Capítulo 15
LA CIENCIA DE LOS DATOS

En el capítulo anterior hemos descrito los conceptos básicos del aprendizaje automático, y dentro de ello hemos nombrado a la ciencia de los datos. Dedicaremos este capítulo a profundizar un poco en esta disciplina por su importancia desde un punto de vista profesional y por su relación con la disciplina denominada *Big Data*[228].

La ciencia de datos es un campo interdisciplinario que ha evolucionado a lo largo del tiempo, incorporando elementos de estadísticas, informática, y análisis de negocios, entre otros. Aunque algunas de sus técnicas y métodos se remontan a siglos atrás en la historia de las matemáticas y las estadísticas, como campo con entidad propia empezó a tomar forma en las últimas décadas; y probablemente será una de las profesiones más impotentes en las siguientes.

15.1 Introducción a la ciencia de datos y aprendizaje automático

La ciencia de datos es una disciplina que se centra en extraer información y conocimiento significativo a partir de grandes cantidades de datos. Como ya hemos dicho, utiliza una combinación de métodos estadísticos, matemáticos y de programación para analizar y comprender los conjuntos de datos, y utiliza estos conocimientos para tomar decisiones informadas y resolver problemas complejos.

En la década de 1960, se empezó a utilizar el término "minería de datos" [*data mining*] para describir los métodos estadísticos e informáticos que eran capaces de extraer información oculta de grandes conjuntos de datos. La explosión de la disponibilidad de datos, impulsada en parte por el auge de la Internet y la reducción en los costes de almacenamiento y procesamiento de datos, ha llevado al desarrollo y popularización de herramientas y técnicas más sofisticadas para el análisis de datos.

En 2001, William S. Cleveland introdujo el término *ciencia de datos* como un campo interdiscipli-

228 *El término Big Data se utiliza comúnmente en muchos idiomas y contextos sin traducción, ya que se ha convertido en un término técnico universal. Sin embargo, si se desea traducir, una forma adecuada en castellano podría ser "Datos Masivos" o "Grandes Volúmenes de Datos". Estas traducciones capturan la esencia de lo que significa Big Data: la gestión y análisis de conjuntos de datos tan grandes y complejos que requieren sistemas y métodos especializados para su manejo y análisis.*

nario que engloba el análisis estadístico de datos, pero también se preocupa por la visualización, la ingeniería, y la toma de decisiones basada en datos. La publicación del artículo "*Data Scientist: The Sexiest Job of the 21st Century*"[229] por D.J. Patil y Thomas H. Davenport en la revista *Harvard Business Review* en 2012 marcó un punto de inflexión, ya que popularizó tanto el término **ciencia de datos** como la función del **científico de datos**.

Si queremos encontrar información oculta (patrones) en los datos, necesitamos herramientas que aprendan automáticamente de los mismos y generalicen ¿Y qué herramientas conocemos que nos permita aprender y generalizar a partir de los datos? El aprendizaje automático.

Ejemplo: Aplicación de la ciencia de datos en la medicina

Imaginemos un conjunto de datos médicos que contiene información sobre pacientes, como edad, género, historial médico, resultados de pruebas y diagnósticos. Un científico de datos podría utilizar este conjunto de datos para identificar patrones que puedan ayudar a predecir enfermedades o evaluar la eficacia de ciertos tratamientos.

Por ejemplo, con estos datos de entrada y utilizando técnicas de aprendizaje automático, se podría desarrollar un modelo que tras analizar los datos de pacientes con diabetes, prediga la probabilidad de que uno en particular desarrolle complicaciones graves en el futuro; simplemente porque los patrones existentes en los datos históricos así lo indican.

El modelo podría tener en cuenta variables como el nivel de glucosa en sangre, el índice de masa corporal (IMC), la presión arterial y el historial familiar. Esto ayudaría a los médicos a tomar medidas preventivas y brindar un **tratamiento personalizado**.

Ejemplo: Aplicación de la ciencia de datos en el comercio electrónico

Imaginemos una empresa de comercio electrónico que recopila datos de transacciones de sus clientes, como información de productos comprados, tiempo de compra, ubicación geográfica y preferencias de compra. Utilizando técnicas de ciencia de datos, la empresa puede analizar estos datos con el objetivo de comprender mejor el comportamiento de sus clientes y mejorar su estrategia de marketing y ventas.

