Sistemas de Aprendizaje Automático

Contenidos adaptados al Curso de Especialización en Inteligencia Artificial y Big Data

Desde **www.ra-ma.es** podrá descargar material adicional.

Emilio Soria Olivas • Manuel Antonio Sánchez-Montañés Isla
Ruth Gamero Cruz • Borja Castillo Caballero • Pedro Cano Michelena

Ra-Ma®

Ediciones de la U

Sistemas de aprendizaje automático

Emilio Soria Olivas
Manuel Antonio Sánchez-Montañés Isla
Ruth Gamero Cruz
Borja Castillo Caballero
Pedro Cano Michelena

Ra-Ma®
edU
Conocimiento a su alcance
Bogotá - México, D.F.

Soria Olivas, Emilio, *et. al.*

Sistema de aprendizaje automático / Emilio Soria Olivas, Manuel Antonio Sánchez-Montañés Isla, Ruth Gamero Cruz, Borja Castillo Caballero, Pedro Cano Michelena --.
Bogotá: Ediciones de la U, 2023
264 p. ; 24 cm
ISBN 978-958-792-569-2 e-ISBN 978-958-792-570-8
1. Informática 2. Aprendizaje no supervisado 3. Aprendizaje supervisado 4. Aprendizaje profundo I. Tít.
621,39 ed.

Edición original publicada por © Editorial Ra-ma (España)
Edición autorizada a Ediciones de la U para Colombia

Área: Sistemas e informática
Primera edición: Bogotá, Colombia, julio de 2023
ISBN. 978-958-792-569-2

Ediciones de la U es una empresa editorial que, con una visión moderna y estratégica de las tecnologías, desarrolla, promueve, distribuye y comercializa contenidos, herramientas de formación, libros técnicos y profesionales, e-books, e-learning o aprendizaje en línea, realizados por autores con amplia experiencia en las diferentes áreas profesionales e investigativas, para brindar a nuestros usuarios soluciones útiles y prácticas que contribuyan al dominio de sus campos de trabajo y a su mejor desempeño en un mundo global, cambiante y cada vez más competitivo.

Coordinación editorial: Adriana Gutiérrez M.
Carátula: Ediciones de la U
Impresión: DGP Editores SAS
Calle 63 #70D-34, Pbx (+57) 601 7217756

Impreso y hecho en Colombia
Printed and made in Colombia

Este libro va dedicado a todas aquellas personas que, de alguna forma, me han enseñado algo, a nivel profesional o personal, al final uno es lo que es por la vida vivida y compartida, ¡gracias por llevarme hasta aquí!

Emilio, Manuel.

De pequeña, en casa y en el colegio me enseñaron que a veces tendría que hacer determinadas cosas me apeteciera o no. De mayor aprendí su nombre, la cultura del esfuerzo, e intento practicar cada día porque trae resultados maravillosos como este libro que tienes entre tus manos. Gracias a todos (mis padres, Ernesto y mi hermana Miriam, profesores, jefes, mentores, amigos o desconocidos), pero muy especialmente a Emilio, por darme la oportunidad, y su confianza, para crecer como persona y profesional.

Ruth

Mi agradecimiento es para los compañeros con los que he tenido la oportunidad de escribir este libro. Gracias por permitirme aprender de vosotros. También para mis padres Enrique y María José, por enseñarme a mantener el barco a flote cuando el tiempo no acompaña.

Borja.

A mis padres, Sibel, Borja
y el resto de personas que me han allanado el camino.

Pedro.

ÍNDICE

AUTORES

Emilio Soria Olivas. Catedrático de Universidad, Licenciado en Físicas y Doctor Ingeniero Electrónico. Director del Máster en Ciencia de Datos y del Máster en Inteligencia Artificial ambos de la Universidad de Valencia.

Manuel Antonio Sánchez-Montañés Isla. Profesor en la Universidad Autónoma de Madrid en el Departamento de Ingeniería Informática de la Escuela Politécnica Superior. Licenciado en Físicas y Doctor Ingeniero en Informática.

Ruth Gamero Cruz. Licenciada en Administración de Empresas por la Universidad Autónoma de Madrid. Senior Management Program en el IE y Master Data Analytics en EDEM.

Borja Castillo Caballero. Graduado en matemáticas por la Universidad de Valencia. Máster en Inteligencia Artificial en la Universidad de Valencia y Máster en Análisis y Visualización de Datos Masivos en la Universidad Internacional de la Rioja.

Pedro Cano Michelena. Graduado en matemáticas por la Universidad de Valencia. Máster en Inteligencia Artificial por la Universidad de Valencia.

