AUTORES

Emilio Soria Olivas. Catedrático de Universidad, Licenciado en Físicas y Doctor Ingeniero Electrónico. Es 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 (Universidad Autónoma de Madrid). Senior Mgmt. Program (IE). IA para la empresa (MBIT). Master Data Analytics (EDEM).

Borja Castillo Caballero. Graduado en matemáticas por la Universidad de Valencia. Realizó un máster en Inteligencia Artificial en la Universidad de Valencia y un 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. Estudió el máster en Inteligencia Artificial por la Universidad de Valencia. Trabaja como científico de datos.

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 de los datos 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 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 mediante el uso de las técnicas de Aprendizaje Máquina/Profundo 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 de del Aprendizaje Máquina/Profundo 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 Máquina/Profundo 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 Máquina 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 ¿son las máquinas más inteligentes que nosotros?

Un rasgo de la inteligencia es la capacidad de aprender ¿aprenden las máquinas? Las máquinas no aprenden, 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. De ahí el término *machine learning*: las máquinas son capaces de crear respuestas adecuadas ante datos desconocidos.

¿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.

¡Empezamos!

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. Hoy los métodos estadísticos siguen siendo muy utilizados por los científicos de datos. 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, la oficina del censo de Estados Unidos fue incapaz de tener todos los datos del censo entre un período y otro. La recogida de los datos y el procesado de éstos era tan lenta que no daba tiempo. La cantidad de datos era tan grande que se hizo inmanejable. Ése fue uno de los primeros problemas de Big Data de la historia y que no podía ser resuelto por la tecnología de aquel tiempo. 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? Son nuevos problemas que han desembocado en nuevas soluciones para recoger estos datos, procesarlos y poder extraer conclusiones.

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.

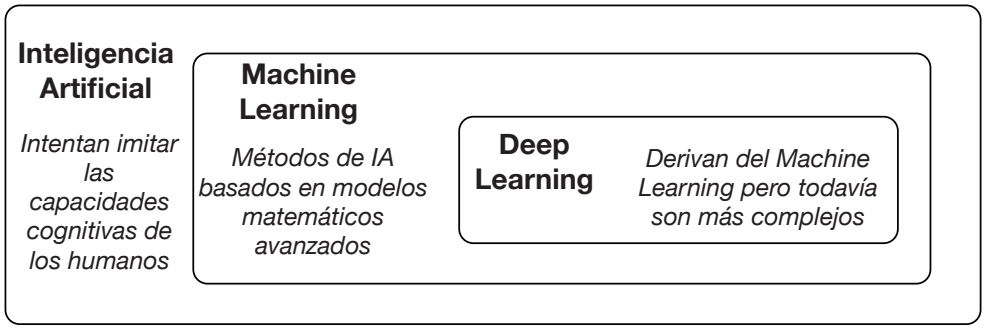

Figura 1. Disciplinas de la ciencia de datos

1.1.2 Inteligencia artificial

La inteligencia artificial, IA, (*AI o artificial intelligence* en inglés), se ha desarrollado a la vez que la informática. En realidad, 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. La IA se definió entonces como las máquinas capaces de realizar tareas que, en el caso de ser hechas por un humano, requieren cierta inteligencia.

Pero dotar a una máquina de una inteligencia similar a la humana, es decir una inteligencia general es un objetivo muy ambicioso, comparable a explicar el origen del universo. 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*. Ocurre algo parecido con otros conceptos de la ciencia de datos como minería de datos o *big data*. Aclararemos las diferencias en los siguientes párrafos.

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 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). 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 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: nos permitirán descubrir, si es que existen, relaciones, modelos, regularidades o patrones que subyacen en un conjunto de datos y que, a priori, desconocemos.

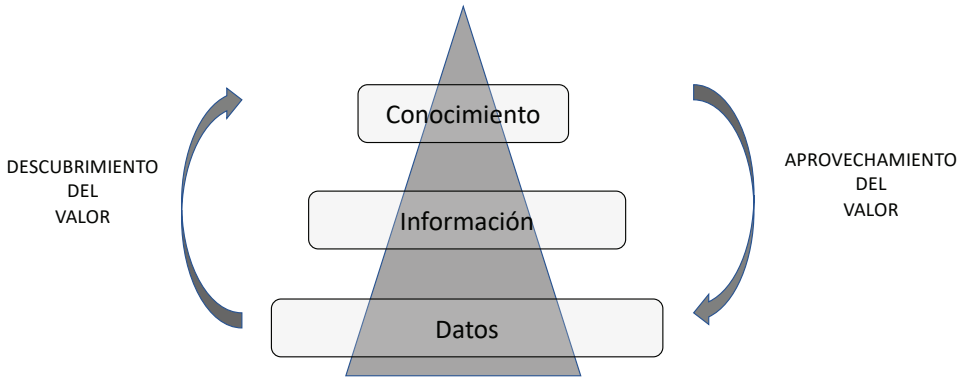

Figura 3. Cadena del valor del dato hasta llegar al conocimiento

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 3.

En este caso también, 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 algortimos 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 Bayes (más compleja) o los algoritmos evolutivos, como el algoritmo genético. Estas técnicas no son exhaustivas, pero son más rápidas en encontrar los hiperparámetros óptimos. Además, la búsqueda basada en Bayes 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, basándose en los datos proporcionados, 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.