Siguiendo con el ejemplo, podrían desarrollar un modelo de recomendación personalizado utilizando algoritmos de aprendizaje automático. Este modelo analizaría el historial de compras de un cliente y los patrones de compra de otros clientes similares para hacer recomendaciones de productos relevantes. Así, la empresa podría ofrecer una experiencia de compra más personalizada y aumentar la satisfacción del cliente.

15.2 Big Data

Además de las herramientas relacionadas con el aprendizaje automático, la ciencia de datos necesi-

229 *Artículo original: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*

ta poder manejar grandes conjuntos de información, probablemente enormes, para conseguir algo de provecho.

> *Con **Big Data** nos referimos a las técnicas y herramientas para el procesamiento de grandes volúmenes de datos, tanto estructurados como no estructurados, que se genera en el mundo cada día y aquellos que hemos almacenado con el tiempo.*

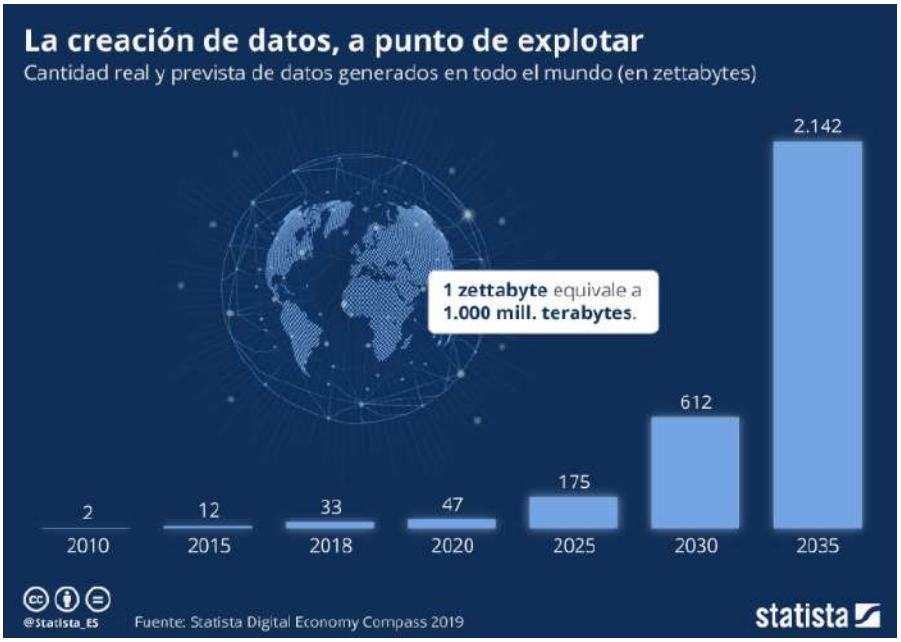

Figura 67: Estimación de la cantidad real y prevista de datos generados en todo el mundo (en zettabytes).

Fuente: indicado en la imagen.

Pero no es la cantidad de datos lo que es importante; es lo que las organizaciones hacen con los datos lo que importa. *Big Data* es una herramienta usada para crear ideas que conduzcan a mejores decisiones y estrategias empresariales y gubernamentales. El concepto de *Big Data* no es absoluto y está en constante evolución, pero inicialmente se caracteriza por las tres V:

- **Volumen**: Se refiere a la cantidad de datos. *Big Data* normalmente implica grandes volúmenes de datos. Estos pueden ser cantidades desconocidas o datos que son demasiado grandes para ser manejados por sistemas de bases de datos tradicionales.

- **Velocidad**: Se refiere a la velocidad con la que se generan nuevos datos y la velocidad a la que estos se mueven. Las redes sociales pueden recibir cientos de miles de actualizaciones por minuto, lo que requiere un manejo en tiempo real.

