaciónes jerárquicas desde bordes simples hasta objetos complejos, se denomina Red Neuronal _____ (siglas: CNN).

5. Microsoft Power BI incorporo en 2026 capacidades de inteligencia artificial generativa bajo el nombre de _____ para Power BI, que permite crear visualizaciónes usando lenguaje natural.

# Aplicaciones del Big data e impacto futuro

El objetivo de este módulo es analizar el impacto que Big Data y la inteligencia artificial tienen en diversos sectores, como salud, finanzas y marketing, explorando casos de uso reales, identificando oportunidades laborales emergentes y comprendiendo cómo estas tecnologías están transformando ocupaciones y procesos en el ámbito profesional.

# 1. Reconocimiento de las aplicaciones del Big Data en el sector público.

Las tecnologías de Big Data e IA ya no son el territorio exclusivo de unas pocas empresas tecnológicas. En 2026 están presentes en los hospitales que diagnostican enfermedades, en los ayuntamientos que optimizan el transporte urbano, en las cadenas de supermercados que predicen la demanda de cada producto, en las organizaciones no gubernamentales que combaten el hambre y en las plantas industriales que reducen su huella de carbono.

Durante décadas, la Administración pública ha acumulado enormes volúmenes de datos sobre los ciudadanos, los territorios y los servicios que gestiona: registros de nacimientos y defunciones, historiales médicos, expedientes tributarios, estadísticas de empleo, datos de tráfico y transporte, registros catastrales. Sin embargo, buena parte de esa información permanecía encerrada en silos departamentales, en formatos propietarios y sin ningún mecanismo que permitiera cruzarla, analizarla y extraer de ella conocimiento útil para la toma de decisiones. El Big Data y las políticas de datos abiertos están cambiando esta situación de forma progresiva pero profunda.

## 1.1. OpenData: ejemplos de aplicación de Big Data en instituciones públicas para la transparencia y eficiencia.

El movimiento Open Data —datos abiertos— parte de un principio simple pero poderoso: los datos generados o financiados con recursos públicos deben estar disponibles para cualquier ciudadano, empresa o investigador que quiera usarlos, de forma gratuita y en formatos que permitan su procesamiento automático. Esta filosofía, que comenzó a consolidarse a nivel institucional a finales de los años 2000, ha dado lugar a portales de datos abiertos en prácticamente todos los países desarrollados.


En España, el Portal de Datos Abiertos del Gobierno —datos.gob.es— alberga en 2026 más de 50 000 conjuntos de datos procedentes de ministerios, comunidades autónomas y ayuntamientos: estadísticas de empleo del SEPE, datos meteorológicos de AEMET, registros de contratación pública, estadísticas judiciales, datos de tráfico de la DGT y centenares de conjuntos de datos más. La Unión Europea gestiona el portal europeo de datos —data.europa.eu— con más de un millón de