# Introducció al cicle de vida de les dades

PID_00265702

Laia Subirats Maté
Diego Oswaldo Pérez Trenard
Mireia Calvo González

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

**Laia Subirats Maté**

Enginyera de Telecomunicacions per la Universitat Pompeu Fabra (2008), màster en Telemàtica per la Universitat Politècnica de Catalunya (2009) i doctora en Informàtica per la Universitat Autònoma de Barcelona (2015). Des de 2009, treballa com a investigadora a Eurecat (Centre Tecnològic de Catalunya) aplicant la ciència de dades a diferents camps com ara la salut, el medi ambient o l'educació. Des de 2016, col·labora amb la UOC com a docent en el màster de Data Science i en el grau d'Informàtica. És especialista en intel·ligència artificial, ciència de dades, salut digital i representació del coneixement.

**Diego Oswaldo Pérez Trenard**

Enginyer electrònic per la Universitat Simón Bolívar (2015), especialització en High Tech Imaging (HTI) per la Universitat Télécom SudParis (2014) i doctor en Senyals, Imatges i Visió per la Universitat de Rennes 1 (2018). Des del 2014, ha treballat com a enginyer de recerca i desenvolupament a l'Institut Nacional de Salut i Investigació Mèdica (INSERM) i al Laboratori de Processament de Senyals i Imatges (LTSI), aplicant coneixements en electrònica i en processament de dades a l'estudi de diferents malalties neurològiques, cardíaques i respiratòries. Des del 2018, col·labora com a docent en el màster de Data Science de la UOC.

**Mireia Calvo González**

Enginyera de telecomunicacions per la Universitat Politècnica de Catalunya (2011), Màster en Enginyeria Biomèdica per la Universitat de Barcelona i per la Universitat Politècnica de Catalunya (2014) i Doctora en Processament de senyals i telecomunicacions per la Universitat de Rennes 1 i en Enginyeria Biomèdica per la Universitat Politècnica de Catalunya (2017). Des del 2012 ha treballat com a investigadora en diferents entorns acadèmics, clínics i industrials, aplicant el processament de dades a l'estudi de diferents malalties cardíaques i respiratòries. Des del 2017 col·labora amb la UOC com a docent en el Màster de Data Science.


L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Isabel Guitart Hormigo (2019)