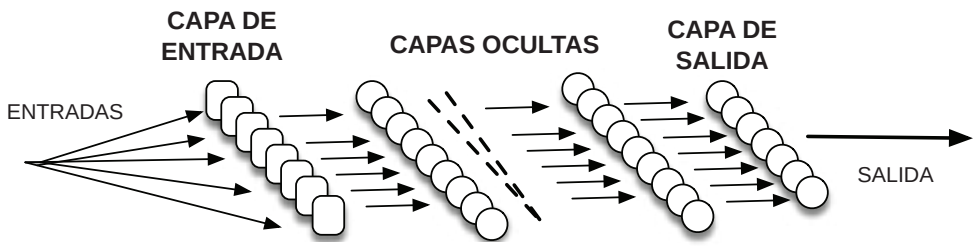

Figura 4. Estructura típica de una red neuronal

Una red neuronal aprende mediante el ajuste de sus parámetros. Como se ve en la imagen, 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 de test, podemos enfrentarla a nuevos datos reales. De ahí la similitud con el funcionamiento del cerebro humano.

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 (de ahí que también se denominen 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 virtuales en la nube Amazon, Microsoft, IBM o Google nos ofrecen sus servicios de computación de forma virtual 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 mensualmente o mediante suscripciones la capacidad de computación y almacenamiento, dejando atrás los problemas de mantenimiento de hardware o la obsolescencia técnica y ajustar dinámicamente su contrato 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 y a negocios.

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, ofrecen 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 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 (IA explicable).

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 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. 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 5.

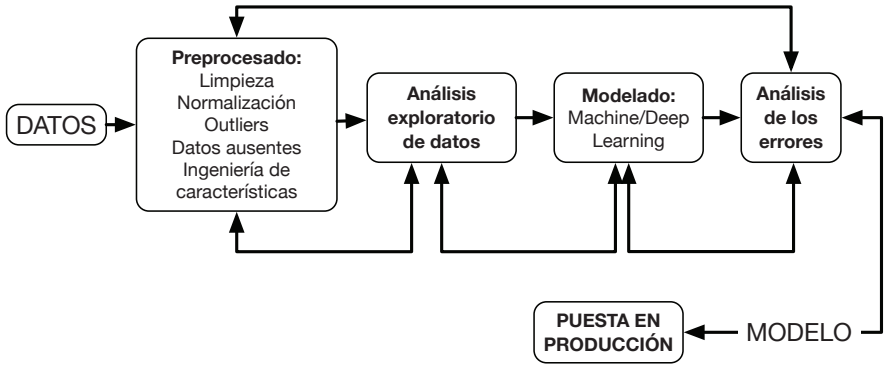

Figura 5. Etapas del análisis de datos

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: son 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. 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

0 y 10-3, la mayoría de los algoritmos dará más importancia a la primera variable más que a la segunda no por su importancia en el problema a resolver sino por el rango de sus valores, por lo que será necesario normalizar las variables antes de empezar a trabajar. La regresión lineal y los árboles de decisión son robustos a este hecho. Uno de los modelos que sufren de este tipo de problemas es el perceptrón multicapa (en ese capítulo se profundiza en el tema).

3. *Detección/caracterización de outliers.* Como hemos explicado en el primer punto, los datos atípicos (*outliers)* pueden llevar a error en las conclusiones, haciendo inestable el modelo matemático. En el caso de los modelos lineales (se ajustan a la media de los valores) la existencia de un valor extremo le afectaría muy negativamente (desvirtuaría esa media) y, el modelo, ante datos nuevos, dará un resultado menos exacto. La figura 6 lo explica visualmente: el punto redondo es el atípico en comparación con el resto de puntos que son "x". El modelo A, sin *outlier*, quedaría mucho más ajustado que el modelo B (con) a la media de la mayoría de los valores.

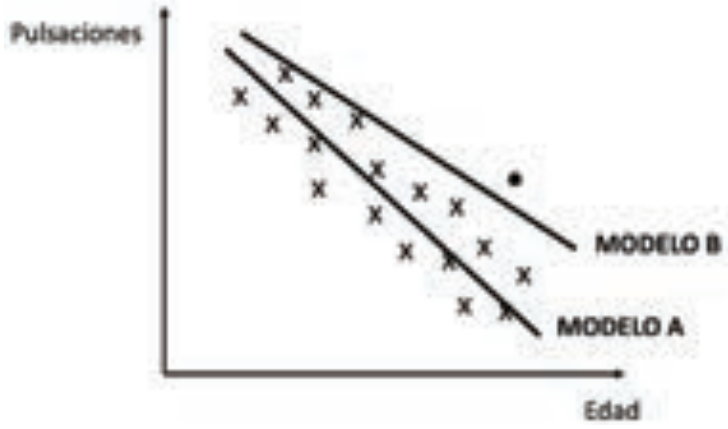

Figura 6. Modificación de un modelo por la presencia de *outliers.*

Para detectar si hay *outliers* en las variables, una aproximación es observar la diferencia entre la media y la mediana de la variable a analizar y, si toma un valor muy diferente a cero, existen *outliers*. Otro procedimiento clásico es definir un rango $[m - 3 \cdot \sigma, m + 3 \cdot \sigma]$ y los *outliers* son los valores fuera del rango (donde m es el valor medio de la variable y σ es su desviación estándar.

Si lo que buscamos son los patrones atípicos, el análisis se complica. Podemos, por ejemplo, utilizar algoritmos de agrupamiento para detectarlos (se generan grupos y los datos atípicos serían los que quedasen fuera). Los algoritmos empleados con este fin se denominan genéricamente 'algoritmos de detección/caracterización de anomalías' y han cobrado una relevancia especial en los últimos años por su aplicación en ciberseguridad, mantenimiento predictivo, reconocimiento de fraudes, detección de *fake news*, etc.

