Introducción a los datos

PID_00272095

Blas Torregrosa García

Tiempo mínimo de dedicación recomendado: 1 hora

Universitat Oberta de Catalunya

Blas Torregrosa García

Ingeniero en Informática y máster universitario en Seguridad de las Tecnologías de la Información y de las Comunicaciones (MISTIC) por la Universitat Oberta de Catalunya (UOC). Especializado en ciberseguridad. Profesor colaborador en el máster de Ciencia de Datos de la UOC y profesor asociado en la Universidad de Valladolid (UVA).

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Ferran Prados Carrasco (2020)

Primera edición: febrero 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción..... 5

1. Concepto de datos..... 7
1.1. ¿Qué es un dato? 7
1.2. Datos, información y conocimiento 8
1.3. Tipos de datos 9
1.3.1. Datos cuantitativos 9
1.3.2. Datos cualitativos 10
1.4. Estructuras de datos 11

2. Ciclo de vida de los datos..... 12
2.1. Generación 12
2.2. Captura 12
2.3. Almacenamiento 14
2.4. Preprocesado 14
2.5. Análisis 15
2.6. Visualización 15
2.7. Interpretación 16

Bibliografía..... 17

Introducción

El objetivo de este módulo es definir qué son los datos y la relación existente entre datos, información y conocimiento. También se introducen los diferentes tipos de datos.

En este módulo se utilizará el ciclo de vida de los datos para describir las diferentes fases por las que pueden pasar estos antes de convertirse en información y luego en conocimiento.

1. Concepto de datos

En primer lugar vamos a hablar un poco de lo que significa el concepto **dato**.

> «Del latín *datum* ('lo que se da').
>
> Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho.
>
> Documento, testimonio, fundamento.
>
> Información dispuesta de manera adecuada para su tratamiento por una computadora.»
>
> Diccionario RAE

Un dato es, en principio, una cantidad o cualidad que describe un atributo de una entidad dentro de un rango de valores posibles. Es un valor «dado» con respecto de algo observado, de acuerdo con la raíz latina que da origen al término (*datum*).

1.1. ¿Qué es un dato?

Supongamos que alguien nos transmite el siguiente dato:

42

Inmediatamente nos aparece la pregunta «¿42 qué?». Con este sencillo ejemplo pretendemos mostrar, de momento, dos cosas:

1) Un dato, sin su **contexto**, carece de significado.
2) El **formato** de representación del dato es importante.

Supongamos que ahora nos dicen «la temperatura del paciente es de 42 grados». Hemos dotado de significado al 42, cuando el dato (42) es la respuesta a una pregunta («¿Cuál es la temperatura del paciente?»). Hemos avanzado un nivel y ya podemos hablar de **información**. Pero aún no somos lo suficientemente precisos. ¿Qué quiere decir que «la temperatura del paciente es de 42 grados»? Pues cosas bien distintas:

- Si son grados Celsius, el paciente tiene fiebre (42 °C).
- Si son grados Fahrenheit, el paciente es un cadáver frío (5 °C).
- Si son grados Kelvin, el paciente es un cadáver congelado a –231 °C.

Así pues, para poder hablar de información con propiedad necesitamos un tercer elemento:

3) Los datos tienen **unidades** y un rango asociado.

El 23 de septiembre de 1999 la NASA perdió el contacto con la sonda espacial Mars Climate Orbiter, un satélite diseñado para estudiar la superficie, atmósfera y clima del planeta Marte. La razón fue que el satélite entró en órbita a una altitud insuficiente, lo que causó su destrucción. El motivo de dicho error fue que una parte del software utilizado para el cálculo de las trayectorias orbitales usaba el sistema métrico decimal, mientras que otros módulos del software usaban el sistema anglosajón de unidades (pies, pulgadas, etc.). Este error costó a la NASA un total de 327,6 millones de dólares, el importe de construir el satélite, lanzarlo al espacio y controlarlo hasta su puesta en órbita, sin tener en cuenta los problemas de imagen y el retraso en la misión original.

1.2. Datos, información y conocimiento

Datos e *información* son dos palabras usadas indistintamente; de hecho, una participa en la definición de la otra, aunque no son sinónimos.

Los **datos** son los hechos o detalles de los que se deriva la **información**.

Información

Del latín, *informatio*: 'concepto' o 'explicación de una palabra'. Acción y efecto de informar (dar forma o describir).

Como hemos visto, los datos individuales en contadas ocasiones son útiles por sí mismos: necesitan tener un contexto para que puedan convertirse en información.

El **conocimiento** es la capacidad de saber, la capacidad de actuar y la capacidad de entender que reside en el cerebro.