Primera edició: setembre 2019

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

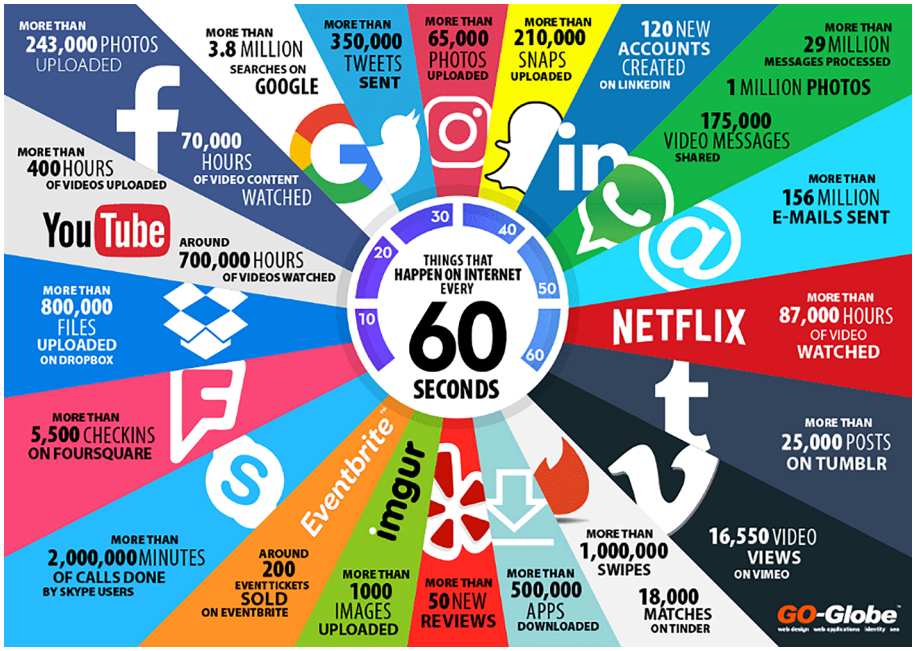
# Índex

# Introducció

Al llarg de la història hi ha hagut diferents revolucions industrials fins arribar a la societat de la informació i el coneixement. En concret, la primera revolució industrial va sorgir al segle XVIII, quan va aparèixer la màquina de vapor. La segona revolució industrial va tenir lloc entre els segles XIX i XX amb la producció massiva d'electricitat. La tercera revolució industrial, o primera revolució de la informació, va aparèixer a la fi del segle XX amb la irrupció de les tecnologies de la informació i comunicació (TIC). Finalment, a principis del segle XXI es va produir la quarta revolució industrial, o segona revolució de la informació, basada en la intel·ligència artificial.

Una de les característiques de la societat de la informació i el coneixement és el nombre creixent de dades generades tant per individus com empreses, també conegut com a *datificació*. Per exemple, Computer Sciences Corporation estima que l'any 2020 hi haurà 44 vegades més de dades de les que hi havia al 2009. Per altra banda, segons go-globe.com, cada 60 segons es genera la quantitat d'informació que es mostra a la figura 1.

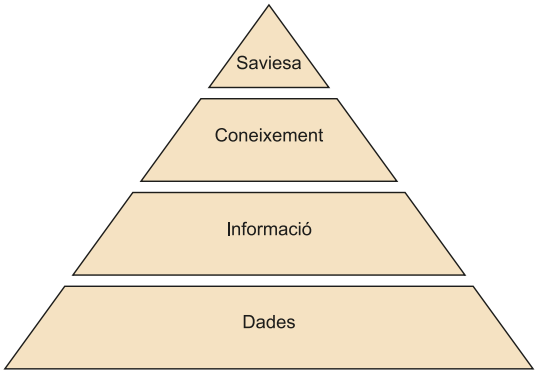
Figura 1. Informació generada a Internet cada 60 segons




Font: https://www.go-globe.com

Per accedir a la saviesa des de les dades, definim la piràmide "Dades, Informació, Coneixement, Saviesa (DICS)", com es mostra a la figura 2. En aquesta piràmide partim de les dades; després, segons el context i la connexió de parts, arribem a la informació. El pas següent és el coneixement i, finalment, l'últim pas és la saviesa.

Figura 2. Piràmide DICS




Dins de l'àmbit de la ciència de les dades, diferents perfils professionals fan possible el trànsit reflectit a la piràmide. A continuació, indiquem alguns exemples:

**Més informació a:**

La infografia "The data science industry. Who does what".

- **Científic de dades** (*data scientist*). Neteja i organitza les dades. És un analista de dades curiós.

- **Analista de dades** (*data analyst*). Recull, processa i realitza anàlisis estadístics de les dades. Intuïtiu amb capacitat per desxifrar problemes i dades.

- **Arquitecte de dades** (*data architect*). Gestiona els sistemes per integrar, centralitzar, protegir i mantenir fonts de dades. Crea plans per a la gestió de dades per integrar, centralitzar, protegir i mantenir les fonts de dades.

- **Enginyer de dades** (*data engineer*). Desenvolupa, construeix, testeja i manté arquitectures. Desenvolupa, construeix, prova i manté arquitectures (com a bases de dades i sistemes de processament a gran escala).

- **Estadístic** (*statistician*). Recull, analitza i interpreta dades qualitatives i quantitatius amb teories i mètodes estadístics. És un professional de la lògica i domina l'estadística.

- **Administrador de base de dades** (*database administrator*). Assegura que la base de dades estigui disponible per a tots els usuaris rellevants, que funcioni correctament i es mantingui segura.

- **Analista de negoci** (*bussiness analyst*). Millora els processos de negoci com a intermediari entre negoci i tecnologies de la informació.

- **Líder de ciència de dades** (*data science leader*). Gestiona un equip d'analistes i científics de dades.

