rac.2026.114.1.org05
RACSG.2026;114(01):71-90
Recibido: 10/10/2025 | Aceptado: 16/01/2026

ORIGINAL

DATOS MASIVOS: DE LA ESTADÍSTICA A LA INTELIGENCIA ARTIFICIAL

BIG DATA: FROM STATISTICS TO ARTIFICIAL INTELLIGENCE

Daniel Peña[1,2,3]

1. Profesor Emérito de la Universidad Carlos III de Madrid.
2. Académico Numerario en la sección de Ciencias Exactas de la Real Academia de Ciencias Exactas Físicas y Naturales de España.
3. Académico Extranjero de la Academia Nacional de Ciencias Exactas Físicas y Naturales de Argentina.

RESUMEN

El advenimiento de los datos masivos en el presente siglo ha transformado de manera profunda los métodos de aprendizaje empírico. Desde la creación a finales del siglo XIX de la Estadística, como la disciplina científica para el análisis de datos, el conocimiento científico se ha adquirido combinando los datos de experimentos y observaciones con hipótesis sobre su generación. Los avances en la digitalización en este siglo han transformado en datos cualquier tipo de información que podemos manejar en un ordenador, incluyendo sonidos, imágenes y textos, y el aprendizaje sobre estos nuevos datos se ha ido produciendo gradualmente, observando muchos ejemplos y sin realizar hipótesis previas sobre su estructura de dependencia. La disponibilidad de volúmenes masivos de información ha impulsado el avance de la Inteligencia Artificial, concebida hoy como un marco general para el auto- aprendizaje automático de todo tipo de datos basado en la computación intensiva. Sus aplicaciones se encuentran hoy en plena expansión y constituyen un factor de transformación sustantiva de la organización social, económica y cultural contemporánea. El presente trabajo examina este cambio en el proceso de aprendizaje y ofrece un análisis de algunas de sus principales implicaciones.

Palabras clave: Aprendizaje automático; Computación; Redes neuronales; IA generativa.

ABSTRACT

The advent of massive data in the present century has profoundly transformed learning methods. Since the creation of Statistics at the end of the 19th century as the scientific discipline for data analysis, scientific learning has been carried out by combining data from experiments or observations with hypotheses about their generation. Advances in digitization in this century have turned into data virtually any type of representation that can be processed by a computer, including sounds, images, and texts. Learning from these new data has gradually been achieved by observing many examples, without prior hypotheses about their dependency structure. The availability of massive volumes of information has fueled the advancement of Artificial Intelligence, conceived today as a general framework for the automatic self-learning of all types of data based on intensive computation. Its applications are currently expanding and constitute a substantive factor in the social, economic, and cultural transformation of contemporary society. This paper examines this process and offers an analysis of some of its main implications.

Keywords: Machine learning; Computation; Neural networks; Generative AI.

Correspondencia
Daniel Peña
Real Academia de Ciencias Exactas, Físicas y Naturales de España
Calle Valverde, 22 · 28004 · Madrid, España
E-mail: Daniel.pena@uc3m.es

1. INTRODUCCIÓN

Los datos han tenido un papel fundamental en la historia de la ciencia. En la antigüedad los datos fueron inicialmente observaciones dispersas, comenzaron a analizarse en el siglo XVII y se convirtieron en el siglo XIX en una herramienta importante tanto en la investigación científica como en la toma de decisiones del gobierno y las empresas. En la actualidad, los datos masivos ocupan un lugar central para el aprendizaje automático en todas las áreas, impulsando el desarrollo económico y social y el avance de la ciencia. Permiten no solo comprobar teorías, sino, también, generar hipótesis y nuevas líneas de investigación, transformando la ciencia en un proceso cada vez más colaborativo, global y dependiente de la información.

Los datos masivos han transformado la economía del siglo XXI. En la segunda mitad del siglo XX las grandes empresas del mundo por capitalización vendían petróleo (Exxon), automóviles (General Motors), aparatos eléctricos y de telecomunicaciones (ATT) y productos de consumo (Procter&Gamble). En la actualidad, las empresas más importantes del mundo son las que han basado su desarrollo en la explotación de los datos masivos de sus posibles clientes. Estos datos han sido obtenidos de buscadores de internet, redes sociales y plataformas de compras on-line y entretenimiento de videos o audios, y se utilizan para crear publicidad personalizada, (Alphabet con Google y Youtube, Meta, con Facebook, Instagram y WhatsApp), para mejorar las posibilidades de los teléfonos móviles, (Apple), para establecer un comercio electrónico mundial, (Amazon), para desarrollar métodos de cálculo masivo con servidores en la nube, (Microsoft), y para diseñar procesadores especializados en cálculos en paralelo para Inteligencia Artificial, (Nvidia). Sin llegar al tamaño de los gigantes tecnológicos anteriores, otras empresas centradas en utilizar los datos de sus usuarios para hacer recomendaciones y mejorar sus políticas, como Netflix en videos, o Spotify en música, se han convertido en líderes de su sector y acumulan una audiencia mundial de cientos de millones de subscriptores.

En las secciones siguientes de este trabajo se analiza el proceso de cambio del papel de los datos en nuestra sociedad y de los métodos utilizados en su análisis. La sección 2 revisa brevemente la historia de la recogida de datos y su creciente influencia en los desarrollos científicos, económicos y sociales. La sección 3 describe cómo los cambios en este siglo en las tecnologías de la información y comunicación han favorecido e impulsado la generación de datos masivos y su tratamiento analítico. La sección 4 analiza los cambios en los métodos de análisis y aprendizaje con los datos y cómo los nuevos datos digitales han propiciado la sustitución en muchas aplicaciones de la estadística por los métodos de aprendizaje automático (Machine Learning, ML). La sección 5 describe la evolución de la Inteligencia Artificial (IA) y sus potencialidades actuales, con especial atención a la IA generativa, que ya ha comenzado a cambiar nuestras vidas. Finalmente, la sección 6 presenta algunas reflexiones sobre el efecto social de estos cambios en nuestro siglo y sus posibles implicaciones futuras.

2. LOS DATOS EN LA CIENCIA Y EN LA SOCIEDAD HASTA NUESTRO SIGLO

A lo largo de la historia, la humanidad ha ido generando conocimiento observando fenómenos deterministas: aquellos que, repetidos bajo las mismas condiciones, producen siempre los mismos resultados, como construir una herramienta. Este es también el proceso de aprendizaje del mundo, por repetición, que utilizan los niños. Sin embargo, la comprensión de los fenómenos aleatorios —donde bajo circunstancias similares, los resultados pueden variar al azar— parecía inalcanzable y hasta el renacimiento los resultados de estos fenómenos se atribuían generalmente a la voluntad divina. Ejemplos de experimentos aleatorios son lanzar un dado y observar el sexo de un recién nacido, la frecuencia de lluvias o la aparición de desastres naturales. Un reflejo de la dificultad humana de comprender el azar se manifiesta en los niños, que buscan siempre una causa única determinista para los fenómenos que no comprenden, como los cambios de humor de sus padres. Esto explica que el estudio riguroso de los fenómenos deterministas, que llevó al avance de la física y la matemática, haya precedido al desarrollo de la estadística.

Los juegos de azar, como manifestación temprana de lo aleatorio, surgieron hace más de cuarenta mil años, según David (1962), quien estudió la abundante presencia del hueso astrágalo de oveja o ciervo, precursor del dado, en las excavaciones arqueológicas. En 3000 a.C., los dados se empleaban tanto en juegos como en ceremonias religiosas (Hasofer, 1967; Kendall, 1956) y los oráculos de Delfos los utilizaban para predecir el futuro. No obstante, aunque muchas leyes físicas y resultados matemáticos se fueron consiguiendo en los primeros quince si-

glos de la era cristiana, los sucesos aleatorios no se consideraron objeto de estudio científico hasta el Renacimiento.

Tenemos que esperar al siglo XVI para que matemáticos y físicos italianos comenzaron a analizar los resultados de experimentos aleatorios simples, como los juegos de azar. En 1553, G. Cardano escribió *Liber de ludo aleae*, donde explicó, por simetría, la equiprobabilidad de las caras de un dado, y, pocos años después, Galileo, en su correspondencia con un jugador, mostró por qué es más difícil obtener una suma de 9 puntos al lanzar tres dados que obtener 10, razonando que, de las 216 combinaciones posibles, 25 generan un 9 y 27 un 10. El estudio de los juegos de azar evolucionó lentamente hasta que a mediados del siglo XVII se formalizó con la introducción de conceptos como la esperanza matemática, o esperanza de ganancia en una apuesta, y el de distribución de probabilidades, o recuento de los sucesos posibles en un experimento aleatorio, y su frecuencia esperada de aparición a largo plazo. Por ejemplo, supongamos que apostamos 10 euros al suceso "número par al tirar un dado" y si sucede recibimos 20 euros, lo que ocurre con probabilidad, 0,5, y en caso contrario perdemos la apuesta, también con probabilidad 0,5. La esperanza matemática de este juego se calcula sumando los productos de las probabilidades por los posibles resultados, es decir, 0,5 (20-10)+0,5(-10)=0. Esta apuesta es equilibrada y, jugando muchas veces, esperamos un resultado neutro, es decir, perder o ganar un poco, con igual probabilidad.

El cálculo de probabilidades surge en la célebre correspondencia entre Pascal y Fermat en 1654 en Francia para resolver el "problema de los puntos" que les planteó un jugador famoso de su época: Chevalier de Meré. Su pregunta fue cómo repartir las apuestas de un juego que debe interrumpirse antes de finalizar, lo que ocurría con frecuencia al estar el juego entonces prohibido en Francia. La solución propuesta por estos científicos, repartir lo apostado en proporción a la ganancia esperada de cada jugador dada su situación en el momento de la interrupción, sentó las bases para el análisis de otros muchos fenómenos aleatorios y se considera el inicio del estudio de las probabilidades como una parte de las matemáticas. Este avance tuvo influencias sociales, con la creación de los primeros seguros marítimos para cubrir accidentes y la aparición de las primas de seguro, que deben ser mayores que la esperanza matemática del riesgo asegurado para que la compañía obtenga beneficios.

En esa misma época, en 1662, John Graunt publicó en Inglaterra su obra Observations, que fue pionera en la aplicación de razonamientos estadísticos a datos demográficos. En este trabajo Graunt obtuvo una estimación de la población inglesa a partir de muestras y calculó, por primera vez, tasas de mortalidad por edad y frecuencias de nacimientos según el sexo, iniciando la demografía. Poco después, en 1687, Newton publica los *Principia*, que presenta una explicación global del universo mediante leyes físicas capaces de generar predicciones verificables mediante mediciones. La teoría de Newton desencadenó un gran interés por la recogida de datos físicos, especialmente astronómicos mediante telescopios, y durante los siglos XVIII y XIX gran parte de las investigaciones empíricas en Física y Astronomía se dedicaron a contrastar sus leyes.