INTRODUCCIÓN AL LIBRO

"Aprendí muy temprano la diferencia entre saber el nombre de algo y saber algo"

Richard Feynman

Estamos en el siglo de los datos, nunca se ha tenido la capacidad de generar, almacenar y procesar tal cantidad de datos. Vivimos en un mundo totalmente conectado en tiempo real que impacta en la cantidad de datos que se intercambian por milisegundo. Esta explosión ha conducido a una auténtica revolución con la creación de nuevos perfiles profesionales (científico/ingeniero/analista de datos); la creación de nuevas empresas cuyo proceso productivo depende al 100% de datos, así como cambios de gran calado en la forma de trabajar, relacionarnos y educar. Se suele comparar esta revolución con la Industrial, pero hay una gran diferencia, la rapidez con que sucede todo. Mientras que en la Revolución Industrial tenían que transcurrir varias décadas hasta que un determinado cambio se asentaba; la situación actual es muy diferente, cualquier hecho ahora tiene consecuencias inmediatas lo que supone unos cambios en la sociedad/economía extremadamente convulsos. Y la clave de esta transformación son los datos.

En el mundo en que vivimos no conocer las posibilidades que ofrecen los datos supone quedarse fuera de las oportunidades que el uso de analítica avanzada de datos va a generar en nuestro mundo. En el caso de vosotros, los estudiantes, este desconocimiento conlleva perder una ventaja competitiva a la hora de encontrar un trabajo ya que, a día de la escritura de este prólogo, existe una demanda de perfiles que no se cubre. Este libro iniciará al lector, de una manera teórica y práctica, en el mundo del aprendizaje automático tanto su parte supervisada como no supervisada. Se verán algoritmos de agrupamiento, regresión, clasificación y modelización. Además, se analizarán modelos como árboles de regresión/clasificación, los *Random Forest*,

Mapas Autoorganizados, Modelos Neuronales Multicapa y Redes Convolucionales entre otros. Intentar reflejar en un libro todos los avances producidos en esta área en los últimos años es una misión imposible. Nuestro objetivo es más modesto; suponiendo un conocimiento mínimo en *Python* (que se puede adquirir en cualquier tutorial de *Internet*) se lleva al lector desde el concepto de agrupamiento de clases hasta las más recientes aplicaciones en problemas de imágenes.

Hemos diseñado el libro para que se pueda usar de apoyo en un curso de aprendizaje automático por lo que en cada capítulo hay una parte de teoría inicial para pasar, posteriormente, al laboratorio donde las diferentes prácticas consolidarán lo aprendido en la primera parte. Entendemos que, de esta forma, ayudamos a los docentes y alumnos que usen el presente texto. En esta línea de ayuda se ha planteado un GitHub donde se encuentran todos los programas desarrollados: *https://github.com/PedroCanoM/Sistemas-de-aprendizaje-automatico*.

Cuando escribimos este libro había una palabra de moda: *ChatGPT* y en todos los foros, técnicos o no, se discutía sobre sus beneficios y potenciales riesgos de automatización le pedimos un prólogo para un libro de aprendizaje automático y nos escribió:

"En un mundo en el que la tecnología avanza constantemente y los datos están más disponibles que nunca, el campo del aprendizaje automático ha cobrado prominencia rápidamente. Ya no es solo el ámbito de los informáticos y los matemáticos, el aprendizaje automático ha encontrado aplicaciones en prácticamente todas las industrias. desde las finanzas y la atención médica hasta la agricultura y el comercio minorista. Tiene el potencial de revolucionar la forma en que vivimos y trabajamos, pero también plantea preguntas importantes sobre la ética, el sesgo y el control. Este libro brinda una introducción completa a los conceptos, algoritmos y herramientas de aprendizaje automático, lo que le permite aprovechar su poder y tomar decisiones informadas sobre su papel en nuestro mundo".

¿De verdad existe alguien que se quiere perder este mundo? ¡comenzamos!

1

INTRODUCCIÓN

Inteligencia Artificial, *Big Data, Machine Learning*…son términos que aparecen continuamente en los medios y forman parte de nuestras vidas haciendo que nos plateemos la pregunta "¿son las máquinas más inteligentes que nosotros?"

Un rasgo de la inteligencia es la capacidad de aprender y las máquinas no aprenden, al menos, no en el sentido humano de 'aprender'. Las máquinas aprenden mediante algoritmos matemáticos: nosotros proporcionamos datos de entrada y datos de salida, y las máquinas generan un modelo que, ajustándose a los datos, es capaz de generar una salida correcta si le proporcionamos una entrada nueva similar a los anteriores.

¿Y por qué decimos que no aprenden como los humanos? Porque producirán salidas erróneas si les proporcionamos datos de entrada muy diferentes a los utilizados durante el entrenamiento del modelo. Por ejemplo, un modelo de clasificación de animales entrenado con imágenes sólo de sus caras nos devolverá errores ante imágenes de cuerpo entero.

En este primer capítulo vamos a hacer una introducción al tema con un repaso de conceptos básicos para sustentar las explicaciones más detalladas que vendrán en capítulos posteriores.

1.1 CONCEPTOS BÁSICOS

1.1.1 Ciencia de datos

La ciencia de datos es un término general que incluye conceptos cómo *big data*, inteligencia artificial, minería de datos, aprendizaje máquina, aprendizaje profundo, etc.