4. *Datos ausentes*. En una determinada variable puede faltar un valor (por ejemplo, si estuviera vacío el campo altura en el patrón María en la tabla 1). Aunque hay algún modelo que puede funcionar con datos ausentes, como los gráficos probabilísticos, la gran mayoría no. Al igual que ocurre con los *outliers* no trataremos igual la falta de valores en la variable o en los patrones.

   Si analizamos las variables, y es una variable discreta, sustituiremos el valor ausente por la moda (el valor de dicha variable que más se repite). Si la variable es continua, lo sustituiremos por la media (o la mediana).

   En cualquier caso, hay que evaluar el porcentaje de datos faltantes porque con estos métodos estamos configurando una realidad 'modificada' que puede ser distinta de la realidad 'real', afectando los resultados del modelo.

5. *Ingeniería de características*. Es una etapa fundamental para los modelos de aprendizaje máquina (y para vuestra información, no lo es -a priori- para los de aprendizaje profundo, de ahí su éxito). Durante esta etapa lo que haremos es a) *crear/eliminar variables*, volviendo al caso anterior, podríamos crear una nueva variable, el IMC o índice de masa corporal y eliminar dos, el peso y la altura; b) *seleccionar/descartar variables*, se escogen aquellas variables relevantes para el modelo a desarrollar y c) *analizar las entradas*, revisamos cada entrada y su importancia porque ésa será la información que utilice el modelo para su ajuste.

1.2.3 Análisis exploratorio de datos (EDA)

Es el momento de captar la información que contienen nuestros datos. En las variables continuas, lo haremos mediante el cálculo de los estadísticos descriptivos más comunes y en las variables discretas, por sus frecuencias. Evaluaremos, mediante contrastes de hipótesis (estadística inferencial), la posible existencia de relaciones lineales o, en caso contrario, la independencia entre las variables y su

normalidad, que visualizaremos para extraer mejores conclusiones. Para finalizar, aplicaremos algoritmos de agrupamiento (*clustering*) para estudiar las posibles zonas de alta densidad de patrones, es decir, comportamientos similares en los datos que nos ayudará a simplificar el problema a resolver.

1.2.4 Modelado

El modelo que desarrollaremos, o de aprendizaje máquina o de aprendizaje profundo, dependerá de la cantidad de datos y del problema a resolver. Y como no existe un único modelo perfecto (teorema *'No free lunch'*) habrá que buscar el óptimo teniendo en cuenta nuestro objetivo y dado el conjunto de datos sin sobreajustar.

El sobreajuste ocurre cuando los modelos memorizan los datos y sus resultados y, ante datos nuevos, no genera resultados correctos. Se lo sabe al pie de la letra, es decir, no es capaz de generalizar.

Uno de los pasos imprescindibles es comprobar la capacidad de generalización de nuestro modelo.

1. La forma más sencilla es mediante la *creación de dos subconjuntos* a partir de nuestro conjunto original de datos: los datos de entrenamiento (70-90% de los datos, según el volumen disponible) y los datos de test (el resto) según la figura 7.

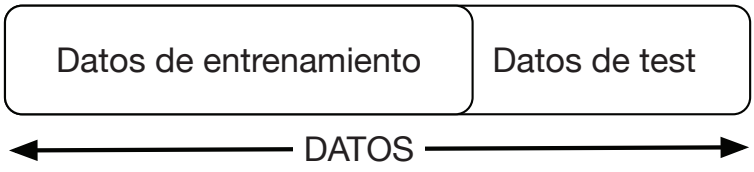

Figura 7. División del conjunto de datos en entrenamiento y test.

Como regla general que no debemos olvidar en cualquier proceso de toma de decisiones basadas en datos, el conjunto de test nunca se utiliza para nada. Una vez realizada la división, nos olvidaremos de ellos y sólo los rescataremos cuando queramos comprobar la eficacia de nuestro modelo entrenado ante datos no observados, para 'testear' su capacidad de generalización. Este potencial problema se conoce como fuga de información (*leakage information* en inglés).

2. Podemos mejorar la estrategia anterior *dividiendo el conjunto de datos en tres*: los datos de entrenamiento se dividen ahora en dos subconjuntos, un conjunto de entrenamiento 'puro' y un conjunto de validación (con una proporción general de 80-20%) según la figura 8. Estos datos de validación se utilizarán para estimar el sobreajuste y el error que puede cometer el modelo.

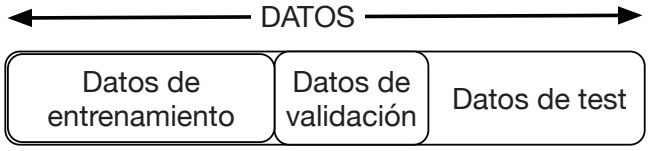

Figura 8. División del conjunto de datos en entrenamiento/validación y test.

La función de error J recogida en la figura 9 mide el ajuste del modelo para el conjunto de entrenamiento y para el conjunto de validación. El error de entrenamiento va disminuyendo al igual que el de validación ajustándose a los datos y mejorando su capacidad de generalización con el paso del tiempo. En el momento en que el error de validación empieza a aumentar es la señal de que el modelo está memorizando los datos de entrenamiento y, por tanto, a perder capacidad de generalización. Debemos elegir el modelo en el mínimo de la función porque ahí tendrá la máxima capacidad de generalización. Esta representación muestra un punto muy claro de "parada del entrenamiento" porque es una simplificación, pero, en la realidad, hay muchas oscilaciones y será necesario utilizar el promedio de varios puntos (por ejemplo) para suavizar dichas oscilaciones.