Un problema clave en las mediciones obtenidas mediante telescopios es el tratamiento de los errores de medición: al medir repetidamente una misma magnitud los resultados no eran idénticos, como consecuencia de los pequeños cambios en las condiciones de medición. Surgió entonces el problema de cómo combinar estas observaciones de la misma magnitud para obtener una estimación más precisa. Este desafío condujo al nacimiento de la teoría de errores, a comienzos del siglo XIX. En *Théorie Analytique des Probabilités*, Laplace introdujo en 1818 una definición explícita de probabilidad y comparó métodos para combinar varias mediciones de una misma magnitud, como la media o promedio de los datos, y la mediana (que los ordena de mayor a menor y elige el del centro). Por su parte, Gauss desarrolló la distribución normal, con su famosa forma de campana, como modelo de los errores de medida en astronomía y propuso un método para ajustar muchas mediciones distintas de una trayectoria a una ecuación lineal que las explique, haciendo mínima la desviación entre lo observado y la ecuación, que es el método de mínimos cuadrados, propuesto también por Legendre en 1805.

Posteriormente, en 1846, Quetelet aplicó estos principios al estudio de datos humanos, sociológicos y económicos y propuso la idea del "hombre medio" como representante de una población, avanzando hacia una teoría general del análisis de datos con incertidumbre en las ciencias sociales. En el siglo XIX la probabilidad se convirtió en la base de la actuaría (ciencia de seguros y pensiones) y se inició el desarrollo de la teoría del riesgo y la fijación de precios en seguros y mercados financieros. Aunque los censos

de población ya se hacían en el imperio romano, durante la segunda mitad del siglo XIX muchos estados comienzan a recoger de forma periódica datos sobre sus ciudadanos y sobre la situación social y económica del país y a mediados de siglo muchos países establecen instituciones con este objetivo. Por ejemplo, en España se crea en 1857 la Junta General de Estadística, precedente del actual Instituto Nacional de Estadística. También aparece la necesidad de unificar los instrumentos de medida y, en 1889 en la primera Conferencia General de Pesos y Medidas en París, se definen las medidas básicas del sistema métrico decimal, como el metro y el kilogramo.

La Estadística se consolidó como la ciencia dedicada al análisis y la interpretación de los datos a finales del siglo XIX, impulsada por el interés de contrastar la teoría de la evolución de Darwin (*El origen de las especies*, 1859). A diferencia de la física newtoniana, esta teoría no ofrece explicaciones deterministas sobre hechos observables, sino que explica los cambios experimentados en promedio por una población de seres vivos para adaptarse a un entorno cambiante. Permite, además, hacer predicciones estadísticas, sobre los cambios futuros. Galton, primo de Darwin, desempeñó un papel fundamental en el desarrollo de métodos estadísticos para contrastar la teoría de la evolución: en *Natural Inheritance* (1889) introdujo, entre otros conceptos, la regresión entre dos variables, analizando las estaturas de padres e hijos. Descubrió la tendencia de los descendientes a acercarse al promedio poblacional: los padres muy altos tienen hijos altos, pero, en promedio, más bajos que sus padres y viceversa, fenómeno que bautizó como regresión a la media. Este hallazgo permitió explicar la estabilidad relativa de los rasgos poblacionales a lo largo del tiempo. El interés de Galton lo llevó a financiar la creación del primer departamento de Estadística en la Universidad de Londres y durante la primera mitad del siglo XX esta disciplina se expandió principalmente en Inglaterra.

Una contribución fundamental a la recogida de datos mediante experimentación y su análisis es debida al estadístico británico R. A. Fisher, que desarrolló el diseño de experimentos para datos agronómicos en el primer tercio del siglo XX. Fisher observó que el método tradicional de experimentación, comparar el efecto de una variable sobre una población, por ejemplo, de un fertilizante sobre un cultivo, con los obtenidos en experimentos anteriores donde esa variable no estaba presente, era ineficiente, al no controlar que las condiciones experimentales sean idénticas en ambos casos. En su lugar, propuso comparar el efecto de la variable con los obtenidos en una población idéntica no sometida a dicha variable. Fisher estudió con detalle cómo crear grupos de control lo más idénticos posibles en agricultura, donde en uno se aplica el tratamiento y en otro no. El trabajo de Fisher tuvo un gran impacto e impulsó la realización de experimentos más eficaces que avanzaron el conocimiento en agricultura, biología y ciencias de la salud. Por ejemplo, en medicina, para ver el efecto de un tratamiento sobre una población tenemos que conseguir un conjunto de individuos homogéneos respecto al efecto de la variable, dividir al azar los individuos de ese conjunto en dos grupos, donde uno recibe el tratamiento y el otro no y actúa como control para comparar los resultados. Estos ensayos aleatorizados han contribuido mucho al avance de la medicina, como veremos más adelante.

Fisher estableció también una metodología general para aprender de los datos experimentales. Se inicia con un modelo conceptual del problema que se traduce en un conjunto de hipótesis sobre los resultados esperados. A continuación, se recogen los datos, se estiman las magnitudes o parámetros de interés y se contrasta si las predicciones de la teoría explican los datos observados. Si es así, la teoría establecida se considera provisionalmente hasta encontrar otra mejor. En caso contrario, se reformula la teoría para que concuerde con los datos y se inicia un nuevo proceso de contrastación con nuevos datos. Este proceso continuo, bien descrito por Box (1976), ha constituido el método científico de aprendizaje desde principios del siglo XX.

Las ideas de Fisher impulsaron la aplicación de la estadística y la recogida de datos en todos los campos. George Box, yerno de Fisher, extendió estas ideas a procesos químicos y continuos, poniendo énfasis en el aprendizaje iterativo basado en experimentos secuenciales para encontrar el punto óptimo de funcionamiento de un proceso, iniciando la experimentación en procesos industriales como método fundamental de mejorar su rendimiento y eficacia. Sus ideas se aplican ahora para decidir los anuncios que nos aparecen en nuestros teléfonos móviles.

En las ciencias de la salud el análisis de datos contribuyó al avance de la medicina y al continuo aumento de la esperanza de vida durante el siglo XX, que ha cambiado de 31 años a principios de siglo, por la alta mortalidad infantil y las enfermedades infecciosas, a 48 años en los años 50, por los antibióticos, la vacu-

nación masiva y la mejora de la higiene, y a 67 años a finales de siglo, más que doblando la vida esperada al inicio del siglo, véase Riley (2001). La estadística ha tenido un papel central en este proceso de mejora de la salud en el siglo pasado y daremos como ilustración solo dos ejemplos. A finales del siglo XIX la estadística Florence Nightingale (véase Bostridge, 2008) fue pionera en demostrar con datos y gráficos recogidos durante la guerra de Crimea en 1855 la importancia de la higiene para la salud, y fundó el cuerpo de enfermeras en el Reino Unido. Para mejorar los primeros procesos de vacunación se utilizaron datos de series temporales y, también, para controlar los brotes de enfermedades infecciosas (Greenwood and Yule, 1920), haciendo progresivamente más eficientes los sistemas sanitarios preventivos.

Es ampliamente reconocido que desde mediados del siglo XX, y hasta la actualidad, la utilización de la estadística para mejorar la salud ha sido decisiva. Ilustremos este hecho con solo tres ejemplos. Austin Bradford Hill mostró la importancia de complementar la intuición médica con datos rigurosos de ensayos clínicos aleatorizados, como el que dirigió en 1948 sobre la tuberculosis en el Reino Unido, y demostró la importancia de los ensayos clínicos para evaluar tratamientos en medicina. Posteriormente, Richard Doll y Austin Bradford Hill, con el trabajo que realizaron en los años 50 para probar la relación causal entre fumar y cáncer, demostraron la importancia de los estudios epidemiológicos basados en muestras grandes para identificar factores de riesgo y buscar relaciones causales. Estos estudios han cambiado para siempre el enfoque de la salud pública. Finalmente, el estadístico Cox (1972) tuvo un papel central en el desarrollo de modelos estadísticos para estimar con precisión la supervivencia de un enfermo a un tratamiento. Desde entonces, la medicina basada en la evidencia (MBE), es decir en los datos, ha contribuido de forma decisiva a mejorar la salud y aumentar la esperanza de vida en todo el mundo. En este siglo han tenido un gran papel los procedimientos de meta-análisis, donde se comparan muchos estudios realizados en centros distintos para obtener conclusiones de todos ellos, véase por ejemplo Borenstein et al (2021).

La aparición del ordenador a mediados del siglo XX impulso nuevas posibilidades de cálculo para el análisis de la información disponible y, también, introdujo la posibilidad de generar nuevos datos para aprender de sistemas mediante simulación, por ejemplo, con el método de Monte Carlo impulsado por Von Neumann. En la segunda mitad del siglo XX el progreso computacional de los ordenadores hizo viable construir modelos estadísticos más complejos para prever, clasificar y agrupar datos y los métodos estadísticos se comienzan a utilizar habitualmente en el campo de la salud, de la economía, y de la industria y la administración, generando pequeños bancos de datos de uso público para la docencia y la investigación. Su importancia en las ciencias económicas fue reconocida con el primer premio Nobel de Economía otorgado conjuntamente en 1969 al estadístico Ragnar Frisch y al físico Jan Tinbergen, por su trabajo en el desarrollo de modelos dinámicos estadísticos para analizar datos económicos.

El aumento paulatino del tamaño de los datos disponibles en la segunda mitad del siglo XX inicia un proceso de reemplazamiento de las hipótesis estadísticas clásicas, que eran imprescindibles para realizar estimaciones efectivas con muestras pequeñas, por métodos que permitan más flexibilidad en el análisis de datos, como los llamados métodos no paramétricos, que tratan de extraer la información de los datos con un mínimo de hipótesis adicionales. Por otro lado, se aprovecha la posibilidad de generar datos simulados, como en el procedimiento Bootstrap, o estimación autosuficiente, propuesto por B. Efron, (Efron, 1979) para calcular la precisión de un estimador por simulación de nuevos datos generados de una muestra, en lugar de utilizar hipótesis matemáticas rígidas.

Los datos disponibles comienzan a aumentar de forma vertiginosa a finales del siglo XX con la aparición de Internet y los primeros sensores capaces de mediciones automáticas. Timothy Berners-Lee a finales del siglo XX inventó la World Wide Web, haciendo posible el uso de Internet para cualquiera con acceso a un ordenador. La aparición de Google en 1998, que se convirtió en el buscador más utilizado al comenzar el siglo XXI, contribuyó a hacer disponibles en internet una cantidad creciente de datos, muchos generados por los propios usuarios de la web mediante teléfonos inteligentes, redes sociales y sensores. Los sensores digitales conectados a los ordenadores se introdujeron en los años 70 del siglo pasado pero el desarrollo de internet y de las redes de comunicación ha llevado en este siglo al llamado *Internet de las cosas*, con miles de millones de sensores activos recogiendo información en todos los campos.