A la taula 1 es mostren els diferents llenguatges que dominen els diferents perfils:

Taula 1. Llenguatges de programació dels diferents perfils de la ciència de dades

| | Científic de dades | Analista de dades | Arquitecte de dades | Enginyer de dades | Estadístic | Administrador de base de dades | Analista de negoci | Líder de ciència de dades |
|---|---|---|---|---|---|---|---|---|
| R | X | X | | X | X | | | X |
| SAS | X | | | X | X | | | X |
| Python | X | X | | X | X | X | | X |
| Matlab | X | | | X | X | | | X |
| SQL | X | X | X | X | X | X | X | X |
| Hive / Pig | X | | X | X | X | | | |
| Spark | X | | X | | X | | | |
| HTML / Javascript | | X | | | | | | |
| C / C++ | | X | | X | | | | |
| SPSS | | | | X | X | | | |
| Java | | | | X | | X | | X |
| Ruby | | | | X | | X | | |
| Perl | | | | X | X | | | |
| XML | | | X | | | | | |
| C# | | | | | | X | | |
| Stata | | | | | X | | | |

# Objectius

En aquest material didàctic es proporcionen les eines fonamentals que permetran assimilar els objectius següents:

1. Entendre què és la generació de dades i el concepte de *societat de la informació*.

2. Ser capaç d'identificar els diferents perfils que intervenen en la ciència de dades.

3. Conèixer el significat i la classificació de les dades segons la seva estructura, nivell d'accés i tipus d'informació.

4. Ser capaç d'avaluar la qualitat de les dades.

5. Conèixer les diferents etapes del cicle de vida de les dades: captura, emmagatzemament, preprocessat, anàlisi, representació i publicació.

# 1. Què són les dades?

Una dada és, en principi, una quantitat o qualitat que descriu un atribut d'una entitat, dins d'un rang de valors possibles. És un valor «donat» al respecte d'alguna cosa observada, d'acord amb l'arrel llatina que dona origen al terme (datum).

**Bibliografia recomanada**

**Minguillón, Julià** (2016). «Fundamentos de Data Science». Editorial UOC.

Les dades poden classificar-se de diferents maneres. En els següents subapartats explicarem els diferents tipus de dades i els aspectes que determinen la seva qualitat de vida.

## 1.1. Classificació de les dades

Les dades es poden classificar segons la seva estructura, nivell d'accés i el tipus d'informació que contenen.

### 1.1.1. Estructura de les dades

Segons la seva estructura podem classificar les dades de la manera següent:

- **Dades simples**. Dades atòmiques indivisibles, amb un significat propi, d'acord amb la definició (un valor d'un atribut). Es tracta de dades reals o cadenes, per exemple, el pes corporal.

- **Dades compostes o estructurades**. Dades que són una combinació d'altres dades simples o compostes, d'acord amb una estructura fixa i coneguda *a priori*. Un exemple de dades compostes seria una radiografia toràcica.

- **Dades semiestructurades**. Dades estructurades que segueixen una estructura parcial o que pot canviar segons el context, o que no segueixen cap estructura. Un exemple seria una pàgina HTML.

### 1.1.2. Nivell d'accés

Pel que fa a la seguretat, segons la Universitat Carnegie Mellon, les dades poden classificar-se en tres nivells:

- **Dades restringides**: les dades s'han de classificar com a restringides quan la divulgació no autoritzada, l'alteració o la destrucció d'aquestes dades podria causar un nivell significatiu de risc per a la universitat o els seus

afiliats. Els exemples de dades restringides inclouen dades protegides per regulacions de privacitat estatals o federals i dades protegides per acords de confidencialitat. El nivell més alt de controls de seguretat s'ha d'aplicar a les dades restringides.

- **Dades privades**: les dades s'han de classificar com a privades quan la divulgació no autoritzada, l'alteració o la destrucció d'aquestes dades podria representar un nivell moderat de risc per a la institució. De manera predeterminada, totes les dades institucionals que no es classifiquen explícitament com a restringides o dades públiques s'han de tractar com a dades privades. S'ha d'aplicar un nivell raonable de controls de seguretat a les dades privades.

- **Dades públiques**: les dades s'han de classificar com a públiques quan la divulgació no autoritzada, l'alteració o la destrucció d'aquestes dades comporti un risc petit o nul per a la institució. Com a exemples de dades públiques s'inclouen comunicats de premsa i publicacions. Això no obstant, cal tenir en compte que es requereix cert nivell de control per evitar la modificació o destrucció no autoritzada de les dades públiques.

A vegades també es classifiquen les dades segons el possible impacte de seguretat que tinguin en l'organització. A la taula 2 es mostren tres objectius de seguretat (confidencialitat, integritat i disponibilitat) classificats segons l'impacte de seguretat de les dades (baix, moderat o alt).

