Prólogo

PID_00284577

Julià Minguillón Alfonso

Tiempo mínimo de dedicación recomendado: 1 hora

Universitat Oberta de Catalunya

Julià Minguillón Alfonso

Licenciado en Ingeniería Informática por la Universidad Autónoma de Barcelona (UAB) en 1995, máster de Combinatoria y Comunicación Digital por la UAB en 1997 y doctor ingeniero en Informática por la UAB en 2002. Desde 2001 ejerce como profesor en los Estudios de Informática, Multimedia y Telecomunicación de la Universitat Oberta de Catalunya (UOC). Pertenece al grupo de investigación Learning Analytics for Innovation and Knowledge Application (LAIKA), donde desarrolla proyectos de investigación relacionados con el análisis y la visualización del comportamiento de los usuarios de entornos virtuales de aprendizaje y redes sociales.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Julià Minguillón Alfonso

Primera edición: septiembre 2021

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Julià Minguillón Alfonso
Producción: FUOC

Índice

1. Prólogo

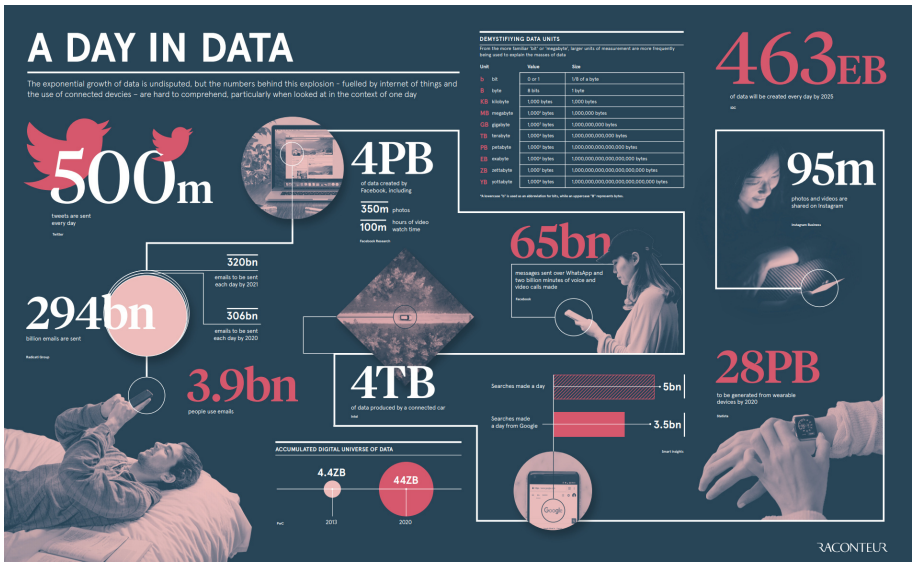
En los últimos años ha habido un incremento sin precedentes en la informatización de las organizaciones y en sus capacidades de transmisión de datos en todos los ámbitos de actividad. Hace diez años, como muestra del incremento en la capacidad de cómputo, teníamos ejemplos como este:

> Los datos generados por una sola de las últimas misiones de la NASA equivalen en número de bytes a cien veces las de todas las misiones de la NASA hasta el momento; y estos datos tienen que analizarse. En concreto, el sistema de satélites Earth Orbiting System (EOS) de la NASA puede generar hasta cincuenta gigabytes de datos cada día.

En la actualidad, estas cifras son risibles. La digitalización ha llegado a cualquier aspecto de nuestra vida cotidiana, tal y como muestra la figura 1, y conceptos como *datos masivos* o *inteligencia artificial* aparecen en todos los ámbitos y sectores de negocio. En el año 2020, se estima que en total se produjeron 59 zettabytes (ZB) de datos.[1] Un zettabyte son 10^{21} bytes, por lo que el total equivale a 59 mil millones de discos duros de un terabyte cada uno. El uso masivo de dispositivos móviles y los datos que se generan en su utilización diaria son, en parte, causantes de este incremento, ya que potencialmente cada usuario (que puede usar uno o más dispositivos) accede a una infinidad de servicios digitales y deja un rastro que puede ser analizado.