Conocimiento

Acción y efecto de conocer (del latín, *cognoscĕre*: 'averiguar por el ejercicio de las facultades intelectuales la naturaleza, cualidades y relaciones de las cosas').

Tabla 1. Comparación datos-información-conocimiento

	Definición	Respuestas
Datos	Símbolos que representan propiedades de objetos y eventos. Propiedades básicas sin refinar ni filtrar que deben procesarse.	
Información	Cuando los datos se han procesado, organizado, estructurado y puesto en contexto, resultando entonces útiles.	Da respuesta a las preguntas «quién», «qué», «dónde» y «cuándo».
Conocimiento	Reside en las personas y resulta de la aplicación de los datos y la información.	Da respuesta a la pregunta «cómo».
Sabiduría	Reside en las personas y es la aplicación del conocimiento.	Da respuesta a la pregunta «por qué».

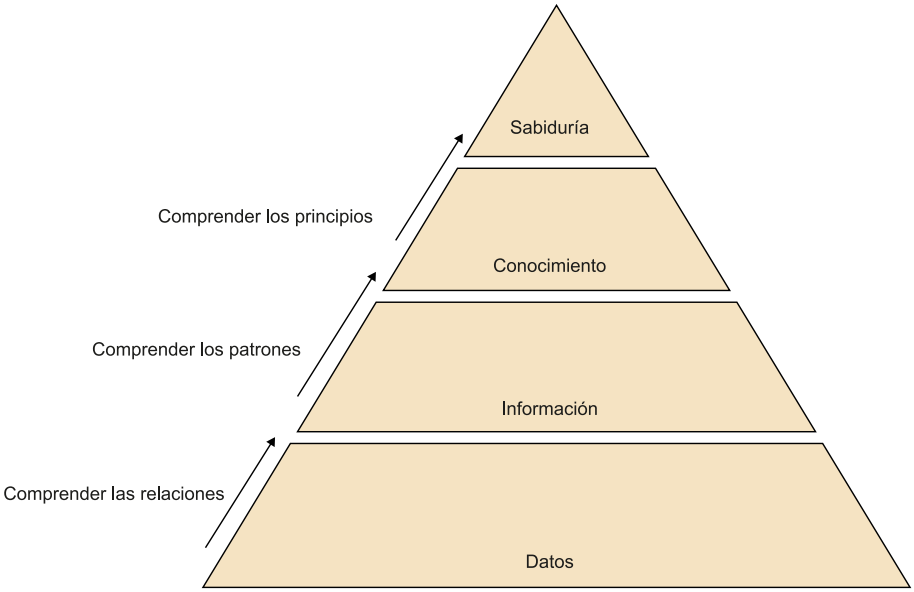
La información y el conocimiento son dos conceptos muy diferentes, aunque ambos se basan en un elemento primordial: los datos. Es lo que se conoce como la **pirámide DIKW** (*Data*, *Information*, *Knowledge* y *Wisdom*).

En la pirámide DIKW los **datos** son el nivel más básico, la **información** añade el contexto, el **conocimiento** añade cómo usar la información, y la **sabiduría** incorpora el porqué (aplicar dicho conocimiento en beneficio propio o común).

Sabiduría

De sabidor: grado más alto del conocimiento adquirido a través del estudio o de la experiencia.

Figura 1. Pirámide o jerarquía DIKW



1.3. Tipos de datos

Un detalle fundamental cuando se trata con datos es que su tratamiento depende de su naturaleza o tipo de dato. Los datos pueden ser cuantitativos o cualitativos.

1.3.1. Datos cuantitativos

Los **datos cuantitativos** (también denominados **datos numéricos**) son aquellos que pueden medirse o cuantificarse (que pueden ser contados).

Pueden ser de dos clases:

1) Datos cuantitativos continuos. Existe un continuo de valores posibles del dato, que no se restringe a valores discretos. Los valores se miden en vez de contarse. Las operaciones disponibles incluyen igual (=), distinto (≠), menor (<), mayor (>), suma (+), resta (–), multiplicación (×), división (÷), etc.

Ejemplos de datos cuantitativos continuos

Altura: 181,3 cm, 194,1 cm, 167,44 cm, etc.

Dinero: 23,53 €, 11,18 $, 1.053,99 ¥, etc.

2) Datos cuantitativos discretos. No admiten valores intermedios. Se enumeran (cuentan) más que se miden. Suelen tomar solamente valores enteros. Las operaciones disponibles incluyen igual (=), distinto (≠), menor (<), mayor (>), suma (+), resta (–), multiplicación (×), división (÷), etc.