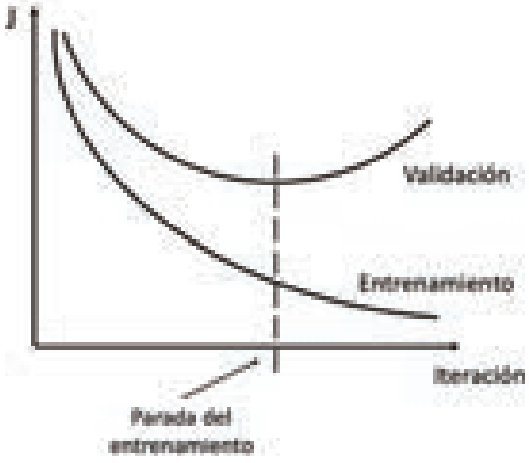

Figura 9. Representación del ajuste del modelo medido por la función J. La 'parada del entrenamiento' se produce al encontrar el modelo óptimo.

En la figura 10 podemos observar el 'modelo original' de inicio que no se ajusta a los datos. Si continuamos entrenando, llegaremos a un 'modelo óptimo' que se acerca al conjunto de entrenamiento representado por cruces (x) y de validación (representado por n) y se correspondería con el mínimo de la función de validación en la figura 1.8. En el caso de que continuáramos ajustando, llegaríamos al 'modelo sobreajustado' que se solapa con los datos de entrenamiento, pero se aleja de los datos de validación, es el problema del sobreajuste explicado anteriormente.

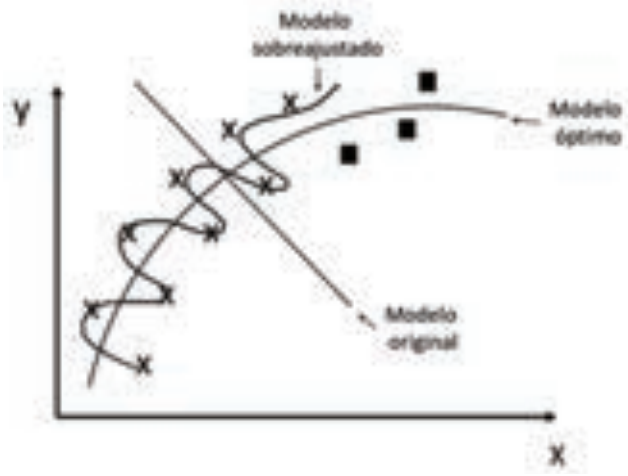

Figura 10. Evolución del ajuste de un modelo. Las cruces (x) representan el conjunto de entrenamiento y los cuadrados (n) el conjunto de validación.

3. Otra opción para evaluar la capacidad de generalización del modelo es mediante *k-folds*. Este método está indicado cuando el conjunto de datos tiene pocos patrones pues proporciona medidas de error más robustas.

   En un primer paso dividiremos el conjunto de datos en dos subconjuntos, entrenamiento y test. Y en un segundo paso, el conjunto de entrenamiento se subdividirá en conjunto de entrenamiento y de validación, tal y como se ha explicado en el punto 2. En un tercer paso (y ésta es la novedad) dividiremos los conjuntos de entrenamiento y validación en subconjuntos (*folds*) de tal manera que 1 subconjunto se usa para validar y el resto para entrenar. Si se divide en 5-*folds* (k=5), según la figura 11, significa que obtendremos 5 valores para la función que queremos optimizar en el conjunto de entrenamiento y 5 valores para el conjunto de validación. Esos valores se promedian para obtener una medida única para los dos conjuntos, entrenamiento y validación.

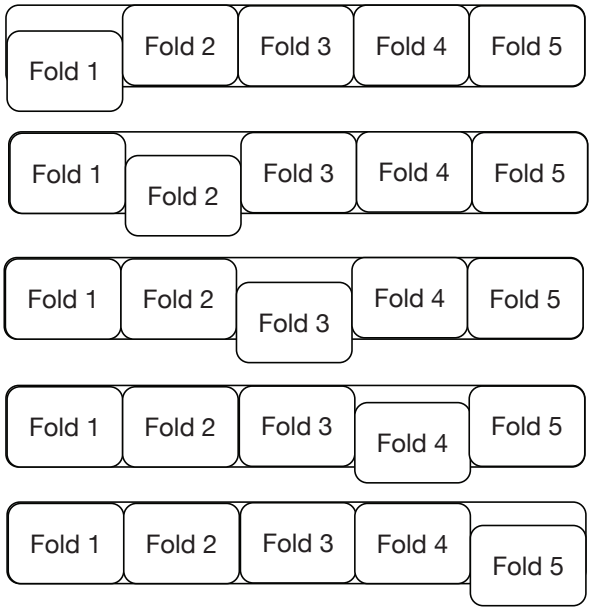

Figura 11. Datos divididos por el método *k-fold (5-fold en este caso).* Los *folds* alineados forman el conjunto de entrenamiento y los *folds* no alineados, el de validación

Este procedimiento se suele repetir varias veces. Como la segmentación de los diferentes subconjuntos creados cada vez es aleatoria y, por tanto, diferente, aumentamos así la fiabilidad de las medidas de funcionamiento obtenidas.

Hemos hecho hincapié en que el subconjunto de test no se debe utilizar en ningún momento durante el ajuste del modelo. Igual ocurre con el subconjunto de validación: cuando estamos en la fase de desarrollo del modelo usando los datos de entrenamiento, la información de este subconjunto no se debe usar. Pero un punto importante a tener en cuenta es que, si vamos a aplicar una transformación a los datos de entrenamiento para el desarrollo del modelo, también debemos aplicar dicha transformación a los de validación y test para homogeneizar. En cualquier caso y, para hacer esto un poco más fácil, no hay que preocuparse porque los lenguajes de programación utilizados (Python, R) hacen estas divisiones en subconjuntos de forma automática y aíslan los datos que no se pueden usar.