Taula 2. Classificació segons l'impacte de seguretat de les dades

| Objectiu de seguretat | Impacte baix | Impacte moderat | Impacte alt |
|---|---|---|---|
| Confidencialitat | La divulgació no autoritzada d'informació pot tenir un efecte advers **limitat** a l'organització. | La divulgació no autoritzada d'informació pot tenir un efecte advers **greu** a l'organització. | La divulgació no autoritzada d'informació pot tenir un efecte advers greu o **catastròfic** a l'organització. |
| Integritat | La modificació o destrucció no autoritzada de la informació pot tenir un efecte advers **limitat** a l'organització. | La modificació o destrucció no autoritzada de la informació pot tenir un efecte advers **greu** a l'organització. | La modificació o destrucció no autoritzada de la informació pot tenir un efecte advers greu o **catastròfic** a l'organització. |
| Disponibilitat | La interrupció de l'accés o ús de la informació o un sistema d'informació puguin tenir un efecte advers **limitat** a l'organització. | La interrupció de l'accés o ús de la informació o un sistema d'informació puguin tenir un efecte advers **greu** a l'organització. | La interrupció de l'accés o ús de la informació o un sistema d'informació pu-guin tenir un efecte advers greu o **catastròfic** a l'organització. |

Els beneficis d'aquestes classificacions són els següents:

- Compliment de les dades i una gestió de riscos més senzilla. Les dades se situen on s'espera en el nivell d'emmagatzematge predefinit i «punt en el temps».

- Simplificació del xifrat de dades perquè no cal xifrar totes les dades. Això estalvia valuosos cicles de processador i tota la conseqüència relacionada.

- Indexació de dades per millorar els temps d'accés dels usuaris.

- La protecció de dades es redefineix quan es millora el RTO (objectiu de temps de recuperació o en anglès *recovery time objective*).

### 1.1.3. Tipus d'informació

Des d'un punt de vista estadístic, les dades poden classificar-se en quantitatives (o numèriques) i qualitatives. Les dades qualitatives poden classificar-se en ordinals (que es poden ordenar) o nominals (no hi ha un ordre). Les dades qualitatives també poden classificar-se en binàries (dicotòmiques) o categòriques (multicotòmiques).

**Exemple**

- Dada qualitativa nominal: les llengües que parles (multicotòmica) o si ets major d'edat (dicotòmica).
- Dada qualitativa ordinal: el grau de formació (elemental, formació professional, graduat, màster, metge, etc.).
- Dades quantitatives: la temperatura.

**Referència bibliogràfica**

**Kristin H. Jarman** (2013). *The art of data analysis. How to answer almost any question using basic statistics*. Hoboken, NJ: Wiley.

Per tractar i resumir dades qualitatives normalment s'utilitza el que s'anomena *anàlisi de freqüència*. Aquesta consisteix simplement en comptar quantes dades hi ha a cada categoria. També és útil mostrar les dades en freqüència relativa (percentatge). Hi ha diferents maneres de visualitzar dades qualitatives com gràfics circulars o gràfics de barres. En els valors ordinals també es poden fer servir percentils, mediana, moda i el rang interquartil per resumir les seves dades. Per a dades quantitatives es poden usar histogrames o gràfic de caixes i bigotis (en anglès *box plot*).

## 1.2. Qualitat de les dades

Hi ha diferents factors que influeixen en la qualitat de les dades: exactitud, completitud, consistència, atemporalitat, unicitat i validesa. A continuació, per a cada factor s'explica la definició, referència, mesura, l'àmbit, la unitat de mesura i les dimensions relacionades.

**Més informació a:**

Veure l'article "The Six primary dimensions for data quality assessment. Defining Data Quality Dimensions".

### 1.2.1. Exactitud

Es defineix com el grau en què les dades descriuen correctament l'objecte o esdeveniment del «món real». Idealment, la veritat del «món real» s'estableix a través de la investigació primària. Això no obstant, com que sovint això no és pràctic, és comú utilitzar dades de referència de tercers, de fonts que es consideren fiables i de la mateixa cronologia, la mesura és el grau de representació

del món real i l'àmbit és la base de dades, la unitat de mesura és el percentatge de dades que passen les regles d'exactitud. Les dimensions relacionades són la validesa, la unicitat i la consistència.