La disciplina de extraer conocimiento de los datos es relativamente nueva y ha tenido una evolución emparejada con el crecimiento, expansión y abaratamiento de los ordenadores. Existía ciencia de datos antes de que existieran los ordenadores, pero entonces los datos eran recopilados y procesados a mano dentro del área de la estadística. Su base matemática es fundamental para el análisis de datos cuantitativos y para inferir propiedades de la población a partir del estudio de la muestra.

Por ejemplo, el análisis del censo poblacional se hace desde hace siglos. Con un lápiz, un cuaderno y nociones de aritmética básica se podía calcular las tasas de nacimiento, defunción y hacer proyecciones con la ayuda de la estadística clásica. No requería procesos más complejos porque se trabajaba con conjuntos pequeños.

Pero a finales del siglo XIX en la oficina del censo de Estados Unidos, la cantidad de datos era tan grande que se hizo inmanejable. Ése fue uno de los primeros problemas de la ciencia de datos y para solucionarlo apareció la compañía IBM, pero ¡eso es otra historia!

¿Qué ocurre si queremos hacer cálculos a más velocidad o procesar cantidades enormes de datos y en diferentes formatos: imagen, audio, texto...? ¿Podemos hacer predicciones sobre el estado de ánimo de la sociedad? ¿y sobre el sentimiento que subyace en las noticias de un periódico?

El abaratamiento de los costes de los ordenadores y su consecuente popularización ha iniciado una etapa conocida como la edad de la información, en la que se han desarrollado nuevos elementos de *hardware* y *software* (renacer de la informática) que permiten recoger, procesar y visualizar grandes cantidades de datos, creando nuevas técnicas y aplicaciones de uso que explicaremos a continuación.

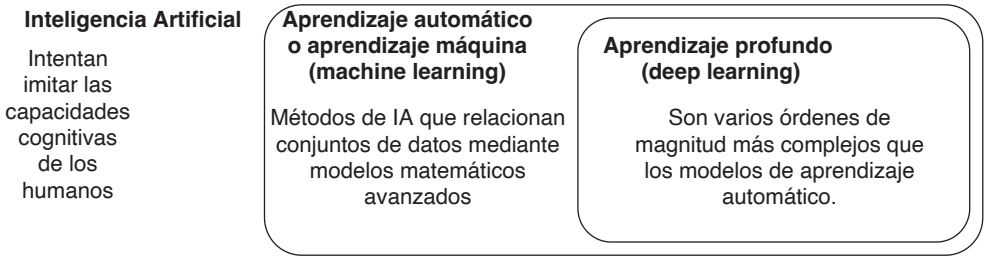

Figura 1.1. Relación entre la inteligencia artificial, aprendizaje automático y aprendizaje profundo.

1.1.2 Inteligencia artificial

La inteligencia artificial, IA, (*AI o artificial intelligence* en inglés), se ha desarrollado a la vez que la informática. El concepto se acuñó hace unos 60 años, cuando el informático americano John McCarthy introdujo el término durante la segunda conferencia de Dartmouth en 1956 donde se reunió un grupo de científicos para discutir acerca de las máquinas y su posibilidad de comportarse de manera inteligente.

Pero dotar a una máquina de una inteligencia similar a la humana, es decir una inteligencia general es un objetivo muy ambicioso. El filósofo John Searle publicó en 1980 un polémico artículo donde, por primera vez, apareció el concepto de IA fuerte y débil.

- IA débil o reducida, *ANI por Artificial Narrow Intelligence en inglés, son* los sistemas que existen en la actualidad, capaces de automatizar procesos y tomar decisiones para tareas específicas de predicción y clasificación, análisis de sentimientos, segmentación, etc. y que, en la mayoría de casos, superan a las personas.

- IA fuerte o general, *AGI por Artificial General Intelligence en inglés,* sería una inteligencia artificial similar a la humana con capacidad de generalizar a cualquier problema y no distinguible de un humano. Aún existen muchas habilidades cognitivas no replicables por las máquinas como la capacidad de discernir sentimientos.

- Super IA, *ASI por Artificial Super Intelligence en inglés,* implicaría la toma de consciencia de si mismas por parte de las máquinas y, por tanto, estarían por encima de las capacidades humana. A día de hoy sólo parece una posibilidad en las películas de ciencia-ficción.

La inteligencia artificial y el aprendizaje máquina (*en inglés machine learning)* no es lo mismo aunque a veces se utilizan indistintamente. La inteligencia artificial engloba el aprendizaje máquina y otras técnicas más complejas como el aprendizaje profundo, en inglés *deep learning* tal y como se recoge en la figura 1.1.

1.1.3 Big data

El concepto *big data* (se usa en inglés aunque en español podría traducirse como datos masivos) se refiere a la captura y tratamiento de un gran conjunto de datos con variedad de formatos y una velocidad de generación de nuevos valores que supera la capacidad de los sistemas de hardware y software convencionales para procesarlos. Son las tres uves del *big data* (3V): Volumen, Variedad y Velocidad.