1.2.5 Análisis de los errores

Hemos determinado ya nuestro modelo óptimo, el que minimiza la función de error del conjunto de entrenamiento y validación y maximiza su capacidad de generalización. Ahora vamos a validarlo, analizando la calidad/el rendimiento del

modelo mediante diferentes medidas de error que, además, podremos usar para comparar con otros modelos y elegir, finalmente, el que mejor se adecue a nuestro problema y datos.

Algunos ejemplos, que se verán más en detalle en el punto 5, son la función de coste o error entrópica y el área bajo la curva (AUC). Ambas analizan el sesgo (*bias en inglés)* y la varianza del modelo.

El *sesgo* es la diferencia entre el valor medio predicho por el modelo y el valor real. Según Pedro Domingos, mide la tendencia a aprender de forma sistémica cosas equivocadas y de la misma forma. Un alto sesgo da lugar a modelos subajustados (*underfitting* en inglés) por lo que el sesgo debe ser bajo para asegurar que el modelo se ajusta a la realidad.

La *varianza* mide la tendencia a aprender cosas irrelevantes o muy peculiares para el objetivo. Una varianza alta significa que el modelo está sobreajustado (*overfitting* en inglés) por lo que debe ser baja para asegurar que el modelo tiene la capacidad de generalizar.

La figura 12 es una representación clásica de estas dos variables mediante una diana donde se muestra los posibles resultados del lanzamiento de 6 dardos. En la primera fila se tiene un sesgo bajo y en la segunda fila, alto. En la primera columna tenemos una varianza baja y en la segunda columna, alta. Se observa que un modelo con bajo sesgo/varianza proporciona buenos resultados, mientras que el sesgo desplaza el valor correcto del modelo a una cierta cantidad y la varianza impacta en la dispersión de los datos.

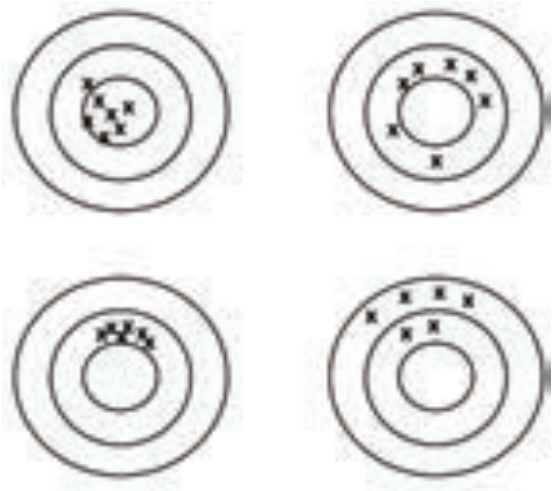

Figura 12. Sesgo y varianza al lanzar 6 dardos

1.2.6 Puesta en producción

Una vez que el conjunto de datos ha pasado todas las etapas previas, deberemos revisar varios puntos antes de que se pueda utilizar de forma masiva en el entorno de la industria/empresa o sector público.

- Proporcionar una excelente experiencia de usuario, contando con experto de esa área (*user experience*, UX en inglés).
- Evaluar la seguridad de acceso al modelo (se puede atacar a los modelos basados en datos).
- Identificar la capacidad del modelo de explicabilidad, imparcialidad, confiabilidad y privacidad. Muy relacionado con las investigaciones actuales sobre la Ética de la IA o del algoritmo.
- Examinar la velocidad de actualización y de respuesta del modelo (fundamental si se tienen datos en *streaming*), el correcto acceso a otras fuentes (clave si las entradas provienen de *web scraping*), etc.

Sin olvidar **el** factor del desplazamiento de los modelos. Con el tiempo, las condiciones del problema que se quiere resolver cambian y el modelo no estaría ajustado a esa nueva casuística. Este problema de analítica se puede resolver con técnicas/algoritmos y no se debe pasar por alto.

1.2.7 Metodología CRISP-DM

Es una metodología nacida en Europa nacida en 1996 desarrollada para proyectos dedicados a extraer el valor de los datos (*Cross Industry Standard Process for Data Mining*). Sus etapas coinciden con las anteriormente explicadas más dos que son clave: establecer la pregunta de negocio a resolver y tener una perfecta comprensión de los datos. Esta metodología *CRISP-DM* aparece, simplificada, en la figura 13.

La pregunta de negocio es la clave y la tiene que plantear el especialista de negocio basándose en tres pilares: ¿Qué se quiere resolver? ¿Cómo se va a comprobar el funcionamiento del modelo (medidas)? ¿Cuáles son los datos relevantes disponibles? La mayoría de los fracasos de los proyectos basados en datos se deben a no haber respondido inicialmente a estas tres preguntas. Todos los casos de éxito en la industria/empresa se corresponden con problemas bien definidos, con datos alineados con el objetivo a resolver, en los que encontrar una solución conllevaba nuevas vías de crecimiento o incluso la desaparición de pérdidas económicas.

Los científicos de datos deben entender en profundidad la pregunta de negocio que se quiere resolver. La aplicación de algoritmos y otras técnicas debe estar basada

en un conocimiento a la perfección del problema y los datos relevantes disponibles. El especialista de recursos humanos, finanzas, operaciones...y el analista de datos deben trabajar en perfecta conexión y retroalimentarse para llevar al éxito del proyecto.