Ejemplos de datos cuantitativos discretos

Número de quejas de los clientes: 1, 2, 5, etc.

Número de personas en el curso: 10, 15, 25, etc.

1.3.2. Datos cualitativos

Los **datos cualitativos** (también llamados **datos categóricos**) son aquellos que no se pueden expresar numéricamente y representan una cualidad o atributo que clasifica o describe a cada sujeto en una de entre varias categorías. Existe un número acotado de posibles categorías.

1) Nominales. Representan categorías sin un orden intrínseco. Las únicas operaciones disponibles son igual (=) y distinto (≠).

Ejemplos de datos cualitativos nominales

Color de pelo: rubio, moreno, castaño, pelirrojo, etc.

Género: hombre, mujer, etc.

Estado civil: casado/a, soltero/a, viudo/a, divorciado/a, etc.

2) Ordinales. Representan categorías en las que existe un orden lógico, precedencia o jerarquía (ya sea natural o asignada según alguna preferencia). La distancia entre las categorías no es, en general, conocida. Las operaciones disponibles son igual (=), distinto (≠), menor (<) y mayor (>). Los datos ordinales pueden ser:

a) Secuenciales: existe un valor inicial (o cero) y todos los valores parten de él.

b) Divergentes: es posible identificar un punto central (o cero) y los datos se encuentran por encima y por debajo del mismo.

c) Cíclicos: los valores se repiten formando ciclos.

Ejemplos de datos cualitativos ordinales

Nivel: bajo, medio, alto

Educación: básica, secundaria, universitaria

Meses del año: enero, febrero, marzo, etc.

1.4. Estructuras de datos

Los datos pueden presentarse de muchas formas diferentes. En el ejemplo de «42», se trataba de un número decimal entero, pero los datos son de naturaleza muy diversa y se pueden clasificar de acuerdo con diferentes criterios, entre otros según su estructura; así, tenemos:

- **Simples**: datos atómicos o indivisibles, con un significado propio, de acuerdo con la definición (un valor de un atributo).

- **Compuestos**: datos que son una combinación de otros datos simples y/o compuestos, de acuerdo con una estructura conocida *a priori*.

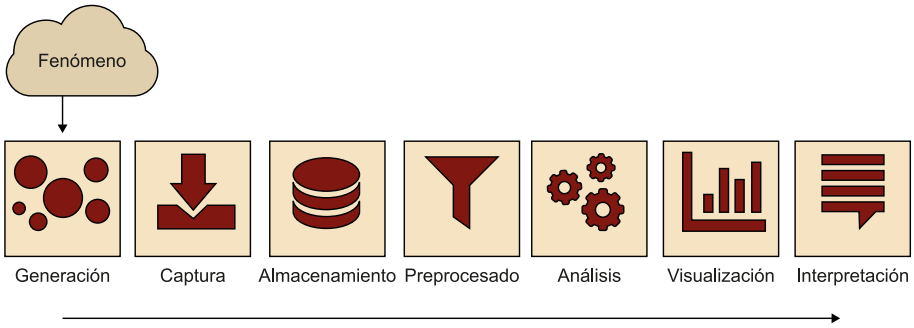
Un **conjunto de datos** (*Dataset*) es una colección de datos relacionados entre sí a los que se puede acceder individualmente o en combinación y que se gestionan conjuntamente.

Los conjuntos de datos han sido capturados y organizados para un propósito concreto. Los conjuntos de datos pueden estar segmentados en varias partes formando subconjuntos de datos separados.

2. Ciclo de vida de los datos

Existe una gestión de los datos, desde que se generan los datos hasta que se interpretan, denominada **ciclo de vida de los datos**, que consta de una serie de fases, cada una de las cuales tiene como objetivo generar valor a partir de los datos. No todas las fases son estrictamente necesarias y, además, pueden solaparse o realizarse simultáneamente en algunos casos (figura 2).

Figura 2. Ciclo de vida de los datos



2.1. Generación

El ciclo comienza con la **generación** de los datos. Supongamos que hay un fenómeno (físico, social, cultural, de comportamiento, etc.) que queremos conocer con más detalle. Este fenómeno va a ser el origen de los datos.

En el mundo actual se generan continuamente datos: cada búsqueda en un motor de búsqueda, cada clic en la web, cada vídeo que se reproduce, mensaje que se envía o lugar que se visita, contribuye a la masiva huella digital que diariamente se genera. Además, existen datos de sensores que monitorizan infraestructuras, cámaras de vigilancia para el control del tráfico y seguridad, así como internet de las cosas (IoT), con más sensores que generan más datos.

IoT

Internet of Things es un concepto de interconexión de objetos a Internet.