Figura 1.2. Datos estructurados vs. datos no estructurados.

Su nacimiento coincide con la la automatización del censo americano por la compañía IBM a finales del siglo XIX, antes incluso de la aparición de los ordenadores. Después, gracias al abaratamiento de los costes de procesamiento y almacenamiento del hardware, hemos podido aumentar el rango de análisis de datos estructurados (números) y empezar a trabajar con datos no estructurados (imágenes, audio, vídeo o texto) como se recoge en la figura 1.2. Amazon, Google, Amazon y Microsoft han desarollado un papel fundamental en la evolución del tratamiento de este tipo de datos por la aparición de sus servicios de computación en la nube.

No existe una definición exacta de qué tamaño, cuántos tipos distintos o a qué velocidad deben producirse para que un conjunto de datos sea considerado *big data*. Imaginemos, por ejemplo, la facturación de chaquetas en un año de El Corte Inglés, los expedientes médicos del hospital de tu ciudad o las transacciones

electrónicas mensuales de un banco a nivel nacional. Pese al gran volumen de datos, estos conjuntos no son estrictamente *big data* porque no cumplen el criterio de variedad ni de velocidad y aunque aplicaremos técnicas de aprendizaje máquina o incluso aprendizaje profundo (ambos dentro de la inteligencia artificial) no será *big data.*

Al analizar *big data* no sólo los ordenadores, ni siquiera los algoritmos o las bases de datos habituales que utilizamos son capaces de obtener buenos resultados: porque no caben, porque no da tiempo o porque los datos son un caos. Para poder procesar y analizar *big data*, se necesitan herramientas y tecnologías especializadas. Esto incluye algoritmos de aprendizaje automático, bases de datos distribuidas y plataformas de procesamiento en paralelo. También se necesitan profesionales altamente cualificados con habilidades en ciencia de datos, ingeniería de software y otras disciplinas relacionadas que se reciclen continuamente: el *big data,* con su crecimiento exponencial y la proliferación constante de nuevos formatos, nos urge a encontrar formas más óptimas de recoger, almacenar y procesar. Más rápido y más barato: el tiempo es oro en nuestra sociedad y los datos, el nuevo petróleo.

El análisis de *big data* puede proporcionar una amplia variedad de beneficios para las empresas. A nivel interno, pueden mejorar la eficiencia de sus operaciones y tomar decisiones más informadas. A nivel externo, proporcionar un mejor servicio a sus clientes, identificar nuevas oportunidades de negocio y desarrollar productos o servicios innovadores. Por eso, la capacidad de analizar *big data* en la actualidad es crucial para el funcionamiento eficiente y el éxito de las organizaciones.

Un claro ejemplo de uso de *big data* es el análisis de transacciones comerciales. Las empresas, al disponer de los datos de compra de sus clientes en su *e-commerce* y en sus tiendas físicas, conocen no sólo los productos que se venden, su precio y cantidad sino también disponen de la tipología de los compradores, de fotos del artículo y de las opiniones de redes sociales (RRSS). Si implantan este tipo de análisis, obtendrán una comprensión más profunda de sus clientes y de cómo están utilizando sus productos para desarrollar estrategias de marketing más efectivas, descubrir si hay una demanda insatisfecha o no cubierta para un determinado tipo de producto y tomar decisiones de inversión basadas en datos.

Otro ejemplo más cercano es el recomendador de Netflix que utiliza técnicas de *big data* para predecir, en función de lo visualizado por los demás usuarios y de tu gusto particular, qué propuesta hacerte, qué serie será la más vista, y lo más importante para el funcionamiento correcto de la aplicación, la carga de procesamiento y el nivel de incidencias de los servidores para poder estimar y distribuir de forma óptima los recursos.

1.1.4 Minería de datos

La minería de datos (en ingles, *data mining*) hace referencia a todas las técnicas que permiten extraer conocimiento de un conjunto de datos para descubrir, si es que existen, relaciones, modelos, regularidades o patrones que subyacen en un conjunto de datos y que, a priori, desconocemos.

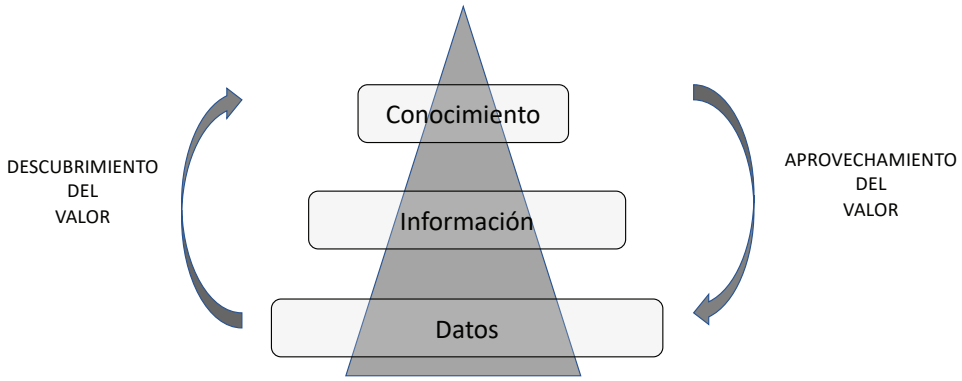

Figura 1.3. Cadena de valor del dato.