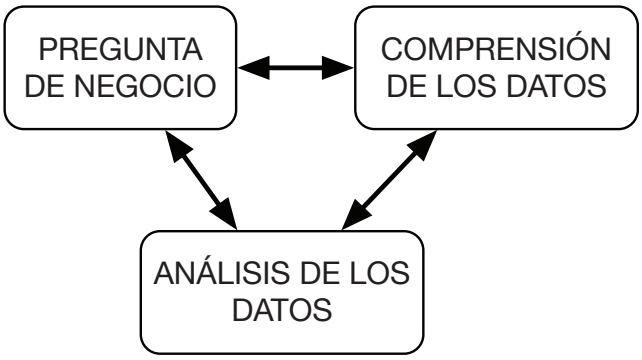

Figura 13. Metodología *CRISP-DM* simplificada.

1.3 ALGORITMOS DE APRENDIZAJE MÁQUINA

Hay muchos tipos diferentes de Aprendizaje Máquina dependiendo de la naturaleza de la tarea T que se desea resolver, la medida de rendimiento P utilizada y la naturaleza de la señal de experiencia E que le proporcionamos (definición de *machine learning* del profesor *Tom Mitchell),* pero de manera clásica los podemos dividir en supervisado, no supervisado y reforzado. También haremos una mención a los más modernos, los semisupervisados (dentro de los no supervisado) y los autosupervisados.

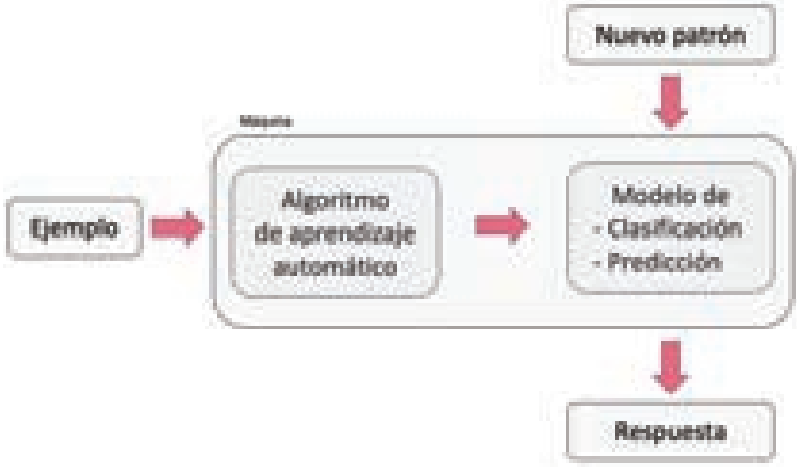

Figura 14. Esquema general de uso de algoritmos de aprendizaje máquina

1.3.1 Aprendizaje supervisado

En el aprendizaje supervisado se utiliza un conjunto de datos etiquetados para entrenar el algoritmo. Los datos etiquetados son aquellos que han sido "marcados" con las respuestas correctas bajo supervisión humana para evitar errores en el modelo. El proceso de etiquetado es caro y laborioso. Las etiquetas se denominan clases y representan los posibles valores que nos podemos encontrar. Por ejemplo, si queremos construir un sistema de reconocimiento de imágenes de animales, entrenaríamos el algoritmo con dichas imágenes previamente etiquetadas y le pediríamos que reconociera los animales en imágenes nuevas.

Sólo es posible aplicarlo si disponemos de datos etiquetados. Se utiliza para clasificar, predecir y, como no, tomar decisiones basadas en esos datos.

- Algoritmos de clasificación. Ante un dato nuevo, el modelo devolverá su etiqueta de clase correspondiente. Resuelven problemas de predicción de clases: si un usuario comprará o no, Test A/B de marketing, correo spam/no spam, o la evaluación de diabetes en pacientes a partir de unos determinados síntomas.

- Algoritmos de regresión. Ante un nuevo dato, el modelo devolverá un valor real (y) que pertenece a los números reales en lugar de una etiqueta de clase. Resuelven problemas de predicción numérica a partir de datos históricos, como determinar el valor de una acción en bolsa la semana que viene o la concentración de ozono troposférico a partir de otras variables meteorológicas.

1.3.2 Aprendizaje no supervisado

En el aprendizaje no supervisado el algoritmo descubre por si mismo patrones y relaciones en los datos (sólo hay entradas), a diferencia del aprendizaje supervisado donde le proporcionábamos datos etiquetados (entradas y salidas). Por ejemplo, supongamos que tenemos los datos de todos los pacientes de la sanidad española. Con los historiales de los pacientes, estos algoritmos podrían agrupar los pacientes que tuvieran un comportamiento médico similar y crear grupos. Luego nos correspondería a nosotros darle un nombre a ese grupo (diabetes, enfermedad crónica, estrés,...) por las características en común de los pacientes. Además, estos algoritmos pueden clasificar a pacientes nuevos automáticamente y proporcionar información de la evolución de su enfermedad. Todo esto sin proporcionar etiquetas, sólo los datos contenidos en los historiales de los pacientes.

Los algoritmos no supervisados serán explicados con todo el detalle en su capítulo:

- Algoritmos de clustering o de agrupamiento de datos.
- Algoritmos de reducción de la dimensionalidad (embeddings y análisis factorial).

1.3.3 Aprendizaje autosupervisado

El aprendizaje autosupervisado (*self-supervised learning* en inglés) es una rama del aprendizaje máquina muy activa en cuanto a proyectos de investigación por las implicaciones que podrían producirse: a un algoritmo de aprendizaje supervisado, le pediremos a partir sólo de los datos de entrada, obtener las salidas (y).