2.2. Captura

Una vez generados los datos, la siguiente fase es la **captura**. Esta fase tiene como objetivo recopilar los datos generados. Por motivos de relevancia, cuestiones de consideración de importancia o de capacidad de procesamiento, no siempre se recopilan necesariamente todos los datos generados. La recopilación se realiza mediante dos mecanismos básicos complementarios:

- **Creación**: se trata de integrar en el propio proceso de generación de los datos un mecanismo que almacene los que se consideren relevantes cada vez que se generen. Por ejemplo, cada vez que un usuario paga con tarjeta de crédito en un establecimiento se genera un nuevo registro que defi-

ne perfectamente qué cantidad de dinero se ha gastado, en qué establecimiento, en qué fecha y hora y con qué tarjeta.

- **Extracción**: se utiliza cuando no es posible intervenir en el proceso de generación de los datos, sino que es necesario ir capturándolos a medida que se van encontrando, idealmente inmediatamente después de su generación. Por ejemplo, es posible capturar los tuits que contienen una cierta palabra clave o *hashtag* tal y como van apareciendo en el flujo (*Timeline*) de tuits de un usuario o en el flujo público, ya que Twitter es un servicio que genera datos en abierto.

Por desgracia, no siempre se tiene acceso al nivel necesario para poder recoger los datos en el momento exacto de su creación, por lo que son más frecuentes las estrategias de captura de datos basadas en la extracción de datos ya existentes o en su generación mediante mecanismos alternativos. Entre ellos:

1) Acceso a los datos mediante un repositorio: como resultado de su publicación en abierto, los datos pueden estar disponibles en un repositorio digital, o simplemente en una web que permita su acceso.

2) Acceso mediante una API (*Application Programming Interface*): en algunos casos sí que es posible utilizar un mecanismo que permite realizar consultas específicas sobre un conjunto de datos, obteniendo solamente aquellos que han sido solicitados de acuerdo con los parámetros de la consulta.

3) Captura de datos mediante *scraping*: cuando no se dispone de una API para acceder a los datos, en ocasiones es posible utilizar herramientas (también llamadas *Bots*) que simulen la navegación de un usuario por páginas web y que extraigan el contenido de las páginas visitadas, analizando su estructura y datos.

4) Extracción de datos de documentos de texto: en ciertos casos se publican datos en diversos formatos, no siempre pensados para su reutilización, que contienen tablas, listas, etc.

5) Formularios: en ocasiones lo más sencillo y eficaz es preguntar directamente a los usuarios de un servicio, recurso o sistema, con el objetivo de recabar datos, tanto del servicio en cuestión como de los propios usuarios. A veces es necesario obtener datos directamente proporcionados por los usuarios porque interesa conocer de primera mano algunos aspectos que no pueden recogerse mediante una encuesta, especialmente el «¿por qué?» y el «¿cómo?», y aspectos difícilmente cuantificables, como emociones o sentimientos.

2.3. Almacenamiento

Los datos capturados tienen que almacenarse en un formato que permita su posterior manipulación, conforme a la representación más adecuada, teniendo en cuenta tanto su tipología como el uso que se vaya a querer efectuar de ellos. En función de su objetivo, frecuencia de acceso y complejidad se puede hablar de:

- **Archivos simples**: los datos se almacenan en archivos (o colecciones de archivos) según ciertos criterios; por ejemplo, el origen de los datos o la fecha de captura.

- **Bases de datos**: se trata de una distribución más o menos compleja que permite representar los datos de acuerdo con su estructura, teniendo también en cuenta las relaciones entre todos los elementos que los componen. Se suele hablar de bases de datos relacionales o SQL, y también de las denominadas no tradicionales o NoSQL, que pretenden dar solución a problemas que tienen que ver con la escalabilidad de las relacionales, especialmente para grandes volúmenes de datos o para bases de datos de grafos.

Ejemplo de archivos simples

Un ejemplo de archivo simple pueden ser los ficheros del registro de actividad de los servidores (*log*) que contienen todas las peticiones que se realizan cuando los usuarios navegan por las páginas web de un servicio en línea.

2.4. Preprocesado

El objetivo de esta etapa es preparar los datos para su análisis posterior, de forma que puedan usarse en procesos de ciencia de datos o de aprendizaje automático, sin tener que preocuparse por aspectos relacionados con su calidad, procedencia, etc. Entre otras operaciones típicas de esta etapa se pueden destacar las siguientes:

1) Fusión: los datos se obtienen de diferentes fuentes, por lo que es necesario combinarlos en una única estructura.

2) Selección/filtrado: consiste en obtener algunos datos que son de interés, según ciertos criterios de búsqueda.