[1] Fuente: https://bit.ly/3xaYZJk

Figura 1. Datos generados diariamente



Fuente: https://www.visualcapitalist.com/wp-content/uploads/2019/04/data-generated-each-day-full.html

De hecho, se acepta que la capacidad de las organizaciones de gestionar y crear conocimiento es una ventaja competitiva, basándose en la información generada dentro de la organización y en la captada de su entorno. Cada vez se dispone de más y más datos sobre el propio funcionamiento de cada empresa u organización, sobre las transacciones con sus clientes y proveedores y sobre las estadísticas de otros fenómenos de interés (por ejemplo, las bases de datos metereológicas sobre condiciones climáticas, sequías, etc., obtenidas por satélite, pueden ser interesantes para quien negocia con productos agrícolas). La idea detrás del concepto *datos masivos* no es solamente disponer de estos (en cantidad), sino también la posibilidad de cruzar conjuntos diferentes que enriquecen los datos originales.

Por ejemplo, las grandes empresas de distribución comercial, en particular las compañías que operan sobre grandes superficies comerciales, están interesadas en conocer y anticipar la evolución de las ventas y pautas de comportamiento de sus clientes, sus diferencias regionales, etc., con el fin de poder anticipar las necesidades, programar ofertas y planificar las compras. Utilizan una gran cantidad de datos procedentes de los movimientos de materiales, así como del análisis de cada una de las compras de cada uno de los clientes (y analizan los datos recogidos en las cajas o terminales de punto de venta). Amazon se ha convertido en el paradigma de empresa digital «rica» en datos. Por ejemplo, en septiembre de 2019 casi 151 millones de usuarios accedieron a la aplicación móvil para realizar sus compras.[2] Con semejante cantidad de datos no resulta difícil pensar qué información puede extraer Amazon de sus clientes para mejorar sus servicios o, simplemente, ofrecerles más productos.

(2) Fuente: https://bit.ly/3wdSBRC

Procesos parecidos de recogida de datos sirven para llevar a cabo la gestión financiera o tomar decisiones con respecto a la concesión de privilegios para los clientes, como tarjetas de crédito o seguros de vida, por ejemplo. Ámbitos como la banca, los seguros, las tarjetas de crédito como la Visa, etc. usan los datos que recogen de sus clientes para detectar, por ejemplo, usos fraudulentos. Esta manera de reconocer el valor que pueden tener los datos se extiende también a otros ámbitos. Por ejemplo, los grandes sistemas hospitalarios recogen y analizan datos con el fin de minimizar los costes debidos a la prolongación innecesaria de la estancia de los pacientes, o bien localizan problemas en el tratamiento o relaciones entre complicaciones posoperatorias y un determinado tipo de intervenciones.

Existe, pues, una extendida lógica de apreciación del valor que los datos pueden tener. Consiguientemente, se han generado nuevas necesidades en las empresas, que se reflejan en la creación de nuevos perfiles profesionales correspondientes a las tareas de lo que se conoce como *minería de datos*. Y esta tendencia continúa y se amplía. La racionalidad que actúa detrás de este proceso también puede aplicarse perfectamente a ámbitos más modestos. De hecho, las pequeñas empresas que adoptan esta manera de considerar los conjuntos de datos, quizá no tan grandes como los que hemos apuntado anteriormen-

te, adquieren rápidamente una nueva forma de ver y organizar sus propios procesos de trabajo para mejorarlos.

Los datos son el reflejo de la actividad de cualquier usuario, proceso, empresa, etc. y permiten extraer un conocimiento sobre esta actividad que puede ser aprovechado para entenderla mejor y optimizarla. Son, de hecho, un activo más de cualquier proceso que pueda usarlos en su propio beneficio. Pero este beneficio está «enterrado» en los datos, es por este motivo que este descubrimiento se asimila a la tarea de la minería, que tiene que retirar toneladas de escombros (los datos) para encontrar el valor buscado (el conocimiento). Con una metáfora similar, Clive Humby decía en 2006 «data is the new oil» (los datos son el nuevo petróleo), resumiendo en una sola frase su valor económico y la necesidad de buscarlos, encontrarlos y extraerlos, como en todo proceso de minería.