Recordemos que en el aprendizaje supervisado se proporcionaban entradas y salidas etiquetadas. Como el 95% de los datos que existen en la actualidad son no etiquetados (redes sociales, webscrapping, datos en tiempo real, etc.) y el proceso de etiquetado es caro y laborioso, el aprendizaje auto-supervisado es fundamental, ejemplo de ello es el modelo de predicción del lenguaje GPT-3 desarrollado por Open AI.

1.3.4 Aprendizaje reforzado

En este tipo de aprendizaje el modelo interactúa con el entorno y crea una serie de acciones determinadas (política) según cada entrada de datos para que se resuelva un problema. No se le dice al modelo la acción que debe realizar (en ese caso sería aprendizaje supervisado) sino que aprende porque se le recompensa o se le castiga según las acciones que realiza y las consecuencias (positivas/negativas) de dichas acciones.

El aprendizaje reforzado es especialmente adecuado para la resolución de problemas que se extienden a largo plazo, como juegos, robótica, gestión de recursos o logística.

1.3.5 Aprendizaje semisupervisado

Sigue un enfoque menos tradicional que el aprendizaje supervisado y no supervisado, mezclando una proporción (menor) de datos etiquetados y una (mayor) de datos no etiquetados en el conjunto de entrenamiento.

El modelo asume que los datos más próximos entre sí tienen la misma etiqueta y se ajustará con un modelo supervisado tomando como referencia todos los valores, incluidos los no-etiquetados, dando un mejor rendimiento que un modelo puro supervisado.

Son algoritmos especialmente indicados en casos con poca disponibilidad de datos etiquetados (por coste o por tiempo) como la detección de anomalías Al igual que el aprendizaje auto-supervisado, está en el punto de mira de muchos proyectos por su potencial impacto ya que el 95% de los datos no se etiquetan.

1.4 PASADO, PRESENTE Y FUTURO

Ésta es una breve historia a través de los hitos más reseñables, según nuestro criterio, de la ciencia de datos y el aprendizaje máquina:

- Mohammed Ibn Musa al-Jwarizmi (780-850), sabio persa autor de tratados sobre matemáticas, astronomía, astrología y geografía, cuya obra Aritmética, traducida al latín como *Algoritmi de número Indorum*, dio nacimiento a la palabra algoritmo.

- El método de mínimos cuadrados de Gauss (1809) y el teorema de Gauss-Markov, como base de las funciones de optimización (descenso por gradiente).

- El teorema de Bayes (1812), nacido para demostrar si la evidencia del mundo natural podía demostrar la existencia de Dios, definió la probabilidad de un determinado evento estudiando los eventos relacionados ocurridos en el pasado.

- El diseño del primer ordenador, o máquina analítica, y el primer programa informático de la historia desarrollados por Luigi Menabrea, Ada Lovelace (1845) y Charles Babbage, considerado padre de la computación.

- En 1880 Herman Hollerith, empleado del censo estadounidense y padre fundador de la compañía que después se conocería como IBM, consigue reducir un trabajo de 10 años a 3 meses con la primera máquina tabuladora.

- En 1926 Nikola Tesla predice la tecnología inalámbrica y que cada hombre llevará un teléfono en su propio bolsillo.

- En 1940 el aumento de bibliotecas a nivel popular, y por tanto del volumen de información disponible, empuja el interés por desarrollar mejores motores de búsqueda.

- Entre 1943 y 1945, el alemán Konrad Zuse creó el lenguaje de programación Plankalkül que no fue utilizado por su complejidad.

- Durante la 2ª guerra mundial, para mejorar los cálculos de tiro de artillería, se inició la construcción del primer ordenador. En 1946 se crea ENIAC (Electronic Numerical Integrator and Computer), una 'calculadora' que pesaba 27 toneladas y ocupaba, sólo, 167 m2. Se tomó la decisión de no apagarlo nunca para reducir sus continuos fallos, por fundidos de las válvulas de vacío, a la mitad a costa de provocar frecuentes apagones en la ciudad de Filadelfia. Su consumo era de 160.000 W.

- En 1950 Alan Turing publica un artículo clave para el desarrollo de la ciencia de datos, Maquinaria computacional e inteligencia, en el que presenta el famoso Test de Turing como respuesta a su pregunta ¿pueden las máquinas pensar? Turing es considerado el padre de la inteligencia artificial.

- En 1956 Fritz-Rudolf Güntsch, físico alemán, crea una forma de procesar la memoria finita como infinita, reduciendo las limitaciones de hardware e introduciendo el concepto de memoria virtual.

- En 1958 Frank Rosenblatt crea el perceptrón que es el origen de las redes neuronales *(tratado en el capítulo de Deep Learning).* Con un modelo de aprendizaje supervisado consiguió que una computadora aprendiera a distinguir las cartas marcadas a la izquierda de las marcadas a la derecha. En ese mismo año Hans Peter Luhn define el término *business intelligence*, la inteligencia de negocio como la habilidad de percibir las relaciones entre los hechos para ir hacia un objetivo deseado.

- En 1962 el estadounidense John W. Tukey, desarrollador de algoritmos complejos y del diagrama de caja y bigotes (Box Plot), cuestiona el futuro de la estadística como ciencia empírica y su evolución a ciencia de datos. El mismo año se presenta *Shoebox* (IBM), la primera máquina que comprende y procesa voz. Fueron los orígenes de los asistentes virtuales (Siri de Apple y Alexa de Google)

- En 1965 se formula la Ley de Moore, una observación empírica por la que Gordon E.Moore, cofundador de Intel, comprobó que cada 2 años aproximadamente se duplicaba el número de transistores por unidad de superficie en circuitos integrados. Fue la antesala de una época de gran progreso tecnológico por el descenso del coste de fabricación, y por tanto de precios, y el aumento continuo de prestaciones. Hay que tener en cuenta que se enunció antes de que existieran los microprocesadores,

inventados en 1971, los ordenadores personales, accesibles al gran público en los años ochenta y la telefonía móvil era sólo un experimento. Fue vigente hasta mediados de la década de 2010 pues el aumento de la densidad tiene un límite provocador por el aumento de temperatura que hay que eliminar sin dañar los transistores.