3) Conversión: frecuentemente los datos están en formatos que dificultan su análisis posterior, por lo que es necesario convertirlos a un formato más manejable.

4) Limpieza: también conocida como *Data Cleaning*, consiste en eliminar todas las inconsistencias en los datos que puedan ser detectadas, con el propósito de garantizar la validez de los datos.

5) Agregación: en algunos casos puede resultar interesante agrupar un subconjunto de datos de forma que se simplifique el conjunto de datos original y se genere una nueva variable que tenga un mayor poder predictivo.

6) Creación de nuevas variables/indicadores (variables derivadas): en ocasiones es necesario realizar cálculos con las variables para, por ejemplo, calcular la relación entre dos variables, generando nuevas variables o indicadores.

Es importante destacar la importancia de esta fase, puesto que la calidad de los resultados obtenidos dependerá, directamente, de la calidad de los datos a partir de los cuales hayan sido obtenidos los resultados.

Es lo que se conoce como *Garbage in*, *Garbage out*, es decir, si se usan datos basura para crear modelos, estos serán, probablemente, también basura. Además, resulta conveniente almacenar los datos limpios para futuras reutilizaciones.

2.5. Análisis

Una vez que los datos ya se consideran válidos, se puede proceder a su **análisis**. Aquí se incluyen todas las técnicas computacionales y estadísticas de análisis de datos con el propósito de: obtener conocimiento, construir clasificadores, construir predictores o inferir causalidad mediante la utilización de algoritmos y métodos de inteligencia artificial, de la minería de datos, del aprendizaje automático y de la teoría de la inferencia estadística.

El objetivo de esta etapa es obtener **modelos** que expliquen cómo son los datos y sus características principales, y poder responder a las preguntas planteadas sobre el fenómeno que ha originado los datos.

El análisis no se limita a la construcción de modelos, sino que debe también explicar el resultado obtenido, mediante una interpretación del modelo y su puesta en contexto con respecto al fenómeno original. Esto incluye también la evaluación del propio modelo, identificando qué variables o características son las más relevantes, la capacidad de generalización ante datos nunca utilizados anteriormente en la creación del modelo, o su capacidad de adaptación a los cambios en los datos.

2.6. Visualización

Los humanos disponemos de un sistema visual muy complejo y avanzado, que incluye desde el ojo hasta el córtex visual, el encargado de procesar toda la información recogida por el primero. Los humanos somos, principalmente,

máquinas de procesado visual, con diversos subsistemas que se encargan de procesar eficientemente diferentes aspectos de dicha tipología de información de manera separada: forma, movimiento, color, etc.

La **visualización** de datos ayuda a presentar los resultados del análisis de una forma clara y sencilla que un ser humano pueda comprender visualizándola. En esta fase, para transmitir mejor los resultados del análisis, es necesario considerar, además de la funcionalidad, la estética y la capacidad de percepción visual humana.

Un aspecto interesante de la visualización de datos es que puede convertirse en una interfaz de navegación de los propios datos, permitiendo ciertas operaciones básicas (selección, agregación, etc.), de manera que sea posible añadir interactividad a la visualización.

De esta forma, es posible basar el análisis de los datos a partir de su visualización, de modo que se combinen la capacidad visual humana para la detección de patrones, tendencias, etc., con la potencia de un sistema informático que permita seleccionar, filtrar o comparar datos.

2.7. Interpretación

Mediante la **interpretación** proporcionamos una explicación de lo que significa la visualización, y generamos un relato que explique el contexto, las implicaciones y las posibles opciones que se extraen de la visualización.

Esta interpretación puede ser publicada en forma de nuevos datos, de manera que sea posible que terceros puedan reutilizarlos con otros propósitos, especialmente los datos ya procesados listos para su análisis, en forma de una o más tablas.

En todo el ciclo de vida de los datos se han obviado las posibles realimentaciones existentes. Inevitablemente, después de presentar algunas observaciones y descubrimientos sobre los datos, se pueden generar nuevos interrogantes y cuestiones que podrían exigir recopilar más datos o realizar otros tipos de análisis.

Bibliografía

Ackoff, R. (1989). «From data to wisdom». *Journal of Applied Systems Analysis* (vol. 16, n.º 1, págs. 3-9).

Chen, M. y otros (2009). «Data, Information, and Knowledge in Visualization». *IEEE Computer Graphics and Applications* (vol. 29, n.º 1, págs. 12-19).

Murray, S. (2013). *Interactive data visualization for the Web*. O'Reilly Media.

Stanton, J. M. (2013). *Introduction to Data Science*. Nueva York: Syracuse University.

Witten, I. H.; Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.