En el lenguaje habitual, la minería es un proceso por el que excavamos para intentar encontrar algún mineral de valor. Si imaginamos los datos como una enorme montaña de información, la minería de datos sería el proceso de tratamiento de esa montaña para encontrar trozos de valor o conocimiento y así lo representamos en la figura 1.3.

La minería de datos y el aprendizaje automático se utilizan de manera intercambiable pero son conceptos diferentes. La minería de datos se refiere al proceso de extraer información valiosa y relevante de (grandes) conjuntos de datos. Abarca técnicas de limpieza, transformación y análisis de datos con el fin de descubrir patrones y características específicas. Mientras que el aprendizaje automático engloba los algoritmos y modelos para conseguir que las máquinas aprendan y resuelvan tareas automáticamente.

1.1.5 Algoritmos y modelos

Los algoritmos son un conjunto de pasos matemáticos, la descripción de unas instrucciones para conseguir un objetivo, en nuestro caso, que la máquina aprenda. Sería el equivalente a la receta de cocina donde se enumeran los ingredientes y los pasos a seguir, los tiempos estimados y el resultado esperado. Existen muchos

algoritmos de aprendizaje máquina que se encuadran, de forma clásica, en algoritmos de aprendizaje supervisado, no supervisado y reforzado. En los últimos tiempos han aparecido nuevas clases, los semi-supervisados y auto-supervisados. La elección de unos u otros depende del tipo de problema que se quiera resolver y de los datos disponibles a analizar y lo veremos en los siguientes capítulos.

Un modelo de aprendizaje máquina consta de dos partes, su estructura (la formulación matemática del modelo, por ejemplo, la profundidad del árbol de decisión o la arquitectura de una red neuronal) y un algoritmo de aprendizaje. El algoritmo utiliza nuestros datos para que el modelo funcione como se desee. Una vez elegido el algoritmo y realizado el proceso de entrenamiento sobre nuestros datos (proceso que describiremos en el punto 2 de este capítulo) obtendremos unos parámetros ajustados a nuestro problema. Este modelo y sus parámetros son capaces de generar respuestas correctas ante nuevos datos.

1.1.6 Parámetros e hiperparámetros

Como hemos explicado anteriormente, los parámetros son lo que caracteriza a nuestro modelo, no son fijados ni predeterminados manualmente por el científico de datos sino que son el resultado del aprendizaje de la máquina, del proceso de ajustar el modelo a nuestros datos de entrenamiento. De su calidad depende la capacidad de nuestro modelo de resolver un problema y, predecir o segmentar correctamente una nueva entrada. Uno de los métodos más usados para estimarlos es el descenso por gradiente, un algoritmo de optimización que veremos en detalle al explicar la regresión lineal (durante la estimación de sus coeficientes), los modelo SVM (durante la estimación de los vectores soporte) o las redes neuronales (durante la estimación de los pesos de la red).

Los hiperparámetros son los valores que los científicos de datos asignamos a la configuración del modelo durante el proceso de entrenamiento. Algunos ejemplos son el tamaño del conjunto de entrenamiento (un 70% es recomendable pero no siempre se puede utilizar este tamaño), el número de iteraciones realizadas durante la fase de entrenamiento y otros coeficientes específicos del modelo. Como no se conocen a priori, inicialmente hay que utilizar unos valores genéricos o basarse en lo realizado en proyectos similares anteriores o actuar según la experiencia. Ajustar los hiperparámetros es una tarea crucial con impacto en el rendimiento final del modelo. Como veremos en el punto 2 el conjunto de datos se dividirá en tres partes: entrenamiento, validación y test. De esta manera, podremos utilizar el conjunto de validación, un conjunto de datos independiente, para evaluar y elegir los hiperparámetros óptimos. El objetivo final es mantener el conjunto de test separado de todo el proceso de estimación, tanto de parámetros como de hiperparámetros.

Una de las técnicas más sencillas para optimizar los hiperparámetros es la búsqueda en cuadrícula. Ésta consiste en definir un rango de valores para cada hiperparámetro y escoger aquella combinación entre todas las posibles que resulten en el mejor rendimiento del modelo. Pero las múltiples opciones pueden alargar este proceso más de lo necesario. Por eso utilizamos alternativas como la búsqueda aleatoria que consiste en probar combinaciones aleatorias de hiperparámetros, la búsqueda basada en métodos bayesianos (más compleja) o los algoritmos evolutivos, como los algoritmos genéticos. Estas técnicas no son exhaustivas, pero son más rápidas en encontrar los hiperparámetros óptimos. Además, la búsqueda basada en métodos bayesianos nos permite aproximar la distribución de probabilidad de los hiperparámetros óptimos, lo que nos ayuda a comprender cómo afecta cada uno de ellos al rendimiento del modelo. En el caso de los algoritmos evolutivos, son muy efectivos cuando los hiperparámetros tienen una gran cantidad de interacciones complejas entre ellos.