- **Variedad**: Se refiere al tipo y naturaleza de los datos. En el pasado, manejábamos principalmente datos estructurados que se ajustaban perfectamente en tablas de bases de datos relacionales. Pero hoy en día, el 80% de los datos del mundo son no estructurados. Esta-

mos manejando diferentes tipos de datos, como texto, imágenes, sonidos, vídeos, registros y transacciones financieras, que están siendo generados y consumidos de forma continua.

El tratamiento eficiente de los datos requiere sistemas de almacenamiento y procesamiento que son significativamente más avanzados que las bases de datos tradicionales. También ha dado lugar al desarrollo de una nueva clase de tecnologías y arquitecturas de datos, algunas las nombraremos en breve.

Aunque el término empezó a popularizarse en la primera década de este siglo, podemos remontarnos a la creación de la ciencia de la computación (y la subsiguiente aparición de la ingeniería informática) en la segunda mitad del siglo pasado:

Años 60-70: Aunque aún no se usaba el término *Big Data*, ya se estaban recopilando grandes conjuntos de datos en sectores como la astronomía y la geología. Muchos de esos datos históricos son empleados hoy en día, aunque en su momento no se sabía el poder de procesamiento que tenemos actualmente.

Años 90: Con el auge de Internet, y la aparición de la *Word Wide Web*, las empresas comenzaron a acumular grandes cantidades de datos. Surgieron los primeros motores de búsqueda, y las empresas comenzaron a reconocer el valor de los datos para entender el comportamiento del cliente.

2001: Doug Laney, un analista de Meta Group, ahora Gartner, creó la definición de *Big Data* a partir de las tres V: Volumen, Velocidad y Variedad, que desde entonces han sido utilizados para describir sus características fundamentales.

2004: Google publicó un artículo sobre el sistema de archivos distribuido GFS, que permitía el procesamiento de grandes cantidades de datos. Este artículo fue la base para el desarrollo de *Hadoop* por parte de Yahoo y actualmente en la Apache Software Foundation.

2005: Lanzamiento de *Hadoop*. Este *framework* de código abierto permitió el almacenamiento y el procesamiento de conjuntos de datos muy grandes, y se convirtió en una tecnología clave para el *Big Data*.

2006-2009: Con la popularización de las redes sociales y el auge de dispositivos móviles, la generación de datos se aceleró de forma exponencial, lo que llevó a un mayor enfoque en el almacenamiento y el análisis del *Big Data*.

2010: Se popularizan las bases de datos NoSQL, diseñadas para manejar tipos de datos no estructurados, lo que facilita aún más el tratamiento *Big Data*.

2012: Como ya hemos dicho, la *Harvard Business Review* publicó el artículo "*Data Scientist: The Sexiest Job of the 21st Century*", lo que impulsó aún más el interés por *Big Data* y la ciencia de datos.

2015 en adelante: El aprendizaje automático y la inteligencia artificial empiezan a jugar un papel mucho más importante en el análisis y la interpretación de *Big Data*. Los conceptos de "cuatro V" y

"cinco V" (agregando Veracidad y Valor respectivamente) comienzan a ser discutidos más ampliamente.

En la actualidad, *Big Data* es una parte integral de diversas industrias como la atención médica, la logística, el marketing y más. Los avances en la computación en la nube y en tecnologías como el Internet de las Cosas (IoT) continúan ampliando sus fronteras.

15.3 El ciclo de vida de la ciencia de los datos

El ciclo de vida de la ciencia de datos se compone esencialmente de la recopilación de datos y su selección, la limpieza de datos, el análisis exploratorio de datos, la construcción de modelos y el despliegue de los modelos creados.

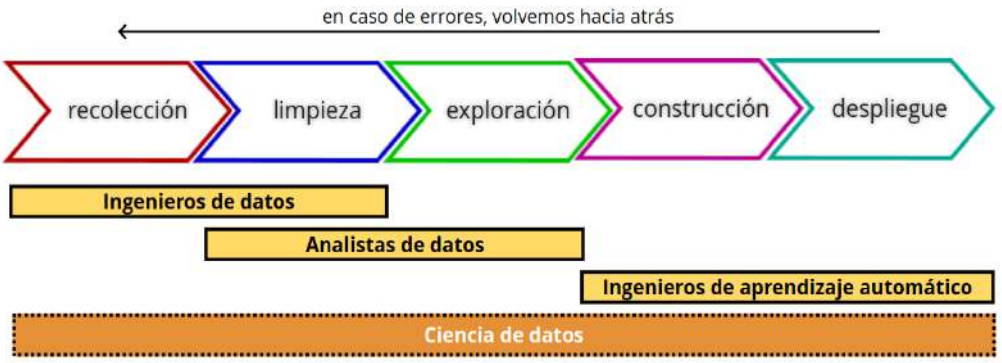

Figura 68: Fases y roles implicados en la ciencia de los datos.