Esta nueva cantidad de datos, obtenidos no solo por mediciones sino también por imágenes, audios o texto digitalizados, forma el llamado Big Data o los da-

tos masivos. En la sección siguiente comentaremos cómo estos datos digitales y masicos han revolucionado la forma de aprender de los información empírica disponible.

Podemos concluir que la disponibilidad de datos ha condicionado históricamente los métodos de análisis. En la Figura 1 se resume el breve recorrido histórico realizado sobre los avances en el tratamiento de los datos hasta finales del siglo XX.

nas alcanzaba unos pocos **Exabytes** $(1\text{EB} = 10^{18}\text{B})$. En la actualidad, más de la mitad de la población de la tierra tiene acceso diario a internet y su uso genera cada día unos 400 millones de TeraBytes $(1\text{TB} = 10^{12}\text{B})$ de información que equivale a un millón de veces todos los libros editados jamás en el mundo (unos 400TB).

Para poner estos datos en una perspectiva más intuitiva, la cantidad diaria de datos generada hoy equivale a que cada ser humano (8.000 millones aproxi-

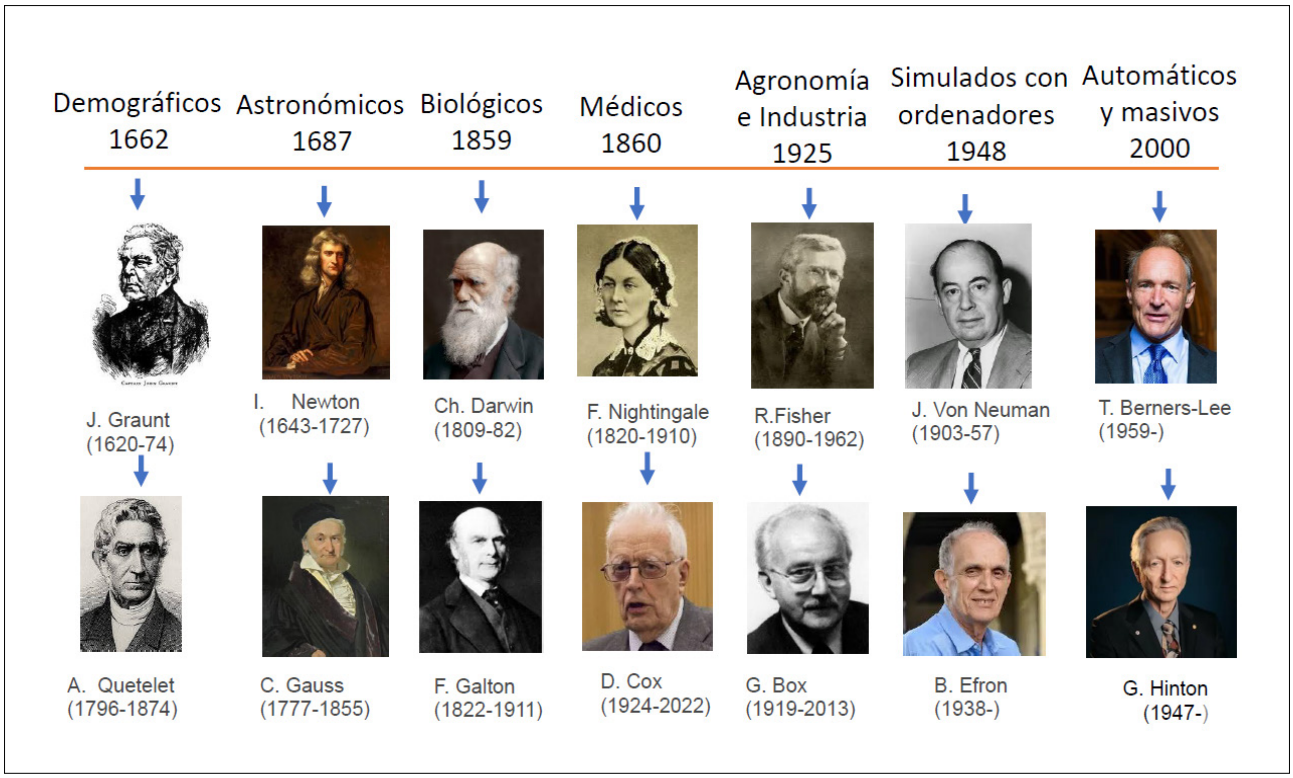

Figura 1. Los datos impulsan el aprendizaje en la ciencia empírica. Fuente: elaboración propia.

3. LOS DATOS MASIVOS Y SU PROCESADO

Desde el inicio del siglo XXI, el volumen de información digital creada, almacenada y compartida ha crecido a un ritmo exponencial, transformando radicalmente la sociedad, la economía y la tecnología. La capacidad de un ordenador para guardar información se mide en Bytes (B), donde cada Byte puede guardar un carácter, como una letra o un dígito. En el año 2000 la mayoría de la información aún se almacenaba en medios físicos, como papel, cintas o discos ópticos, y el total de datos digitales generados a nivel mundial apemadamente en la actualidad) produjera cada día una información igual a dos veces y media todos los libros de la biblioteca del congreso de EEUU (unos 20TB). Hoy, más del 90% de la información mundial existe exclusivamente en formato digital, y, una gran parte de ella, es generada automáticamente por máquinas.

Nuestra capacidad de guardar datos en un ordenador personal, o en un teléfono móvil, se ha multiplicado por 1000 en este siglo. El disco duro de un PC en el año 2000 podía almacenar 10 GigaBytes $(1\text{GB} = 10^{9}\text{B})$, que es el espacio que ocuparían 10.000 libros de unas 300 páginas de texto, (un libro digital promedio

ocupa entre 1 MB y 3 MB, $(1\text{MB} = 10^6\text{B})$ pero esto varía mucho, desde menos de 1 MB para libros de texto puro, hasta más de 200 MB). Hoy tenemos discos duros para PC de 36 TB que nos permiten almacenar todo el contenido de cualquiera de las más grandes bibliotecas del mundo.

Además, se han producido avances espectaculares en la velocidad de transmisión de datos y de cálculo. Vamos a referirnos aquí únicamente a los avances de carácter general que nos afectan como usuarios, (de internet y de ordenadores personales) y no a los desarrollados en centros punteros de computación, que son mucho mayores que los indicados a continuación.

Respecto a la transmisión de datos, hemos pasado de una velocidad promedio en el año 2000 de un megabyte por segundo a un gigabyte en la actualidad, lo que supone multiplicar por mil la velocidad media de transmisión. Pero el cambio más importante es en la velocidad de computación, que se ha multiplicado por 1.000.000. La velocidad de un PC en el año 2000 era de unos $2\text{x}10^9$ flops, es decir podía realizar dos mil millones de operaciones de coma flotante en un segundo. Tomando de nuevo como referencia una población de la tierra de ocho mil millones de personas, la capacidad de cálculo de un PC en el año 2000 era equivalente a tener a ¼ de la población mundial $(8\text{x}10^9 \text{ personas})$ realizando una operación cada segundo con una calculadora manual. Hoy podemos utilizar un ordenador PC CES (NVIDIA) con una velocidad de 300TFLOP= $3\text{x}10^{15}$ flops, equivalente a la capacidad de cálculo de todos los habitantes de 1 millón de planetas como la tierra.

La combinación de abundancia de datos, gran capacidad de almacenaje y rapidez de transmisión y computación ha transformado los métodos de análisis. Los métodos estadísticos clásicos se crearon para aprender "artesanalmente" de los datos disponibles mediante su análisis cuidadoso y se han ido adaptando gradualmente al aumentar el volumen de datos para construir reglas de predicción y clasificación automáticas. El objetivo principal de un análisis estadístico es comprender la estructura de las variables y las relaciones entre ellas, y utilizar después esa relación para prever o clasificar. Los métodos están enfocados para trabajar con datos agregados, donde las relaciones entre las variables suelen ser lineales, o con desviaciones pequeñas de la linealidad, y los modelos más utilizados son de fácil comprensión e interpretación.

Los actuales datos masivos son de carácter muy desagregado y su análisis no aspira a comprender la relación entre las variables, que puede ser muy compleja, por ejemplo, entre los píxeles de una imagen, sino generar métodos automáticos para la clasificación y la predicción. Este es el objetivo de los métodos creados en este siglo con el nombre de aprendizaje automático, o Machine Learning, que veremos en la sección siguiente.

4. EL APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING, ML)

Se conocen como métodos de aprendizaje automático (ML) los desarrollados para analizar datos masivos basados en alta computación que se iniciaron a finales del siglo XX. Estos métodos ocupan hoy un lugar central en el desarrollo de la inteligencia artificial (IA). Inicialmente abordaban tres problemas principales. El primero es la predicción: queremos prever una variable respuesta en función de un conjunto de variables explicativas de la misma. Por ejemplo, la duración de una enfermedad en función de datos clínicos y las condiciones del paciente, o las ventas de ropa de abrigo en función de las ventas pasadas y de la temperatura. El segundo es la clasificación supervisada o discriminación, que aparece para clasificar un objeto en un conjunto de clases bien definidas. Por ejemplo, disponemos de muchas fotos etiquetadas como de hombres o de mujeres, y queremos asignar una nueva foto de forma automática a uno de los dos grupos. El tercero es la clasificación no supervisada, llamada en estadística métodos de agrupamiento o clustering. En ese caso tenemos elementos diversos y queremos encontrar cuántos grupos existen. Por ejemplo, dado un conjunto de fotografías queremos clasificarlas en grupos de forma automática por su semejanza, de manera que los grupos finales encontrados pueden incluir de forma predominante a personas, animales, paisajes, objetos, etc.

Para resolver estos problemas se han aplicado distintas herramientas, muchas de ella desarrolladas en estadística, pero también se han creado nuevos procedimientos de aprendizaje especialmente adaptados a los datos masivos de imágenes y textos, que no habían sido objeto de análisis estadísticos detallados. Véase Peña (2025a) para una revisión de métodos de aprendizaje automático para problemas económicos y empresariales. En este trabajo vamos únicamente a comentar los dos métodos más importantes de

aplicación general en todos los campos para resolver problemas de análisis de datos masivos. Uno especialmente orientado a la clasificación y el segundo a la predicción, aunque ambos pueden adaptarse, como veremos, para casi cualquier aplicación.