Figura 2. «Data is the new oil», según Clive Humby, 2006.

Actualmente, la paradoja reside en el hecho de que, aunque hay más capacidad para generar y tratar de forma automática y rutinaria los datos, es decir, de estar en condiciones óptimas para obtener buenos análisis y extraer conocimiento, el volumen creciente de los datos dificulta su tratamiento rápido y eficiente para la toma de decisiones. Por este motivo, es muy importante el desarrollo de nuevas herramientas que aceleren el proceso de análisis y que lo adapten tan dinámicamente como sea posible a las necesidades cambiantes de quienes tienen que tomar decisiones: un reto fenomenal para todos los que trabajan en estadística, algorítmica, inteligencia artificial, aprendizaje automático, bases de datos y sistemas de información. En este sentido, desde hace unos años se habla de *automated machine learning* (AutoML), con la esperanza de poder automatizar algunos de los procesos inherentes a la minería de datos, aunque siempre bajo la supervisión de un experto. Aunque ciertos procesos de preprocesado y análisis de los datos son siempre los mismos, captar la naturaleza de los datos y comprender el problema que se debe resolver es un elemento clave de la minería de datos. Las herramientas (modelos, algorit-

mos, etc.) que se deberán utilizar serán diferentes en cada caso, dependiendo del tipo de datos y el objetivo a alcanzar.

Para alcanzar un buen nivel de conocimientos en esta nueva actividad llamada *minería de datos*, se debe ser capaz de combinar conceptos y métodos procedentes de diferentes disciplinas. Algunos ejemplos pueden ser los siguientes:

- **La programación.** Como proceso que es, la minería de datos necesita resolver diferentes problemas de forma más o menos automatizada, desde la obtención y almacenamiento de los datos, su preprocesado y análisis, hasta la publicación de los conjuntos de datos y la evaluación de los modelos construidos. Algunas de estas tareas pueden llevarse a cabo mediante el uso de pequeños programas que las automaticen, como por ejemplo, encontrar el mejor ajuste de los parámetros de un modelo.

- **Las bases de datos.** Se dedican a mejorar la captación, la organización y la consulta de datos, pero no se había propuesto de buen principio la disposición correcta para la posterior extracción de conocimiento de ellos. Es necesario saber acceder a bases de datos mediante las consultas adecuadas para extraer aquellos datos sobre los cuales quiere realizarse el análisis.

- **La estadística.** Tiene una probada tradición en la creación de modelos a partir de datos, pero no se había planteado la problemática de extraerlos a partir de bases de datos con una organización complicada, o de interactuar de forma continua con el usuario o de crear modelos más simbólicos o algorítmicos que numéricos. La estadística se fundamenta en las propiedades de los datos, las distribuciones subyacentes, los tests que permiten validar o refutar hipótesis, o la creación de modelos de regresión, aspectos que se integran en muchos de los modelos de minería de datos.

- **El aprendizaje automático.** Es una subdisciplina de la inteligencia artificial que se había encargado de extraer conocimiento de observaciones y ejemplos, pero que había simplificado las fuentes de datos, obviando las oportunidades de aprendizaje implícitas en las complejidades propias de las bases de datos reales. La minería de datos incluye el aprendizaje automático como una herramienta más (igual que la batería de herramientas estadísticas).

- **La visualización de datos.** El sistema visual humano es muy eficiente detectando tendencias, patrones, *outliers*, etc., lo cual se puede utilizar para dirigir un análisis preliminar en una dirección u otra. En la actualidad se plantea a menudo hacer una primera inspección visual de los datos para entender mejor su naturaleza. La visualización de datos también es una herramienta eficaz para interpretar mejor los modelos construidos y, especialmente, medir la incertidumbre de las predicciones o las clasificaciones realizadas.

En conjunto, la minería de datos es una actividad que permite aprender muchas cosas de diferentes campos e integrarlas bajo una misma perspectiva al servicio de los objetivos prácticos concretos. Por lo tanto, en esta asignatura se persigue dar a conocer las diferentes aplicaciones de la minería de datos, los procesos que la componen, los métodos que pueden utilizarse en cada caso y, sobre todo, desarrollar una actitud, que podríamos llamar *de disposición alerta o atenta*, o incluso *exploratoria*, para detectar cuándo es necesario emprender un proyecto de minería de datos y cómo organizarlo.