1.1.7 Aprendizaje máquina o automático

El aprendizaje máquina o automático (*machine learning, ML,* en inglés) es la disciplina dentro de la ciencia de datos que permite que las máquinas aprendan sin ser programadas con reglas específicas. Aplica la estadística para inferir propiedades y otros métodos matemáticos para detectar patrones en los datos y, a partir de ahí, hacer predicciones e incluso tomar decisiones.

Por ejemplo, cuando tecleas en Google "noticias de última hora" aparece una lista con todos los resultados de búsqueda en tiempo real. Cada cierto tiempo los resultados de esa lista se actualizan basados en el número de clicks que recibe cada una de las páginas. El algoritmo de Google, basado en aprendizaje máquina, reconocerá las preferencias de los usuarios y moverá las entradas en el ranking. Y lo interesante es que no hay ninguna persona que controle ese movimiento: el algoritmo evalúa y adapta la ordenación"aprendiendo" del comportamiento humano.

Los algoritmos de aprendizaje máquina existen desde hace varias décadas, pero el desarrollo de la tecnología, el incremento del poder de cálculo y de almacenamiento de datos, ha hecho posible su presencia universal no sólo en ordenadores, sino en pequeños dispositivos electrónicos (teléfonos móviles o incluso microprocesadores). Este uso masivo ha mejorado su rendimiento y ampliado el rango de tareas que realizan de forma óptima, como la lectura comprensiva, la traducción o la escritura, el reconocimiento de vídeo, la identificación de objetos…En ciertas actividades incluso superan nuestras capacidades: en tareas muy repetitivas, en el manejo de muchas variables simultáneamente y en la identificación de patrones en conjuntos de datos muy grandes.

Supongamos la siguiente secuencia de pares (0,0) (3,6) (6,12) (9,18) (12, ¿?) Para un humano, es bastante sencillo identificar cuál sería la pareja del número que iría en la quinta línea. Como el segundo número siempre es el doble del primero, inferimos que la cifra esperada es 24. ¿Qué pasaría si en lugar de tener estos datos tuviéramos 200.000 filas de todas las transacciones hechas con tarjetas de crédito de un banco? ¿Seríamos capaces a simple vista de detectar cuál es la transacción fraudulenta? Al ser humano le resulta muy difícil procesar esa gran cantidad de datos. Es aquí donde los ordenadores y el aprendizaje máquina hacen un trabajo excepcional: además de proporcionar una respuesta correcta, será mucho más rápido, incluso proporcionando la decisión de autorizar o no dicha transacción en tiempo real.

La clave es que estos algoritmos pueden llegar a desarrollar criterios óptimos para tomar decisiones, evolucionando y mejorando en función de los datos que reciben. No es un aprendizaje de inferencia como el humano, pero a mayor número de datos procesados sí se produce una mejora automática de los parámetros del modelo que llevará a mejores resultados.

El aprendizaje máquina, como cualquier programa informático, necesita un humano que lo programe y supervise, éste es el trabajo de los científicos de datos. Si entrenamos un algoritmo para reconocer imágenes de gatos, el algoritmo aprenderá a detectar gatos, sin programar mediante reglas. Pero será incapaz de detectar correctamente perros u otros animales que no sean gatos, eso no es lo que ha aprendido. El algoritmo no es capaz de aprender lo que es un gato o lo que es un perro. Aunque, a mayor número de imágenes analizadas, más capacidad tendrá de identificar un gato entre imágenes de zorros o de cachorros de tigre. El científico de datos elige qué algoritmo es el más adecuado para resolver el problema con los datos disponibles y de configurarlo matemáticamente, ajustando parámetros y minimizando funciones de error, para procesar mejor los datos y con mejores resultados.

1.1.8 Aprendizaje profundo

El aprendizaje profundo (*deep learning, DL* en inglés) se popularizó en 2012, cuando las grandes compañías hicieron públicos los excelentes resultados de la aplicación de redes neuronales al análisis de imágenes y voz principalmente. Pero están con nosotros desde hace más de 50 años.

Las redes neuronales artificiales (*artificial neural network, ANN* en inglés), o simplemente redes neuronales, analizan los datos mediante capas de procesamiento, tal y como hace el cerebro humano, donde los datos se van procesando a través de distintas capas de neuronas. Las redes neuronales se representan típicamente como

puntos interconectados. Cada conexión tiene un valor o peso numérico (parámetro) que se va modificando en base a la experiencia.

Los datos, números, imágenes o sonido, entrarían por la primera capa y se distribuirían entre las neuronas de esa capa que harían un primer procesado y los enviarían a la siguiente capa. A medida que los datos van pasando de capa a capa, queda menos de la imagen o sonido original y quedan datos más abstractos, información útil: cada capa aprende de la capa anterior.