4.1 Las máquinas de vector soporte (SVM)

Una herramienta muy importante para la clasificación supervisada o discriminación son las máquinas de vector soporte, Support Vector Machines (SVM) en inglés, que fueron introducidas por Vapnik y sus colaboradores (Cortes y Vapnik, 1995) a finales del siglo pasado. Este método aporta tres novedades importantes con relación a los métodos previos existentes. La primera es que puede aplicarse para clasificar cualquier tipo de datos digitalizados, sin necesitar hipótesis respecto a las características de las variables. La segunda es que la regla de clasificación entre dos grupos se construye utilizando solo los datos que podrían ser dudosos entre ambos, en lugar de considerar todos los datos, mejorando la clasificación global. La tercera es que hace posible clasificar con reglas de separación no lineales entre grupos: si no es posible separar los datos con una regla lineal, el problema se lleva a un espacio de mayor dimensión, introduciendo funciones no lineales de las variables, para hacer posible esta clasificación (Boser et al. ,1992).

Esta tercera idea se ilustra en la figura 2. En la parte izquierda tenemos elementos de dos clases, círculos rojos y estrellas azules, definidos por dos variables (x,y). Estos elementos no pueden separarse con una recta horizontal o vertical que utilice solo una sola variable, pero sí es posible separarlos utilizando las dos variables, como índica el gráfico. Si no fuese posible separarlos en dos dimensiones podemos introducir más variables y trasladar los datos a espacio de mayor dimensión mayor donde podamos construir una regla lineal de clasificación. Esta situación se ilustra en el lado izquierdo de la figura, donde vemos que no es posible separar los puntos rojos y azules con una línea recta en el plano, pero si llevamos los puntos a un espacio de dimensión tres entonces sí que existe un plano que permite una perfecta separación de ambos conjuntos.

4.2 Las redes neuronales (NN)

Las redes neuronales tienen una larga historia que se inicia a mediados del siglo pasado, pero se han ido generalizando y ampliando hasta convertirse en la herramienta central para resolver problemas en inteligencia artificial. La idea inicial de una red neuronal es imitar el funcionamiento del cerebro. Según descubrió Ramón y Cajal, nuestro órgano de pensamiento está formado por neuronas conectadas entre sí. Rosenblatt (1958), un psicólogo americano, concibió un modelo de aprendizaje, el perceptrón, que recibe un conjunto de p señales de entrada, que representaremos por los símbolos $x_1, x_2, \ldots, x_p$, y obtiene una respuesta o variable de salida. Esta salida se construye en dos pasos. En el primero, las variables de entrada se combinan de forma lineal, es decir, se crea una nueva variable resumen de todas ellas dando pesos, $b_1, b_2, \ldots, b_p$, a cada uno de las variables de entrada y sumándolos para obtener la variable $h = b_o + b_1x_1 + \ldots, x_pb_p$, donde el parámetro b_o se

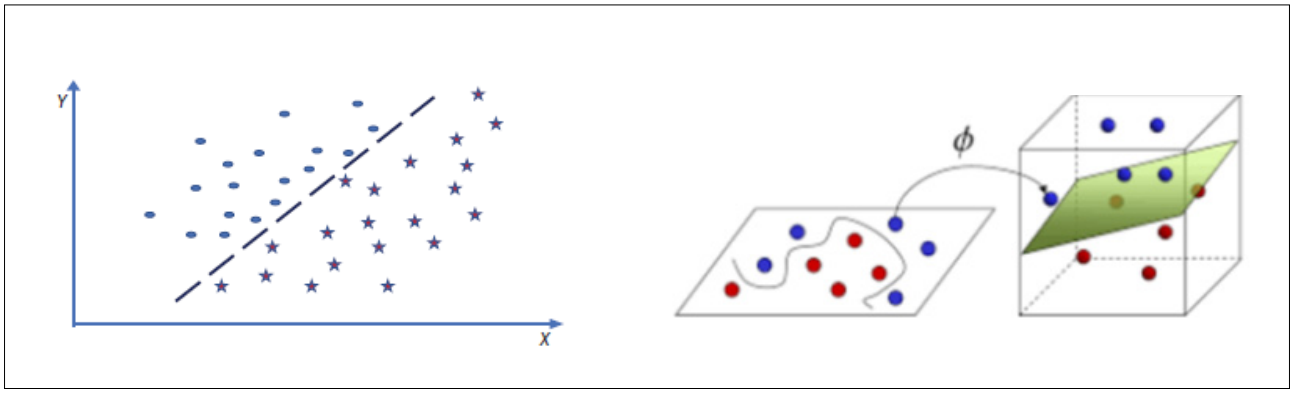

Figura 2. Dos ejemplos de separación de datos bidimensionales. En la parte izquierda es posible una separación lineal. En la derecha la separación en el plano es no lineal pero al aumentar la dimensión y situarnos en un espacio de dimensión tres podemos separarles con un plano.

denomina sesgo y los parámetros b_i determinan el peso de cada variable de entrada x_i en la variable resumen, h. El segundo paso obtiene la variable respuesta, y, aplicando una función no lineal a la variable resumen de la entrada h, y=f(h). Esta función no lineal suele llamarse función de activación, y determina el tipo de salida de la neurona.

La figura 3 presenta dos ejemplos habituales de funciones de activación. La elegida en el perceptrón fue la función logística, utilizada en estadística unos años antes por Berkson (1944) para la clasificación en problemas médicos, en un modelo idéntico al perceptrón con una única neurona utilizado para clasificar un paciente o prever la probabilidad de una enfermedad en función de un conjunto de variables explicativas del paciente, $(x_1, x_2, \ldots, x_p,)$. Otra función muy utilizada es la función lineal truncada o ReLU, rectified linear unit, donde la salida toma el valor cero si la variable de entrada es negativa o cero y el valor de la entrada si esta es positiva. De esta manera, si la variable resumen de las entradas en la neurona es negativa, la neurona no esta activa y no produce respuesta, mientras que si es positiva la salida es la variable resumen sin modificar. Observemos que el valor del parámetro b_o , que se estima con los datos como los otros parámetros, va a determinar a partir de qué punto la variable resumen es directamente la salida de la neurona. Para problemas de clasificación se utiliza mucho la función logística, o sus variantes, mientras que para problemas de predicción es más común en la actualidad la función ReLU, que aproxima una función no lineal por tramos lineales, de forma similar a los splines introducidos en estadística por Wahba, (1990).

Los pesos de la neurona, $b_0, b_1, b_2, \ldots, b_p$, se calculan iterativamente minimizando el error cometido. Por ejemplo si disponemos de n valores de dos variables de entrada $x(1,i), x(2,i)$, y de la variable de salida, $y(i)$, donde i=1,...,n y suponemos una función de activación logística los pesos se determinan tomando un valor inicial para los pesos, calculando el error cometido, viendo que cambios en los pesos reducen el error y modificando su valor para reducir el error. Este ciclo se repite hasta que no sea posible mejorar.

Una red neuronal artificial, RNA, es un desarrollo del perceptrón combinando varias neuronas para construir una respuesta con todas ellas. La figura 4 representa un ejemplo con una red de tres neuronas. Cada una de las ellas produce una variable resumen, (h_i),a la que se aplica la función de activación para formar la variable salida de la neurona y las salidas de las tres neuronas se combinan linealmente para proporcional la previsión de la variable y.

En resumen, una red neuronal artificial, RNA, se basa en dos principios. El primero es resumir todas las entradas a cada neurona por una variable resumen que es una combinación lineal de todas ellas. Esta idea proviene de la estadística y es la base del análisis factorial y de los métodos de reducción de la dimen-

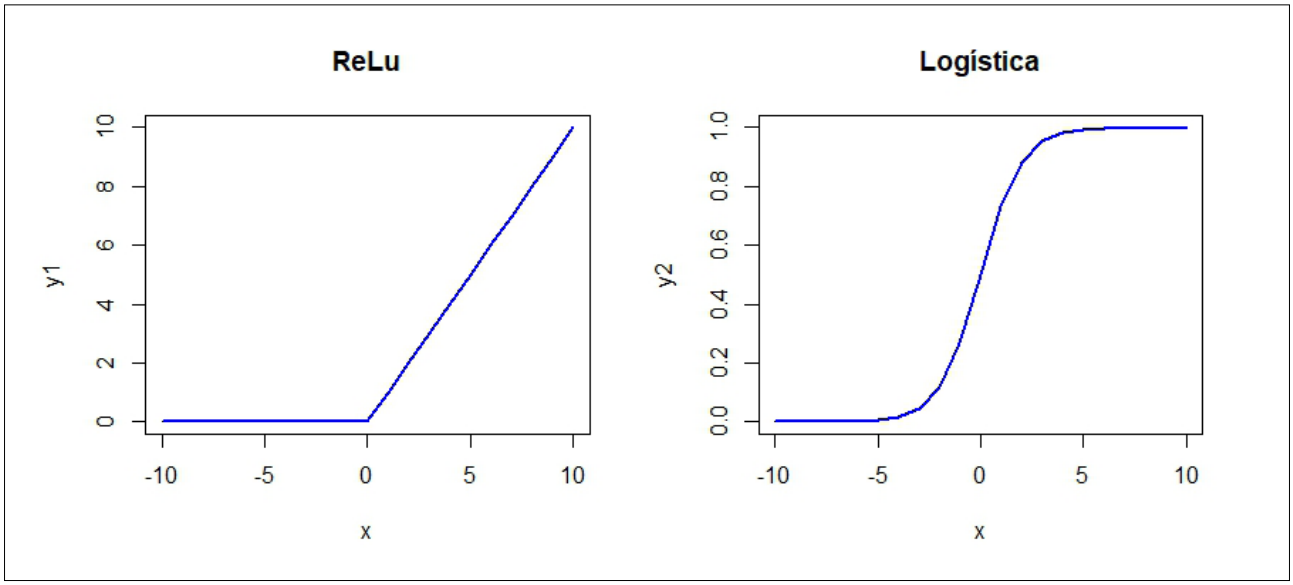

Figura 3. La función de activación ReLU y la sigmoide o logística, funciones de activación habituales en redes neuronales.