Recolección y selección de datos: Los datos pueden provenir de diversas fuentes, como bases de datos, archivos CSV, archivos de texto, APIs, sensores, redes sociales, entre otros muchos. Por ejemplo, para analizar el comportamiento de los usuarios en un sitio web, se pueden recopilar datos de registro de servidores web y datos de seguimiento de clics.

Pero también los datos pueden tener diferentes formatos: Estos pueden poseer un formato estructurado, por ejemplo organizados en tablas con filas y columnas, como una hoja de cálculo (ver la Figura 61). Pero, por otro lado, tenemos datos no estructurados, que no siguen un formato específico, como por ejemplo correos electrónicos, imágenes, vídeos o texto libre. Por ejemplo, los *tweets* de Twitter (X) son datos no estructurados.

Limpieza y transformación de datos: En esta etapa, se eliminan datos duplicados, se resuelven valores faltantes o inconsistentes, y se transforman los datos al formato adecuado. Por ejemplo, si se tienen datos de temperatura en grados Fahrenheit y se requiere en grados Celsius, se realiza una transformación para convertirlos.

Hay también otras transformaciones más complejas, por ejemplo convertir el texto a su correspondiente vector de características (*embeddings*), tratar con datos categóricos, no numéricos, etc.

Exploración y visualización de datos: El Análisis descriptivo consiste en calcular estadísticas básicas, como promedio, mediana, desviación estándar, y generar visualizaciones para entender la distribución de los datos. Por ejemplo, se puede calcular la edad promedio de los estudiantes en un conjunto de datos y visualizarlo en un histograma. Se trata de comprender los datos. Para ello también se lleva a cabo una identificación de patrones y relaciones, mediante técnicas de visualización, como gráficos de dispersión y mapas de calor, se pueden identificar patrones y relaciones entre variables. Por ejemplo, se puede visualizar la relación entre el salario y la experiencia laboral en un conjunto de datos para determinar si existe una correlación.

También es importante el análisis de *outliers*, que son valores atípicos que pueden distorsionar el análisis. Por ejemplo, si se analizan datos de ventas y se encuentra una venta extremadamente alta, puede ser necesario investigar si fue un error o un evento extraordinario.

Finalmente se **construye y despliega el modelo**. Para ello se comienza por la selección de características, el cual es un proceso crucial en la construcción de modelos de aprendizaje automático. Consiste en identificar las variables o características más relevantes y significativas para el problema en cuestión. Algunas técnicas comunes utilizadas en la selección de características incluyen la correlación, la eliminación de características redundantes y el análisis de importancia de características.

Una vez que se han seleccionado las características, es necesario entrenar el modelo de aprendizaje automático. Esto implica proporcionar algoritmos de aprendizaje automático con un conjunto de datos de entrenamiento, que consta de ejemplos etiquetados, para que pueda aprender patrones y relaciones entre las características y las etiquetas.

Una vez que el modelo ha sido entrenado, es esencial evaluar su rendimiento para comprender qué tan bien ha aprendido (Figura 69) y su capacidad de generalizar ante la presencia de nuevos datos. Esto se logra utilizando métricas de rendimiento que miden la precisión, la exhaustividad, la exactitud o cualquier otro criterio relevante para el problema en cuestión. Además, como hemos tratado antes, es común dividir el conjunto de datos en conjuntos de entrenamiento y prueba, para evaluar el rendimiento del modelo en datos no vistos durante el entrenamiento.