La red neuronal más simple tendría tres capas: capa de entrada, capa de procesamiento o capa oculta, y capa de salida. Los datos entran por la capa de neuronas de entrada. La capa oculta, una o varias, procesan los datos para abstraer las características que queremos. El resultado final lo muestra la capa de salida.

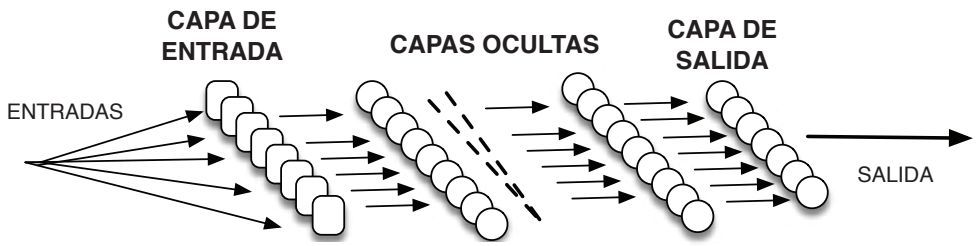

Figura 1.4. Estructura típica de una red neuronal.

Una red neuronal aprende mediante el ajuste de sus parámetros. Como se ve en la figura 1.4, cada neurona de la red se interconecta con las demás mediante unas conexiones. Dichas conexiones tienen unos parámetros asociados, valores o pesos, que se inicializan de manera aleatoria pues no tienen otra información, serían como el cerebro de un bebé. Cuando le proporcionamos el primer dato a la red neuronal, con esos pesos iniciales, nos dará una salida distinta a la realidad. El algoritmo matemático será el que irá ajustando dichos pesos hasta que el resultado sea correcto. Esto es lo que se conoce como entrenamiento de la red. Cuando la red prácticamente acierta casi todos los ejemplos, podemos enfrentarla a nuevos datos reales.

Cuando las redes neuronales empezaron a aplicarse a la minería de datos tenían muy pocas neuronas y capas ocultas porque los ordenadores de la época estaban limitados para procesar tantas conexiones y parámetros neuronales. A medida que la capacidad de computación se ha ido abaratando, las redes neuronales han incrementado el número de capas y conexiones (también denominados grados de libertad) con un mayor grado de ajuste al actualizarse, dando respuestas más fiables.

1.1.9 Infraestructura y aplicaciones. Servicios en la nube

Hasta hace relativamente poco, con el concepto *hardware* nos referíamos a ordenadores físicos o servidores, en una habitación o sala específica para ellos. Actualmente lo habitual son ordenadores en la nube. Amazon, Microsoft, IBM o Google nos ofrecen sus servicios de computación sin tener acceso a su hardware físico. Igual que al pagar por la electricidad, ahora pagaremos por el uso de máquinas virtuales. La nube ofrece disponibilidad total para consumir potencia de computación y almacenamiento según nuestras necesidades. Estos servicios se denominan IaaS (infraestructure as a service).

Las empresas privadas e incluso la administración pública están reemplazando el hardware tradicional por infraestructura en la nube. Han eliminado sus propios centros de datos y servidores; ahora alquilan la capacidad de computación y almacenamiento, dejando atrás los problemas de mantenimiento de hardware o la obsolescencia técnica ajustando dinámicamente su demanda a las necesidades del momento. Además, la tecnología en la nube libera a los científicos de datos de la configuración del hardware, permitiendo que se centren en los datos y la optimización de los algoritmos. Eso sí, limitado a las herramientas que el proveedor deje disponibles.

Además de los servicios de computación en la nube también existen servicios de *software*, bases de datos o herramientas de análisis virtuales ofrecidas por esos mismos proveedores que hacen más sencillo y accesible el procesamiento de datos (ya no hay que alquilar una máquina virtual para montar de cero tu BBDD, sino que te ofrecen dicho servicio solo o combinado con otros). Estos servicios se denominan PaaS y SaaS (Platform/Software as a service). Este aumento de oferta y el abaratamiento de los precios han hecho que muchas compañías puedan plantearse aplicar técnicas de aprendizaje máquina o de inteligencia artificial a sus problemas de negocio.

Respecto al software para los científicos de datos, hay dos tendencias: los que programan mediante línea de comandos (Python y R son los lenguajes más extendidos) y los que utilizan interfaces gráficas. El uso de lenguajes de programación especialmente orientados a la minería de datos, ofrece una mayor flexibilidad, potencia y adaptación para realizar cualquier tarea. Por otro lado, para iniciarse y evitar los problemas de la programación, las herramientas de data mining con entorno gráfico permiten de una manera más intuitiva realizar el preprocesado de los datos, el análisis y la configuración de los algoritmos (Rapidminer, WEKA, Orange o KNIME, de uso libre, y SAS, en su versión *freeware* y de pago). Existen también alternativas de *business intelligence* (Tableau, Power BI, Qlik view) cada vez más potentes con nuevos módulos específicos de ML para predecir y segmentar.