Finalmente, es importante recordar un hecho que es clave para entender mejor lo que queremos obtener mediante un proceso de minería de datos. Se trata del concepto *garbage in, garbage out* (GIGO), que expresa que, si los datos son basura, los modelos construidos para, por ejemplo, elaborar predicciones, también serán basura. Con toda la batería de herramientas disponibles, es relativamente fácil extraer conocimiento de cualquier conjunto de datos, pero hay que tener en cuenta que este puede ser irrelevante o incluso erróneo, si los datos de entrada contienen mucho ruido o son, simplemente, también erróneos. Es por esta razón que conceptos como el mencionado AutoML no pueden usarse sin supervisión ni sin un conocimiento de los datos y de su contexto.

1.1. Contenidos del material docente

Este material docente está organizado en este prólogo de presentación y siete módulos, seis de carácter más teórico donde se introducen los conceptos que componen el temario y uno final de carácter práctico en forma de estudio de caso, que sirve como hilo argumental para ejemplificar los conceptos del resto de módulos.

Así, el primer módulo describe el proceso de minería de datos, relacionándolo con el ciclo de vida de los datos y desarrollando esta idea de proceso más allá del simple uso de modelos y técnicas. Este módulo da una visión general e introduce conceptos que se desarrollarán en el resto de módulos, incluyendo referencias bibliográficas a los trabajos originales donde se definieron por primera vez conceptos relacionados con el proceso de minería de datos y las tecnologías usadas para ello, así como trabajos más recientes que muestran la evolución del ámbito. Una simple búsqueda del término *data mining* en la base de datos Scopus devuelve un total de 184.447 publicaciones (hasta mayo de 2021), de las cuales 560 son libros, siendo solamente 2.328 publicaciones anteriores al año 2000, lo que muestra el creciente interés en el ámbito.

Después hay dos módulos dedicados a la fase de preparación de los datos, tanto por lo que respecta a su preprocesado como a la extracción de caracte-

rísticas que pueden ser relevantes para resolver el problema planteado por el conjunto de datos original. A continuación, se presentan diferentes modelos típicos usados como herramientas básicas de la minería de datos, organizados en métodos no supervisados y métodos supervisados, para la detección de grupos similares y la clasificación y predicción, respectivamente. El último módulo teórico introduce cómo evaluar los modelos construidos, de acuerdo a su naturaleza y el objetivo a alcanzar, de forma que sea posible comparar y escoger entre diferentes modelos.

Por su parte, el caso de estudio describe paso a paso un proceso de minería de datos realizado sobre un conjunto de datos ligado a una pregunta que quiere responderse, siguiendo el esquema descrito anteriormente, desde su preprocesado hasta su validación e interpretación. Aunque aparece al final, este módulo puede consultarse juntamente con el resto de módulos más teóricos para entender mejor los conceptos introducidos en cada uno de ellos. Es crucial darse cuenta de que la minería de datos es más una metodología y un proceso continuo que simplemente una técnica o un conjunto de técnicas. Por lo tanto, es muy importante reforzar la mencionada actitud exploratoria exprimiendo el caso práctico que actúa como hilo conductor a lo largo de los diferentes módulos y los otros ejemplos que se presentan en la asignatura, de tal forma que pueda interiorizarse este proceso de forma adecuada.

Finalmente, aunque algunos de los conceptos básicos propios del ámbito de la minería de datos se establecieron hace ya unas cuantas décadas, no deja de ser un ámbito de rápido desarrollo debido al incremento continuo de capacidad de cálculo y de almacenamiento, por lo que es necesario tener una visión global que permita detectar estos avances. Para ello, recomendamos consultar el espacio de recursos de ciencia de datos, el cual va incorporando nuevos elementos, en forma de catálogo en línea. Dicho espacio puede consultarse en: http://datascience.recursos.uoc.edu/es/.