**Exemple**

Una infermera, atès que havia arribat recentment d'Europa i encara no estava familiaritzada amb el format de datació nord-americà, va entrar totes les dates de naixement amb el format europeu (DD / MM / AAAA), en lloc del nord-americà (MM / DD / AAAA). Les dades van passar la fase de validació del sistema ja que els valors es mantenien dins del rang, però no eren exactes: el sistema va entendre que tots els infants les dades dels quals es van entrar aquell dia havien nascut el 5 de gener del 2014.

### 1.2.2. Completesa

Es defineix com la proporció de dades emmagatzemades davant del potencial «100 % complet». La referència són les regles de negoci que defineixen el que representa el «100 % complet». La mesura és l'absència de valors en blanc (nul) o la no presència de valors en blanc. L'àmbit és la base de dades i la unitat de mesura és un percentatge. Les dimensions relacionades són la validesa i l'exactitud.

**Exemple**

Percentatge de pacients que tenen tots els elements de dades mínimes i bàsiques, segons el que es defineix per l'estàndard, sense valors en blanc.

### 1.2.3. Consistència

És l'absència de diferència, a l'hora de comparar dos o més representacions d'una cosa amb la seva definició. La referència és cada ítem, l'àmbit és la base de dades i la unitat de mesura és un percentatge. Les dimensions relacionades són l'exactitud, la validesa i la unicitat.

**Exemple**

Els tres camps que s'utilitzen per documentar els resultats d'una prova d'audició són: «oïda esquerra», «oïda dreta» i «general». El camp general està dissenyat per adaptar-se a casos específics, quan part de la informació sobre alguna de les dues orelles no està disponible; aquest camp ha de ser un valor calculat automàticament en funció dels resultats de les dues orelles. Per exemple, general és OK si i només si esquerra i dreta són OK. L'avaluació de valors en aquests tres camps s'ha de realitzar per assegurar la consistència.

### 1.2.4. Puntualitat

Es defineix la puntualitat (o atemporalitat) com el grau en què les dades representen la realitat des d'un punt requerit en el temps. La referència és el temps de l'esdeveniment real que ha estat obtingut i la mesura és la diferència en el temps. L'àmbit és qualsevol ítem relacionat amb la base de dades i la unitat de mesura és el temps. Una dimensió relacionada amb aquesta és l'exactitud.

**Exemple**

Diferència horària entre la finalització de la prova d'audició d'un pacient, la visita de diagnòstic i la introducció en el sistema de la informació sobre aquesta visita.

### 1.2.5. Unicitat

Ens assegura que res no es registra més d'una vegada. És l'invers d'una avaluació del nivell de duplicació, la referència és la mateixa i l'àmbit és el conjunt de dades. Es mesura com a percentatge i una dimensió relacionada amb aquesta és la consistència.

**Exemple**

Percentatge de valors duplicats d'un conjunt de dades d'audició.

### 1.2.6. Validesa

Es defineix com l'ajustament a la sintaxi predefinida (format, tipus, rang). La referència és la base de les dades, metadades o les regles de documentació, segons els tipus permesos (cadena, sencer, punt flotant, etc.), el format (longitud, nombre de dígits, etc.) i rang (mínim, màxim o contingut dins d'un conjunt de valors permesos). La mesura és la comparació entre les dades i les metadades o la documentació. L'àmbit és totes les dades i la unitat de mesura és el percentatge de dades vàlides o invàlides. Dimensions relacionades amb la validesa són l'exactitud, la completesa, la consistència i la unicitat.

**Exemple**

Un exemple de validesa pel que fa a l'element de dades és el tipus i la severitat de la pèrdua auditiva, que s'han de triar d'una llista donada de valors permesos.

# 2. Cicle de vida de les dades

Amb els importants avenços en informàtica i en tecnologies relacionades, i la seva implementació cada vegada més gran en totes les àrees del coneixement, quantitats significatives de dades amb diverses característiques són cada dia capturades i emmagatzemades en bases de dades. Per altra banda, l'avanç del poder de còmput dels ordinadors no s'ha quedat enrere. Aquesta simbiosi ha permès que grans quantitats d'informació digitalitzada sigui fàcilment capturada, emmagatzemada i processada.