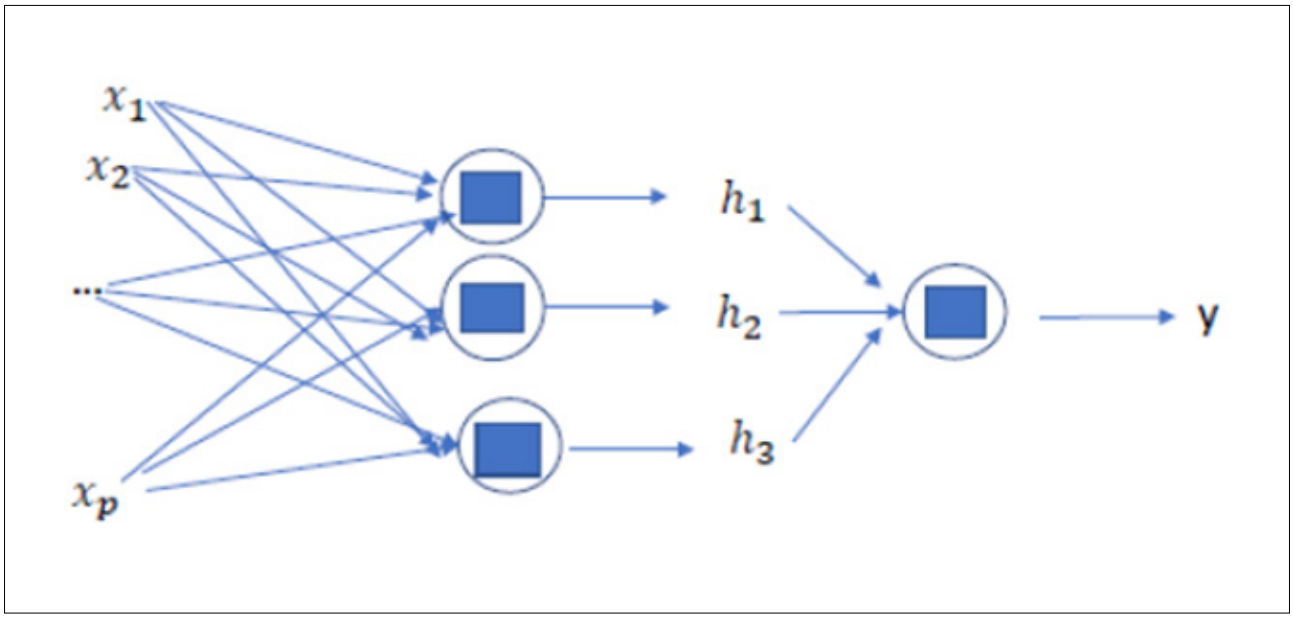

Figura 4. RNA con tres neuronas.

sión, como componentes principales. Estos métodos suponen que una variable respuesta depende de distintas combinaciones lineales de conjuntos de variables explicativas que llamamos factores, que se construyen por su capacidad para interpretar y prever la variable respuesta. En estadística las variables resumen son generalmente linealmente independientes entre sí, mientras que las salidas de las neuronas no están sujetas a esta restricción. La segunda idea central en las RNA es suponer en cada neurona una relación no lineal entre la entrada y la salida y construir la salida final por superposición de todas las salidas de las neuronas intermedias. Esta formulación tiene la ventaja de la generalidad, y el inconveniente de hacer más complejo la interpretación de la relación entre las variables de entrada y la de salida, que no puede descomponerse, en general, entre una respuesta lineal, con cierto peso, y un parte no lineal complementaria, pero en la actualidad se está trabajando en redes neuronales híbridas que permiten esta interpretación (véase por ejemplo Hajirahimi, and Khashei, 2019 and Mahtout, and Ziel, 2026).

4.3 El aprendizaje profundo o Deep Learning.

El siguiente paso en el avance de las redes neuronales para modelar relaciones complejas entre las entradas y las salidas es utilizar distintas capas de neuronas, como se representa en la figura 5, donde las tres neuronas de la primera capa proporcionan respuestas, h_{11}, h_{12}, h_{13}, que son las entradas de la segunda capa. En ella se combinan las dos respuestas generadas por las dos neuronas de esta capa, h_{21}, h_{22}, para crear la variable de salida, y, o respuesta final de la red. El desarrollo de estas redes no se produce hasta finales del siglo pasado cuando Rumelhart, Hinton y Williams (1986) encuentran un algoritmo eficiente (backpropagation) para calcular todos los pesos que definen las salidas de las neuronas minimizando el error de predicción. Además, Cybenko (1989) y Hornik (1991) demostraron que una red neuronal con varias neuronas en una única capa y función de activación tipo sigmoide puede, al aumentar el número de neuronas, aproximar cualquier función continua. Con estas bases durante nuestro siglo el crecimiento de las redes con muchas capas, o Deep learning, aprendizaje profundo, ha sido espectacular.

Además de este tipo de redes RNA, llamadas FF o Feed-Forward porque la información de las variables de entrada fluye siempre hacia adelante, que se utilizan mucho para la predicción, existen otros tipos de redes neuronales adaptados a objetivos específicos. Para el análisis y clasificación de imágenes se utilizan mucho las redes de convolución, para datos temporales, y, para variables dinámicas, las redes recur-

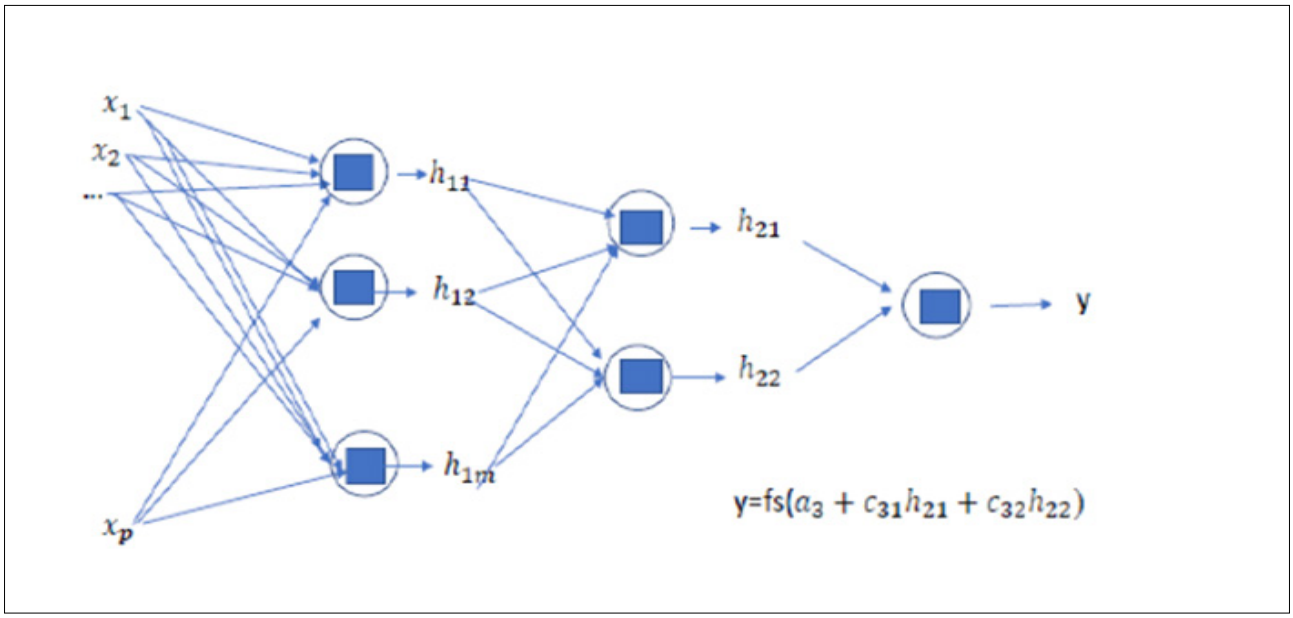

Figura 5. RNA con dos capas y m neuronas en la primera y dos en la segunda.

rentes. Más recientemente se han introducido para el procesamiento de lenguaje natural las redes de transformación, o Transformers, que están teniendo muchas otras aplicaciones en la actualidad. Vamos a describir muy brevemente cómo funcionan las dos primeras y explicaremos el funcionamiento de los Transformes en la sección siguiente dedicada a la Inteligencia Artificial, ya que son el núcleo de la llamada IA generativa.

Las redes neuronales de convolución (CNN) se utilizan para la clasificación de imágenes y datos espaciales. Se basan en el descubrimiento neurológico de Hubel y Wiesel en 1959-60, premiado con el Nobel en Medicina, que encontraron que las neuronas de la corteza visual responden a señales complejas como líneas, bordes y orientaciones específicas. La idea de las CNN es identificar los objetos en una imagen buscando líneas, bordes y patrones específicos, como formas circulares o rectangulares.

Supongamos que queremos clasificar una imagen digital y, para simplificar la exposición, supondremos inicialmente que la imagen está representada por una matriz cuyas celdas son los píxeles de la imagen y el contenido de las celdas es la intensidad de gris, que variará de intensidad 0 para el color blanco al 100% para el color negro. Supongamos que la imagen es una matriz cuadrada con 100 píxeles por lado y un total de 10.000 píxeles. Hay varias formas de construir digitalmente imágenes en color. Un sistema muy utilizado es el RGB, donde el color se representa por tres matrices, que contienen la intensidad de color rojo (red), verde (green) y azul (blue), requerida para reproducir el color de cada píxel. El análisis de imágenes en color es similar y los filtros que vamos a describir se aplican a cada una de las tres matrices separadamente.

El primer paso es extraer las características de la imagen, o detección de sus elementos. Por ejemplo, la imagen puede representar personas, paisajes o animales. Cada uno de estos elementos tiene formas geométricas definidas, como círculos, asociadas a ojos y rostros, o líneas, para brazos y piernas. Para extraer esas características se utilizan filtros o convoluciones. Un filtro es una matriz pequeña, por ejemplo, de 3x3 píxeles que se aplica a un trozo de la imagen y proporciona una medida de similitud entre el trozo de imagen analizado y el filtro. A continuación, se mueve el filtro por la imagen en horizontal y en vertical, hasta obtener la similitud entre cada posible trozo de la imagen de 3x3 píxeles y el filtro considerado. La medida de similitud utilizada suele ser simplemente una suma ponderada, o convolución, de los valores en los píxeles de la imagen y el filtro. Estas similitudes se guardan en una matriz de menor tamaño, proporcionando un mapa de la situación de los trozos de la imagen más similares al filtro.

Podemos aplicar distintos filtros para detectar distintos elementos y resumir toda esta información en matrices de pequeña dimensión que incluyen la información de situación e intensidad de un rasgo específico de la imagen definido por el filtro aplicado. A continuación, las matrices resultantes de aplicar los filtros, que resumen los rasgos importantes de la imagen se utilizan como variables explicativas de una red Feed-Forward donde las variables respuesta son los indicadores de cada clase para clasificarla. Para ello hay que comparar los rasgos de la imagen con los de una base muy amplia de objetos clasificados, como rostros, animales o montañas, para calcular las probabilidades de que pertenezcan a cada clase y clasificarla en la más probable. La figura 6 ilustra este proceso.