- En 1966 se empieza a usar comercialmente el código de barras, un método de líneas paralelas de distinto grosor y espaciado, patentado en 1952, para identificar automáticamente los vagones del ferrocarril estadounidense.

- En 1974 el danés Peter Naur utiliza ampliamente el concepto ciencia de datos en lugar de ciencias computacionales en sus publicaciones, promoviendo su uso intensivo en el mundo académico.

- En 1980 basándose en el bajo coste de almacenamiento de datos (por el que no se eliminan los obsoletos o no útiles), I.A. Tjomsland hizo una analogía a la primera ley de Parkinson sobre el crecimiento exponencial de los datos "los datos se expanden para llenar el espacio disponible" dejando entrever su valor potencial para la toma de decisiones en la empresa.

- En 1989 Erik Larson habla por primera vez de Big Data en un artículo donde habla del correo basura que recibe.

- El 23 de agosto de 1991 usuarios accesos al CERN pueden acceder a la red que conecta ordenadores. Nace la *World Wide Web*, internet tal y como lo conocemos hoy, como un sistema de red con interconexiones mundiales accesible para todos. Ese mismo año nace el lenguaje de programación Python. El neerlandés Guido van Rossum (super fan de los Monty Phyton) creó este lenguaje orientado a objetos y multiplataforma, de alto nivel y con una sintaxis simple, aumentó la productividad de los programadores.

- En 1994 una compañía japonesa subsidiaria de Toyota presenta la versión evolucionada del código de barras que permite la lectura del código a alta velocidad.

- En 1997 Jeff Wu dio una charla llamada ¿Estadística=Ciencia de datos? donde describió la triada formada por la recolección, el análisis y modelado de datos, y la toma de decisiones; solicito que la estadística fuese renombrada como ciencia de datos. Ese mismo año Google lanza su sistema de búsqueda en internet y Michael Cox y David Ellsworth, investigadores de la NASA, utilizan por primera vez el término *Big*

Data para exponer la incapacidad de los sistemas informáticos para el procesamiento de los datos por su continuo crecimiento. Es el renacer del aprendizaje máquina, marcado a nivel mundial por la derrota del campeón mundial de ajedrez, el ruso Garry Kasparov, frente al ordenador de IBM Deep Blue.

- En 1999 Kevin Ashton imaginó la posibilidad de que los datos fueran recopilados sin nuestra ayuda con su considerable reducción de costes. Era el nacimiento del IOT o internet de las cosas.

- En 2001 Doug Laney, de Gartner, en su artículo *3D Data Management* define las 3 V's del *Big Data*: volumen, velocidad y variedad; se populariza el concepto SaaS (software as a service).

- Google publica en 2003 sus sistemas de indexación y almacenamiento para su motor de búsqueda, GFS (Google File System) y MapReduce, pilares para el lanzamiento, en código abierto, de Hadoop en 2006 por Doug Cutting. Su procesamiento distribuido, virtualmente ilimitado, de los datos fue clave para abordar la cantidad masiva de información de la web.

- En 2004, la cadena de gran consumo americana Wal-Mart obliga a sus mayores proveedores a instalar etiquetas RFID en sus artículos. En 2022 lo extiende a todos sus proveedores. Es el mayor impulso a esta tecnología que apareció durante la 2ª guerra mundial y, mediante la inserción del chip, controla la localización del artículo en cualquier fase de la cadena de producción y suministro. Su uso se ha extendido a sectores tan variados como el deportivo y el sanitario.

- En 2005 nace la Web 2.0, pasando de un contenedor de información a la red social donde los usuarios interactúan y son creadores de contenido. Se vende MySpace por 580M$ y Facebook admite a estudiantes fuera de la universidad de Harvard. Nace el fenómeno de las redes sociales, RRSS.

- En 2006, el científico Geoffrey Hinton acuñó el término Deep Learning para explicar nuevas arquitecturas de Redes Neuronales profundas. Nace Twitter, la red social de los 140 caracteres.

- En 2007 más de 330.000 programadores estaban inscritos en AWS, Amazon Web Services, los servicios de computación en la nube ofrecidos por Amazon. Proporciona software como servicio (SaaS), plataforma como servicio (PaaS) e infraestructura como servicio (IaaS)

y es compatible con muchos lenguajes, herramientas y marcos de programación diferentes. En 2008 Microsoft anuncia su competidor, Azure, y Google el suyo, Google Cloud Platform.

- En 2009 aparecen las *GPUs* , *Graphic Processing Units,* que reducen el tiempo y coste de los modelos de Deep Learning.
- En 2010 la navegación por internet desde dispositivos móviles supera la de los ordenadores. Es el triunfo del i-Pad/Apple, de Android/Google y de las aplicaciones móviles sobre el cable.
- En 2014 la terminología *Big Data* sale del informe *Gartner Hype Cycle* sobre tecnologías emergentes y entra la geolocalización de los datos como tecnología de gran valor para las empresas.

Los principales hitos a partir de este año pertenecen al campo del aprendizaje profundo y se tratarán en dicho capítulo.

Y los próximos años seguro que nos deparan otros avances espectaculares, ¡comencemos el camino!