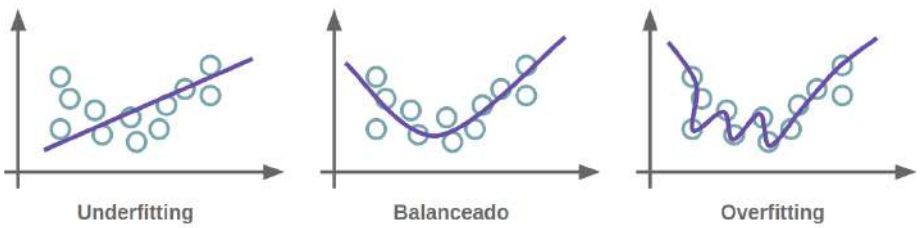

Figura 69: Tipos de aprendizajes según e resultado: subentrenamiento, balanceado y sobreentrenamiento.

Fuente: Wikimedia Commons

El modelo, después de una evaluación sistemática y rigurosa, se implementa en la estructura y canal preferidos de la organización. Este es el último paso en el ciclo de vida de la ciencia de datos. Cada paso de este ciclo debe realizarse con cuidado. Si algún paso se realiza incorrectamente y, por lo tanto afecta el siguiente paso, todo el esfuerzo realizado se desperdicia y debemos volver a una etapa anterior; incluso volver a empezar. Desde la necesidad inicial de la organización, hasta la implementación del modelo, a cada paso, se debe prestar la atención, el tiempo y el esfuerzo adecuados.

15.4 Aplicaciones de la ciencia de datos

La ciencia de datos, y sus técnicas y tecnologías asociadas, tienen una amplia gama de aplicaciones en diversos campos, lo que ha llevado a importantes avances en medicina, finanzas, marketing, transporte, entretenimiento y muchos otros sectores. Veamos algunos ejemplos de aplicaciones concretas para ilustrar cómo estas disciplinas están transformando diferentes áreas de la sociedad.

15.4.1 Medicina

Diagnóstico médico: Los modelos de aprendizaje automático pueden analizar datos clínicos y de pacientes para ayudar en la detección temprana y el diagnóstico preciso de enfermedades. Por ejemplo, se han desarrollado algoritmos capaces de detectar enfermedades cardíacas, cáncer de piel y retinopatía diabética mediante el análisis de imágenes médicas.

Medicina personalizada: La ciencia de datos puede ayudar a crear terapias personalizadas para pacientes. Los modelos pueden analizar el historial médico, los factores genéticos y otros datos para identificar tratamientos óptimos y predecir la eficacia de ciertos medicamentos.

15.4.2 Finanzas

Detección de fraudes: Los algoritmos de aprendizaje automático pueden analizar grandes volúmenes de datos financieros para identificar patrones y anomalías que puedan indicar actividades fraudulentas. Estos modelos pueden ayudar a las instituciones financieras a prevenir el fraude en tarjetas de crédito, seguros y transacciones bancarias.

Predicción del mercado: La ciencia de datos se utiliza para analizar datos históricos y en tiempo real del mercado financiero, lo que permite predecir las tendencias y realizar inversiones más informadas. Los modelos de aprendizaje automático pueden ayudar a los inversores a tomar decisiones basadas en patrones y señales del mercado.

15.4.3 Marketing

Segmentación de clientes: Los algoritmos de aprendizaje automático pueden analizar los datos de los clientes, como sus preferencias, comportamiento de compra y datos demográficos, para segmentarlos en grupos con características similares. Esto permite a las empresas personalizar sus es-

trategias de marketing y ofrecer productos y servicios específicos a cada segmento.

Recomendación de productos: Los sistemas de recomendación utilizan algoritmos de aprendizaje automático para analizar el historial de compras, las preferencias y el comportamiento de los clientes con el fin de ofrecer recomendaciones personalizadas de productos o servicios. Ejemplos populares de esto son las recomendaciones de películas o música en plataformas de streaming como Netflix y Spotify.

15.4.4 Transporte

Conducción autónoma: La ciencia de datos y el aprendizaje automático son fundamentales para desarrollar vehículos autónomos. Los algoritmos de visión por computadora y el procesamiento de sensores permiten que los automóviles interpretan su entorno y tomen decisiones en tiempo real, lo que mejora la seguridad y la eficiencia en el transporte.