Estas redes clasifican nuestras fotos en los teléfonos móviles y reconocen personas en ellas. Fueron introducidas por LeCun et al. (2002) aplicando convoluciones, es decir, promedios ponderados de píxeles y el algoritmo de retro-propagación para redes profundas para reconocer automáticamente dígitos. Estos modelos, desarrollados a partir de 2012, tienen millones de parámetros (60 millones el desarrollado en al artículo seminal de Krizhevsky, Sutskever y Hinton, 2012) y su desarrollo ha sido posible por el gran aumento en velocidad de computación. Las redes CNN se están perfeccionando por la aparición de las redes de Tansformers, como comentaremos a continuación, y se aplican mucho en medicina, para el análisis de radiografías, resonancias, microscopía, etc, así como en el reconocimiento de rostros, vehículos o escritura, formando el núcleo de la llamada visión artificial.

Un tipo de redes profundas desarrolladas para datos temporales o secuenciales son las redes recurrentes. Tienen en cuenta que las predicciones de una variable temporal, como la temperatura ahora, y(t), pueden depender de los valores actuales de las variables explicativas, x(t), pero también de sus valores previos x(t-1), x(t-2) etc. La figura 7 presenta el esquema de estas redes. En cada momento t los vectores de entrada se combinan linealmente para formar variables de estado s(t) y una parte de ellas h(t) se transmite como input en el estado t+1, como ocurre con los modelos autorregresivos de series temporales.

Estas redes RNN clásicas con estructura autorregresiva tienen dificultades para incorporar dependencias con muchos retardos. Véase por ejemplo Bengio et al (1994). Las redes Short Long Term Memory (SLTM) incorporan una celda de memoria y tres compuertas (olvido, entrada, salida) que controlan el flujo de información y permiten conservar datos relevantes durante períodos más largos (véase por ejemplo, Chung et al, 2014, Peña and Tsay, 2021 y Staudemeyer and Morris, 2019). Las redes Gated Recurrent Unit (GRU) introducen una compuerta recurrente (véase Dey and Salem, 2017).

Existen otras redes que tienen en cuenta una dependencia más compleja entre los elementos de una secuencia que la puramente temporal, como ocurre con la dependencia de las palabras en el lenguaje. La más utilizada es la red Transformers que se diseñó para el procesamiento de lenguaje natural y la traducción de idiomas. Estas redes constituyen actualmente el núcleo de muchas de las aplicaciones más populares de la IA y se exponen en la sección siguiente.

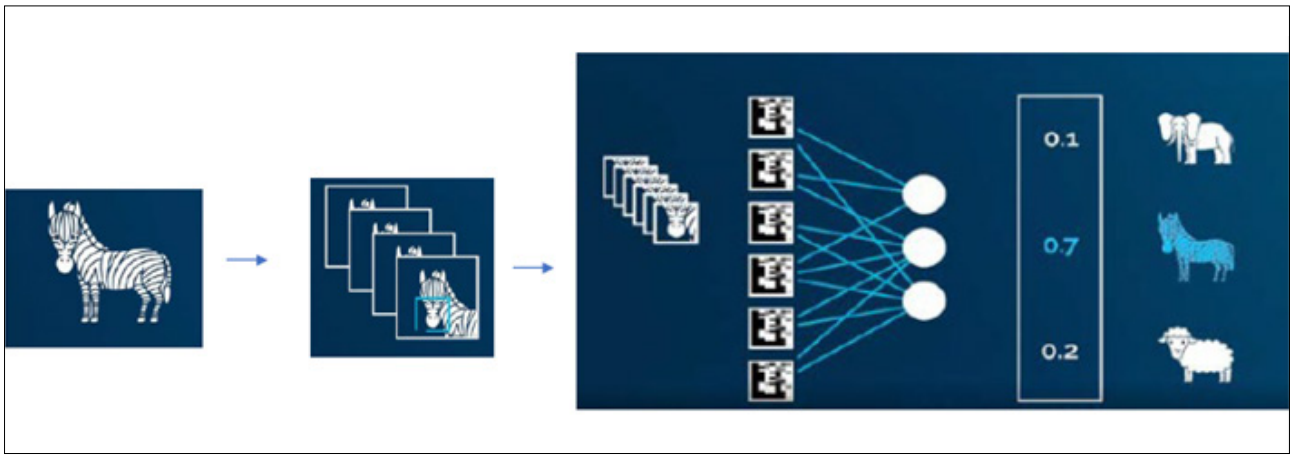

Figura 6. Imagen simplificada de una CNN. Se aplican filtros que por convolución (agregación) detectan rasgos de la imagen y los rasgos encontrados se utilizan como variables en una RNA clásica para clasificar la imagen.

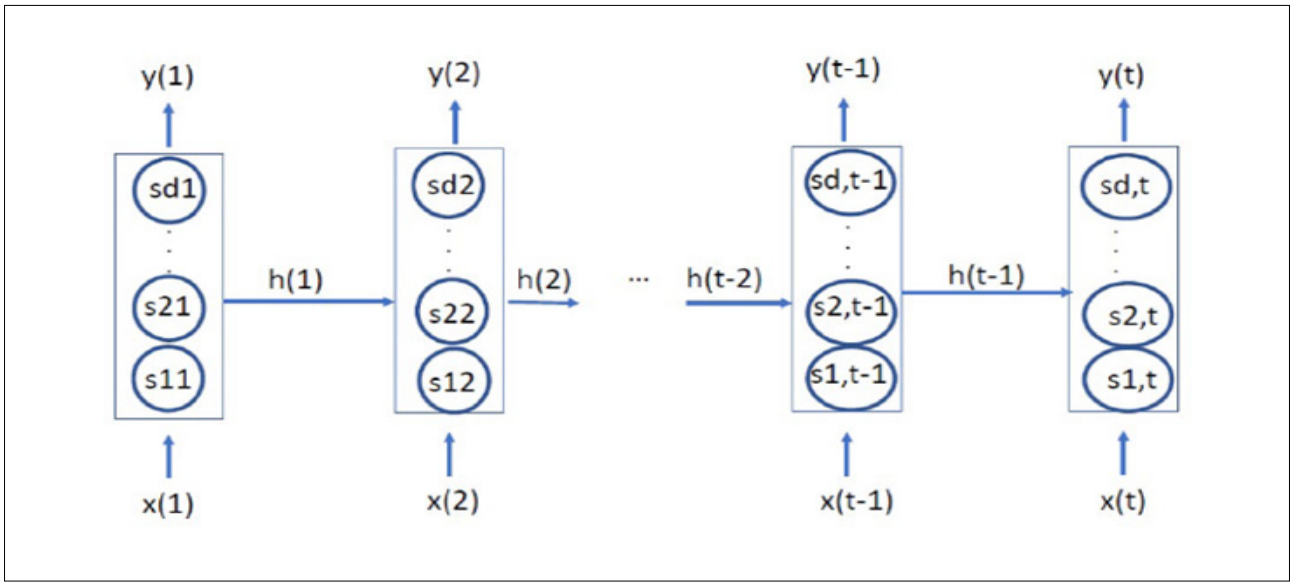

Figura 7. Imagen simplificada de un Red Recurrente Inicial. La información de los instantes previos al que se desea prever se incorpora también a la predicción.

Las redes neuronales han cambiado el paradigma estadístico clásico de análisis de datos establecido por Fisher. En la actualidad muchas investigaciones comienzan sin ninguna teoría previa, sino explorando una base de datos para entender las relaciones entre las variables y los grupos que puedan existir entre ellos. Como consecuencia de este análisis exploratorio pueden surgir modelos cuya validez se determina por su capacidad predictiva de los datos existentes. Su mayor limitación es que con ello se generan modelos predictivos con frecuencia difíciles de interpretar y que funcionan como cajas negras capaces de hacer buenas predicciones. En los últimos años muchos trabajos han explorado métodos de interpretación de los resultados que permitan elaborar teorías que amplíen nuestro conocimiento. Véase por ejemplo Murdoch et al (2019), Molnar (2020), Allen et al. (2023) and Duato et al. (2025).

5. LA INTELIGENCIA ARTIFICIAL (IA)

5.1 Breve historia de la IA

La Inteligencia artificial (IA) nace con los ordenadores en la segunda mitad del siglo XX con el objetivo de crear sistemas que puedan emular las capacidades de los seres humanos, como hablar y entender el lenguaje, ver y comprender imágenes y escribir y responder preguntas por escrito. El nombre de IA se utilizó por primera vez en una reunión de informáticos en Dartmouth (EEUU) en 1955 (véase McCarthy et al., 2006). Un avance importante fue el programa ELIZA, uno de los primeros de procesamiento del lenguaje natural (NLP), desarrollado en 1966 por Joseph Weizenbaum , científico informático del MIT, Weizenbaum (1966). El programa simulaba una conversación con un humano (chatbox) utilizando patrones de texto para generar respuestas. Su módulo más famoso fue DOCTOR, que imitaba a un psicoterapeuta de forma convincente, ya que muchos usuarios creyeron que ELIZA realmente "entendía", la conversación y "comprendía" el estado de ánimo del interlocutor, lo que sorprendió incluso al propio autor del programa que fue de los primeros en alertar en los riesgos asociados a este tipo de aplicaciones de la IA, Weizenbaum (1976).

La inteligencia artificial evoluciona lentamente en los años siguientes, aunque se producen avances en el campo de los sistemas expertos de decisión para el diagnóstico médico, como por ejemplo el MYCIN, desarrollado en la Universidad de Stanford entre 1972 y 1980 por Edward Shortliffe, un estudiante de doctorado en informática médica, o el desarrollado en nuestro país para la diagnosis de la ictericia por Peña (1980). También, se producen avances en campos considerados entonces como parte de la IA, como la robótica. Sin embargo, los resultados tienen poca repercusión por dos dificultades principales. La primera

era la dificultad de estimar muchos parámetros con la capacidad de cálculo disponible, y un avance importante para estimar redes neuronales profundas ha sido el algoritmo de retropropagación (backpropagation) descubierto por Rumelhart et al. (1986). La segunda consistía en la escasez de datos para entrenar sistemas capaces de aprender con muchos ejemplos, lo que limitaba los avances en campos como la visión o la audición artificial mediante reconocimiento de voz y de imágenes. Por ejemplo, los sistemas iniciales de reconocimiento de voz se diseñaban para que una persona concreta se grabase repetidamente, para proporcionar al sistema ejemplos suficientes para entrenar de forma efectiva al algoritmo utilizado para identificar las palabras.

Muchos de los esfuerzos en IA se concentraron en programar máquinas capaces de jugar al ajedrez y otros juegos donde existían, o podían generarse fácilmente, muchos datos de entrenamiento. Un éxito importante se produjo en 1997 cuando el programa Deep Blue, desarrollado por IBM, derrotó a Kasparov, campeón del mundo de ajedrez. Este programa se basaba en calcular un número gigantesco de variantes que evaluaba con criterios diseñados por expertos. La figura 8 resume la evolución de la IA en el siglo XX y podemos concluir que a finales del siglo pasado la inteligencia artificial era una herramienta prometedora, pero de escasa utilidad práctica.