El uso de una u otra alternativa depende de factores como la disponibilidad de esas herramientas en las plataformas de trabajo, la necesidad de profundizar en los algoritmos o de utilizar algoritmos estándar, los conocimientos del programador en ciencia de datos y de un factor que hasta ahora no habíamos nombrado, el grado de explicabilidad que deseamos que tenga el modelo.

1.2 ANÁLISIS DE DATOS. ETAPAS

Lo primero es identificar las variables y patrones contenidos en nuestros datos. Por ejemplo, en un registro médico, un paciente sería un patrón y cada característica (el peso o la altura) una variable; así María sería un patrón y su peso 58 kg, sería su valor de la variable. Para un caso de segmentación de imágenes entre perros y gatos, cada imagen sería un patrón y las variables serían los *pixeles*.

En la tabla 1.1 aparece el resumen donde cada fila sería un patrón (paciente) y cada columna es una variable (peso, altura, etc.).

Paciente	Peso	Altura	Hemoglobina (g/dl)
María	58	1,62	11
Quique	75	1,80	10
Alicia	64	1,65	11,2
Pablo	75	1,83	10,7
Merche	65	1,80	10,3
Jorge	72	1,79	12

Tabla 1.1. Tabla de análisis de datos médicos.

Una vez que tenemos recogidos los datos a analizar, con sus patrones y variables iniciamos el proceso de análisis de datos según la figura 1.5.

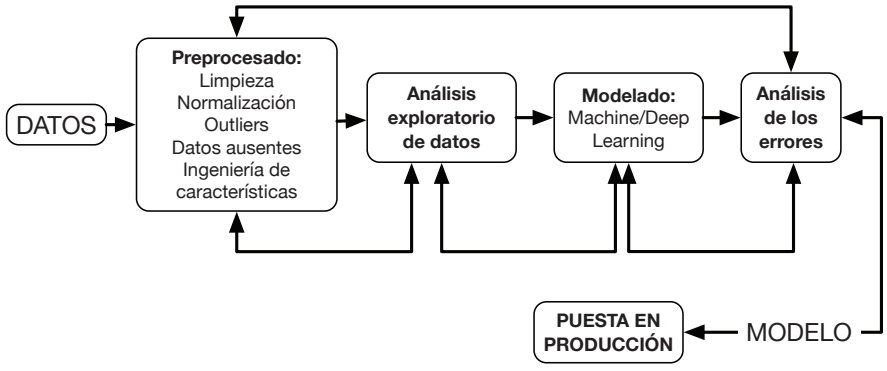

Figura 1.5. Etapas del análisis de datos.

1.2.1 Datos

Conocer nuestros datos es el punto de partida de cualquier problema basado en datos. Tendremos que clasificar nuestras variables entre categóricas o continuas por el tratamiento posterior (distinto) que les vamos a dar.

- Categóricas: por ejemplo, los colores, porque sólo existen una cantidad finita de opciones (rojo, verde, negro, etc.). O el 'nivel de dolor' porque lo definimos como bajo, medio, alto. En el primer caso, se denomina variable categórica nominal (da igual el orden entre los valores) y en el segundo, variable categórica ordinal (porque sabemos que el medio está entre el alto y el bajo y el alto arriba del todo).

- Continuas: existen (casi) infinitas posibilidades para estas variables. Sería el número de horas que pasamos en una red social, la tasa de alcohol en sangre, o el nivel de ingresos por ciudadano.

1.2.2 Preprocesado

Ésta es la fase a la que dedicaremos el 80% del tiempo total invertido en la resolución de cualquier problema basado en datos.

1. *Limpieza de datos*. Un conjunto de datos "ideal" es la tabla 1.1. donde cada variable tiene un valor (no hay datos ausentes, en inglés *missing values)*, no hay datos incorrectos y están en un rango similar (no hay ruido ni datos atípicos, en inglés *outliers*).

   Diferencia entre Dato Incorrecto y Dato Atípico *(outlier en inglés)*: en la variable altura del paciente, un dato incorrecto sería 4 metros (es imposible) mientras que un dato atípico sería 2m. (no es común). Otro ejemplo muy frecuente de dato incorrecto sería tener simultáneamente en la misma variable altura, unos valores expresados en centímetros (162, 202, ...) y otros en metros (1,54; 2,05). Los datos incorrectos hay que eliminarlos porque producen errores graves en las conclusiones obtenidas. La forma más sencilla de detectarlos es conocer bien la variable y establecer sus umbrales máximos y mínimos. En el caso de tratar imágenes o series temporales, donde puede existir interferencias por una relación temporal o espacial entre las variables, habrá que aplicar técnicas de procesado de señales/imágenes para reducir o eliminar dicho ruido.

2. *Normalización.* La diferencia de rangos entre variables puede impactar negativamente el resultado de algunos modelos de aprendizaje máquina Si tenemos una variable con valores entre 0 y 106 y otra con valores entre