Optimización de rutas: Los modelos de aprendizaje automático pueden analizar datos de tráfico, patrones de movilidad y otras variables para optimizar las rutas de transporte y reducir la congestión. Esto puede mejorar la eficiencia del transporte público y reducir los tiempos de viaje.

15.4.5 Entretenimiento

Recomendación de contenido: Las plataformas de entretenimiento utilizan algoritmos de aprendizaje automático para recomendar contenido a los usuarios. Por ejemplo, servicios de streaming como Netflix utilizan modelos que analizan el historial de visualización, las calificaciones y las preferencias de los usuarios para ofrecer recomendaciones personalizadas de películas y programas de televisión.

Generación de música y arte: Los modelos de aprendizaje automático pueden ser entrenados en grandes conjuntos de datos musicales o artísticos para generar composiciones musicales originales o incluso obras de arte. Estas aplicaciones creativas muestran cómo la ciencia de datos puede expandirse más allá de los ámbitos tradicionales.

15.5 Tendencias y futuro de la ciencia de datos y aprendizaje automático

En el campo de la ciencia de datos y el aprendizaje automático, existen varias tendencias y avances que están dando forma al futuro de estas disciplinas. Vamos a destacar las dos más importante a medio plazo que inciden en la ciencia de los datos:

Aprendizaje automático en tiempo real: A medida que los avances en capacidad de cómputo y algoritmos permiten un procesamiento más eficaz y eficiente, se están desarrollando aplicaciones de aprendizaje automático en tiempo real. Esto implica tomar decisiones y realizar predicciones mucho más rápido y cuando se necesiten, lo cual es fundamental en áreas como el comercio financiero, la detección de fraudes y la detección de intrusiones en sistemas de seguridad. Por ejemplo, las em-

presas de comercio electrónico utilizan algoritmos de aprendizaje automático en tiempo real para personalizar las recomendaciones de productos en función del comportamiento de compra del usuario en tiempo real.

Automatización de la ciencia de datos: Se trata del desarrollo de herramientas y algoritmos que pueden realizar tareas de análisis de datos de manera autónoma, sin necesidad de una intervención humana constante. Esto permite a los científicos de datos centrarse en tareas más complejas y estratégicas. Un ejemplo de esto es la automatización de la selección y ajuste de modelos, donde los algoritmos pueden explorar diferentes opciones de modelos y seleccionar automáticamente el mejor modelo para un conjunto de datos dado.

RETOS DEL CAPITULO 16

1. Dialoga con tu IA favorita: Investiga y describe en tus propias palabras qué es la ciencia de datos y cómo se relaciona con el aprendizaje automático.
2. Pregúntale a tu agente conversacional favorito, qué son los términos o conceptos:
   - Big Data.
   - Análisis de datos.
   - Minería de datos.
   - Análisis exploratorio de datos (EDA).
   - Visualización de datos.
   - Feature Engineering.
   - Hadoop.
   - SQL y no-SQL
   - Gestión de datos.
   - Data Lake.

   No te quedes con una simple definición; busca entender y relacionar todos ellos.
3. Una técnica muy habitual es el *web-scraping*. Busca información de qué es y pregunta a tu IA favorita que te describa cómo hacerlo. ¿Te atreves a pedirle que genere un código para que lo haga en una página de un periódico?
4. ¿Qués es una fuente de datos no estructurados? Dialoga con tu IA favorita para que te muestre cómo se tratan estos datos en el ámbito del aprendizaje automático.
5. Busca en la web ejemplos de empresas que usen la Ciencia de los Datos en su día a día.
6. ¿Qué es un/una ingeniero/a de datos? Indaga que objetivos tiene.
7. ¿Qué es un/una analista de datos? Enumera que objetivos tiene.
8. ¿Qué es un/una ingeniero/a de aprendizaje automático? Aprende que objetivos tiene.
9. En la *web* hay cantidad de cursos *online*. Busca los que se correspondan con la Ciencia de los Datos.
10. Busca e indaga que carreras universitarias o ciclos formativos debes estudiar para poder trabajar en el ámbito de la Ciencia de los Datos.