El crecimiento de la IA se produce en nuestro siglo por la aparición de los datos masivos y la digitalización de imágenes, textos, audios y videos, junto al enorme crecimiento ya comentado de nuestra capacidad de cálculo y de transmisión de la información. Los trabajos de Hinton, LeCun y Bengio sientan en el periodo 2000-2012 las bases de las redes profundas, Deep learning, para la visión artificial y el procesado del lenguaje natural. Basados en ello se crean programas de gran impacto práctico, como Alexnet para el reconocimiento de imágenes, o de traducción automática, con Google Brain, o el programa AlphaFold2, que ha revolucionado la biología estructural al predecir estructuras de proteínas con precisión casi experimental y que llevó a sus creadores, Demis Hassabis y John Jumper, al Premio Nobel de Química en 2024. También se basa en aprendizaje profundo el programa AlphaZero, para el juego del Go, mucho más complejo que el ajedrez, que en 2017 derrotó al campeón del mundo en esta especialidad. En este caso el programa no utilizó la fuerza bruta, sino que aprendió jugando contra sí mismo millones de veces,

Figura 8. Breve historia de la IA en el siglo XX.

utilizando métodos de aprendizaje profundo como los explicados en la sección anterior, y modelos probabilísticos para generar jugadas al azar. Esto le permitió utilizar estrategias ganadoras que no habían sido habituales en las partidas jugadas entre seres humanos.

El avance con más repercusión de la IA en los últimos años ha sido la aparición de los modelos generativos, que crean textos, imágenes o sonidos, a petición del usuario. Estos modelos se basan en las redes de Transformers, introducidas por Vaswani et al. (2017), que incluyen un mecanismo de atención para captar las relaciones entre las palabras en una secuencia escrita. Esta idea ha sido clave para la aparición de los grandes modelos de lenguaje o Large Lenguage Models (LLM), como ChatGPT , Gemini o DeepSeek, que han revolucionado las aplicaciones de la IA, y cuyos fundamentos describimos en la sección siguiente.

Las figuras 8 y 9 resumen la evolución de la IA durante el siglo pasado y el actual. El lector interesado en una introducción simple a la Inteligencia Artificial puede consultar Peña (2026), que incluye muchas referencias, y los videos desarrollados en Peña (2025b).

5.2 Los modelos generativos de IA

El avance con más impacto público de la IA han sido los modelos generativos de lenguaje, que se utilizan para traducir texto entre idiomas o generar nuevos textos a demanda del usuario. La tabla 1 presenta un resumen de algunos de los modelos más utilizados y sus propiedades y ha sido realizado con GPT-4.

Estos sistemas se han diseñado para tener en cuenta dependencias flexibles, como las que ocurren en el lenguaje, que son muy distintas de las dependencias cronológicas que aparecen en series temporales. En el lenguaje el orden de las palabras es importante para la comprensión, pero la relación entre las palabras no depende de su distancia, sino de su significado, que se clarifica al ver las palabras que preceden a cada una y forman el contexto en que aparecen. Por ejemplo, la palabra banco tiene un significado muy distinto si estamos hablando de ríos, de muebles, de finanzas o de almacenaje de información.

En los modelos de lenguaje natural las palabras se transforman en vectores de alta dimensión y este proceso se denomina Word embedding, o representación vectorial de palabras. Este proceso, iniciado

Figura 9. Breve historia de la IA en el siglo XXI.

Modelo	Razonamiento	Código	Conversación	Apertura
GPT-4	★★★★★	★★★★	★★★★★	✗
Claude	★★★★	★★★	★★★★★	✗
Gemini	★★★	★★★	★★★★	✗
LLaMA	★★★★	★★★★	★★★	✓
Mistral	★★★	★★★	★★★	✓
DeepSeek	★★★★★	★★★★★	★★	✓

Tabla 1. Propiedades de los modelos generativos de lenguaje más utilizados. Fuente: realizada con GPT-4.

por Bengio et al (2003), se lleva a cabo de manera que los vectores que representan palabras similares, es decir que tienen significados parecidos y tienden a aparecen en el mismo contexto, estén próximos. La proximidad entre dos vectores se mide por el producto escalar, que es equivalente a considerar la correlación entre sus coordenadas para vectores estandarizados a longitud (o módulo) unitaria. El vector que representa a cada palabra pueden interpretarse como la valoración de la palabra según unos factores que reflejan su significación dentro del lenguaje. Estos factores del lenguaje son las dimensiones de los vectores, y palabras similares tendrán coordenadas parecidas en los factores más importantes para representarlas.

La forma inicial de realizar el Word embedding fue considerar todas las palabras del idioma, definir un intervalo de proximidad, por ejemplo, seis palabras, y aplicar en cada texto disponible para entrenamiento una ventana de longitud seis y contar para cada palabra las frecuencias relativas con las que el resto de las palabras del idioma aparecen próximas a ella.

Con esta operación la longitud del embedding de cada palabra es siempre igual al número de palabras del idioma y cada coordenada representa la proximidad o frecuencia de aparición de la palabra con respecto a todas las demás. Esta matriz tiene mucho ceros y valores pequeños, porque muchas palabras jamás, o muy raramente, aparecen con otras. Podemos resumir su información representando cada palabra por un vector de dimensión mucho menor, cuyas coordenadas representen factores semánticos. Por ejemplo, podría aplicarse componentes principales y considerar los componentes como factores del lenguaje, pero en los modelos de lenguaje el embedding se construye con el objetivo de prever la siguiente palabra y los factores no tienen que ser ortogonales, como en componentes principales.

Inicialmente los métodos de generación de lenguaje y de traducción automática asignaban un embedding fijo a cada palabra con independencia del contexto, como los sistemas Word2Vec o GloVe, pero la aparición de los Transformers y de su mecanismo de atención entre las palabras asigna un embedding variable a las palabras que depende del contexto donde aparecen, haciendo su representación mucho más eficaz. Además, el embedding de las palabras se adapta al objetivo y es parte del proceso que se realiza. Por ejemplo, el embedding de una palabra puede ser distinto si queremos generar lenguaje prediciendo la siguiente palabra de una secuencia dadas las anteriores , que si queremos traducir un texto.

Los modelos generativos actuales generan las palabras en función de un contexto definido por la pregunta que les hacemos. El embedding se hace teniendo en cuenta las probabilidades de cada palabra condicionada a las que forman la pregunta y su longitud, o número de coordenadas del vector, depende del sistema. Por ejemplo, los modelos actuales tienen una dimensión de embedding de entre 4000 y 12000 dimensiones aproximadamente. Para ello se utilizan redes Transformers seguidas de redes Feed-Forward para prever la palabra siguiente condicionada al contexto, que está definido por las palabras previas a ella. Para ello se establecen medidas, basadas en correlaciones, de la relación de cada palabra con las que le rodean. La misma idea puede utilizarse para otras acciones generativas de imágenes, videos o sonidos.

Es interesante señalar que la idea de análisis factorial, introducida por Charles Spearman (1904) como técnica estadística para estudiar correlaciones entre tests cognitivos y hacer una teoría factorial de la inteligencia, aparece hoy, un siglo más tarde, como una importante herramienta en los métodos de inteligencia artificial: para resumir la información cuantitativa de las variables de entrada en una red neuronal y, también, para vectorizar las palabras en los modelos de lenguaje.

Las figuras 10 y 11 resumen el funcionamiento de los modelos generativos de lenguaje. Estos gráficos tratan de explicar la lógica general, y no los detalles de cada uno de los sistemas. Las variables de entrada son las palabras de la frase que definen el contexto de generación de las palabras siguientes y a cada una se le asigna un embedding inicial general que va a modificarse para adaptarla al contexto. Estas palabras se interpretan con el mecanismo de atención, o Red Transformer, por su nombre inicial debido a Vaswani et al. (2017). Su esquema de funcionamiento se presenta en la figura 11. El contexto modifica el embedding de las palabras en función de las que la acompañan. Para ello se definen tres variables para cada palabra que van a tener dimensión menor que el embedding inicial, y que se refieren a distintas partes del embedding, es decir a distintos factores del lenguaje, para clarificar por el contexto la interpretación de las palabras. Estas variables son el *Value*, que va a ser el la modificación de una parte del embedding de cada palabra teniendo en cuenta la información proporcionada por las otras del contexto, las *Queries*, que definen qué información sobre cada palabra queremos buscar en las otras palabras del contexto y las *Keys*, que indican la información que estas otras palabras pueden aportar a la que consideramos.

Utilizando las correlaciones entre las Keys y las Queries se modifican las partes del embedding definidas en el Value, adaptando su significado al contexto. Esto se repite muchas veces para distintos tipos de Keys, Queries y Values, que afectan a distintas parte del embedding inicial, es decir a distintos factores del lenguaje, y con ello se obtiene el embedding final de cada palabra adaptada al contexto.

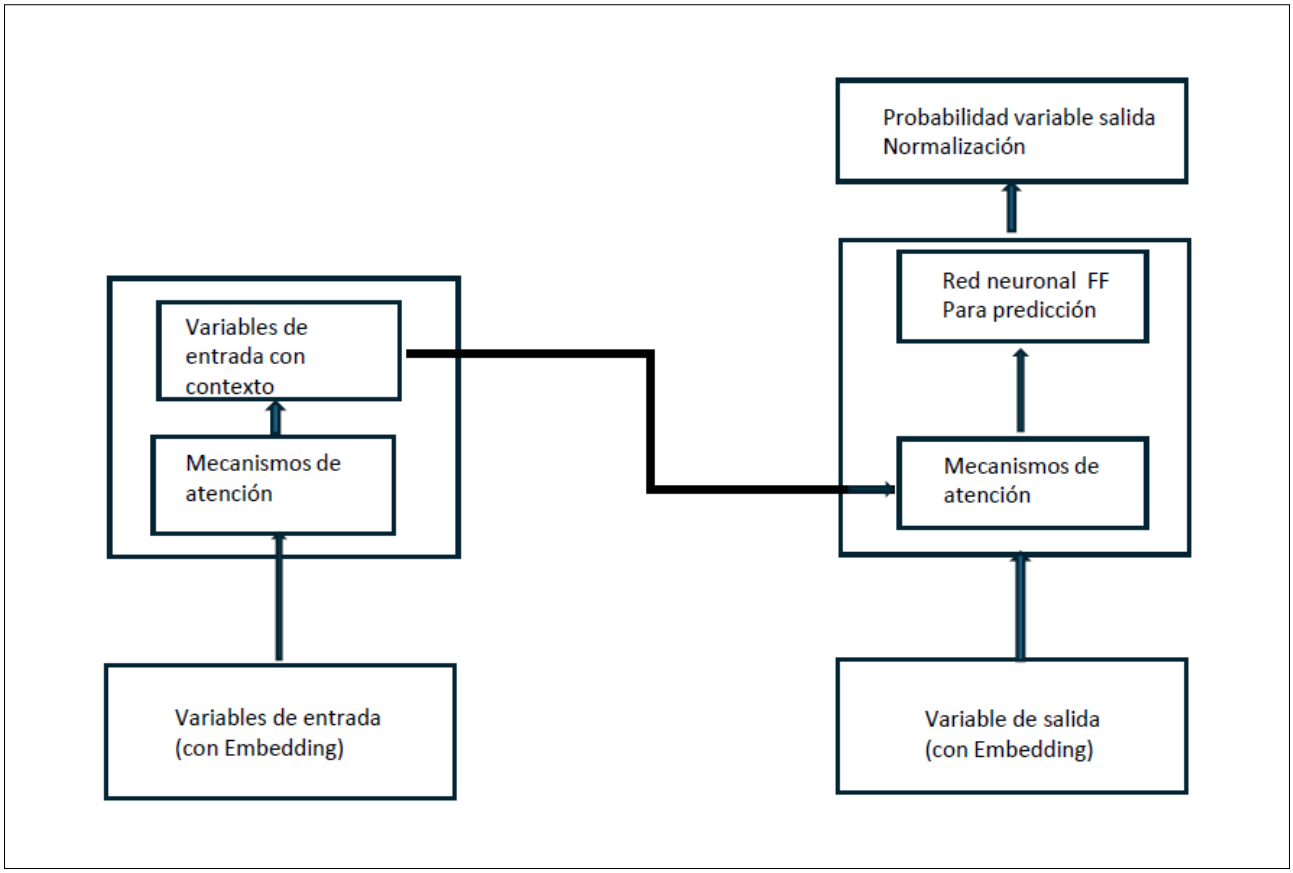

Figura 10. Esquema de funcionamiento simplificado de un LLM.

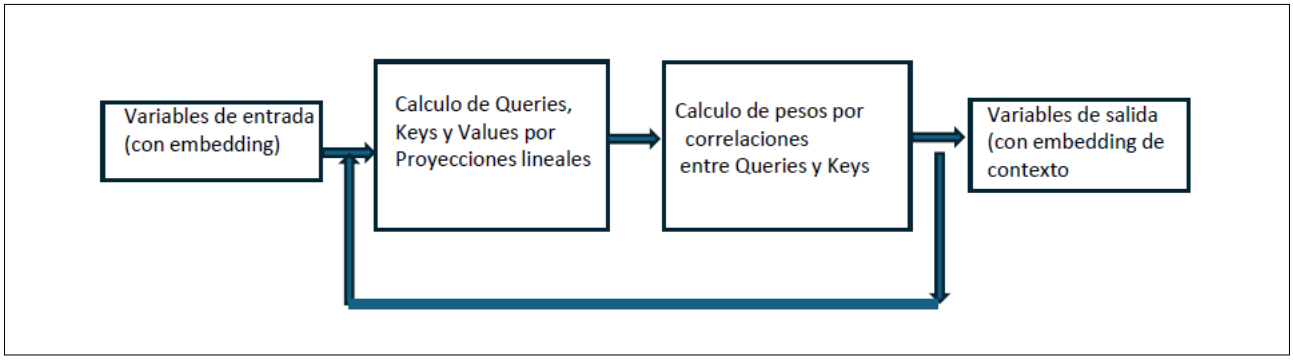

Figura 11. El mecanismo de atención del Transformer.

Una vez adaptadas las palabras de entrada se hace lo mismo con las posibles palabras de salida, o de continuación de la entrada, para adaptar su embedding al contexto de entrada mediante otro mecanismo de atención que tiene en cuenta la relación entre las palabras de entrada y salida. Después toda esta información pasa a una red neuronal del tipo ANN Feed-Forward, que se utiliza para prever las probabilidades de las siguientes posibles palabras condicionadas a las anteriores de la frase.

6. CONCLUSIONES

La Inteligencia artificial se está convirtiendo en el aglutinante de varias disciplinas desarrolladas en el siglo XX y XXI. Engloba los métodos de aprendizaje automático, ML, pero también muchos de los métodos utilizados en Investigación Operativa y Optimización. No coincide con la Informática, pero está influyendo ya mucho en su desarrollo y sus aplicaciones, ni tampoco con la Estadística, que conservará su campo específico de aprender de experimentos puntuales y de diseñar procedimientos efectivos de recoger datos evitando sesgos de medida, pero estará también muy influida en el futuro por los avance y necesidades probabilísticas de la IA. Además, los métodos estadísticos pueden ayudar mucho para interpretar los utilizados en IA (véase Allen et al, 2023) y prevenir sesgos introducidos por la recogida automática de datos masivos.

Un problema en el desarrollo de la IA son los enormes recursos energéticos y ambientales que requiere. Véase por ejemplo Bolón-Canedo et al, (2024). Como las grandes empresas tecnológicas son las únicas que disponen de recursos de esta magnitud los avances en IA se realizan sobre todo por ellas y en su beneficio, en lugar de hacerse en las universidades y centros públicos de investigación, donde el conocimiento se pone a disposición de toda la humanidad. Este proceso ha incrementado ya la desigualdad en este siglo, de forma muy preocupante, y puede hacerlo de manera muy grave en el futuro si los gobiernos no toman medidas para impedirlo. Una forma de avanzar de forma más sostenible es utilizar métodos híbridos en IA, combinando los métodos estadísticos interpretables y de bajo coste con las redes neuronales profundas para problemas específicos de interés general con alto valor global añadido.

Se ha hablado mucho de los riesgos de la IA. Sus métodos pueden utilizarse para generar noticias falsas, difundir bulos y manipular a la opinión pública propagando mensajes que dividan y enfrenten a las personas. También puede tener un gran potencial destructivo en manos delictivas. Es una amenaza para muchos trabajos, al sustituir a las personas por agentes informáticos y procesos automáticos y puede llevar a una sociedad menos crítica, menos culta y con mayor desigualdad económica. Frente a estos riesgos, existe el potencial de grandes avances en medicina, en el cuidado de la salud y en el aumento de la esperanza de vida, así como de avances científicos en muchas áreas que pueden reducir, y quizás incluso revertir, el impacto del cambio climático y proporcionar ventajas generales para toda la humanidad.

AGRADECIMIENTOS

Agradezco mucho a un evaluador anónimo sus comentarios, que han mejorado sustancialmente la versión inicial de este trabajo. También agradezco el continuo apoyo de FUNCAS, Madrid, a mi investigación sobre Big Data e Inteligencia Artificial.

CONFLICTO DE INTERESES

El autor de este artículo declara no tener ningún tipo de conflicto de intereses respecto a lo expuesto en el presente trabajo.

REFERENCIAS BIBLIOGRÁFICAS

1. Allen, G. I., Gan, L., and Zheng, L. (2023). Interpretable machine learning for discovery: Statistical challenges and opportunities. *Annual Review of Statistics and Its Application*, *11*.

2. Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*, 1137:1155.

3. Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157:166.

4. Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, *39* (227), 357:365.

5. Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., and Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, *599*, 128096.

6. Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

7. Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 144:152.

8. Bostridge, M. (2008) *Florence Nightingale: The Woman and Her Legend*, Penguin Books.

9. Box, G. E. P (1976). Science and Statistics. *Journal of American Statistical Association*, 71, 791:799.

10. Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273:297.

11. Cox, D. R. (1972). Regression models and life-tables. *J R Stat Soc Series B*, 34(2), 187:220.

12. Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 2(4), 303:314.

13. Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

14. David, F. N (1962). *Games, Gods and Gambling: A History of Probability and Statistical Ideas*. Charles Griffin.

15. Dey, R., and Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)* (pp. 1597:1600). IEEE.

16. Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68:77.

17. Duato, J., Mestre, J. I., Dolz, M. F., Quintana-Orti, E. S., and Cano, J. (2025). Decoupling structural and quantitative knowledge in relu-based deep neural networks. In *Proceedings of the 5th Workshop on Machine Learning and Systems*, 39:45.

18. Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* 7, 1:26.

19. Greenwood, M., and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of Royal Statistical Society*, *83*(2), 255:279.

20. Hasofer, A. M (1967). Random Mechanisms in Talmudic literature. *Biometrika*, 54, 316:21.

21. Hajirahimi, Z., and Khashei, M. (2019). Hybrid structures in time series modeling and forecasting: A review. *Engineering Applications of Artificial Intelligence*, *86*, 83:106.

22. Hornik, K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4(2), 251:257.

23. Kendall, M. G. (1956). The beginnings of a probability calculus. *Biometrika*, 43, 1:14.

24. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

25. Mahtout, B. E., and Ziel, F. (2026). Electricity Price Forecasting: Bridging Linear Models, Neural Networks and Online Learning. *arXiv preprint arXiv:2601.02856*.

26. McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI magazine*, *27*(4), 12:12.

27. Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

28. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071:22080.

29. LeCun, Yann, et al. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, 11, 2278 : 2324.

30. Peña, D. (1980). A decision analysis model for a serious medical problem. *Management Science*, 26, 7, 707:718.

31. Peña, D. and Tsay, R. S. (2021). *Statistical Learning for Big Dependent Data*. Wiley.

32. Peña, D. (2025a). *Investigación Económica con Datos Masivos* (Coordinador). Fundación Ramón Areces.

33. Peña, D. (2025b). Comprender la IA: *Big data y aprendizaje estadístico automático*. Funcas: https://www.funcas.es/articulos/big-data-aprendizaje-estadistico-automatico-e-inteligencia-artificial/ YouTube: https://www.youtube.com/aylist?list=PLQR2nth3aeQRdcUKsIuMzKIbFEeeSgzEC

34. Peña, D. (2026*). Comprender la Inteligencia Artificial*. Funcas, Madrid.

35. Riley, J. C. (2001). *Rising Life Expectancy: A Global History*. Cambridge University Press.

36. Rosenblatt, F. (1958). *Two theorems of statistical separability in the perceptron*. Washington, DC, USA: United States Department of Commerce.

37. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533:536.

38. Staudemeyer, R. C., and Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.

39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

40. Wahba, G. (1990). *Spline models for observational data*. Society for industrial and applied mathematics.

41. Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36:45.

42. Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. American Psychological Association.

Si desea citar nuestro artículo:

Peña D. Datos masivos: de la estadística a la Inteligencia Artificial. RACSG.2026;114(01):71-90 rac.2026.114.1.org05