No obstant això, obtenir algun tipus de coneixement a partir d'aquest enorme volum de dades resulta una tasca àrdua i complexa. La raó és que, normalment, una col·lecció de dades «en brut» no atorga informació rellevant. El valor d'aquestes dades s'obté després d'un processat que permeti extreure informació útil per facilitar una presa de decisions o per a una correcta comprensió d'un fenomen que governi a la font de les dades.

**Bibliografia recomanada**

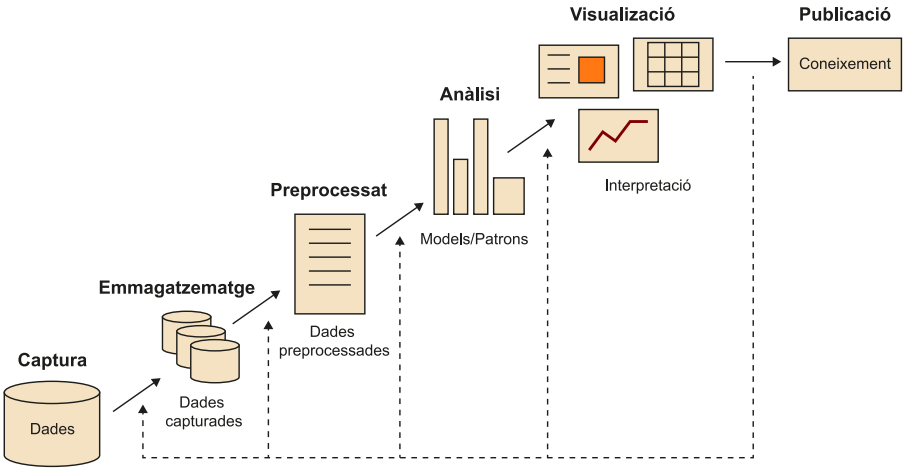
**Han, Jiawei**; **Kamber, Micheline**; **Pei, Jian** (2012). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.
**Riquelme, José Cristóbal**; **Ruiz, Roberto**; **Gilbert, Karina** (2006). «Minería de datos: conceptos y tendencias». *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial* (vol. 10, núm. 29, pàg. 11-18).

L'extracció de coneixement a partir d'un grup donat de dades per a la resolució d'un cert problema és un procés iteratiu. En aquest sentit, es defineixen sis fases o etapes típiques en el cicle de vida de les dades: captura, emmagatzematge, preprocessat, anàlisi, visualització i publicació. Cadascuna d'aquestes etapes té un objectiu, generant-hi valor a partir de les dades en cadascuna. Tot i que no totes les etapes són estrictament necessàries. En els apartats següents s'explicarà més en detall cadascuna d'aquestes fases (veure figura 3).

**Bibliografia recomanada**

**Fayyad, Usama**; **Piatetsky-Shapiro, Gregory**; **Smyth, Padhraic** (1996, març). «From data mining to knowledge discovery in databases». *AI Magazine* (vol. 17, núm. 3, pàg. 37-54).

Figura 3. Etapes típiques en el cicle de vida de les dades



## 2.1. Captura

La primera fase del cicle de vida de les dades té com a objectiu l'adquisició de totes les dades que es puguin generar durant un procés donat. Aquesta fase és coneguda com a captura i pot subdividir-se segons el tipus de mecanisme utilitzat per a aquesta adquisició: creació i extracció.

**Bibliografia recomanada**

**Minguillón, Julià** (2016). «Fundamentos de Data Science». Editorial UOC.

### 2.1.1. Creació

Aquest mecanisme consisteix en la implementació d'una rutina o algoritme dins del procés que està generant les dades, amb l'objectiu d'emmagatzemar les dades que es considerin rellevants a mesura que aquestes es vagin generant. En altres paraules, les dades són capturades en temps real cada vegada que es generen.

Aquest mecanisme pot ser implementat quan es té accés al procés o sistema i s'hi pot intervenir, ja sigui utilitzant rutines preexistents que ens permetin l'obtenció de les dades o amb la creació de noves rutines que puguin ser afegides al procés.

**Exemple**

En una estació meteorològica es desitja obtenir dades de pluviometria en una zona específica per poder comparar-les amb altres zones d'interès. Per a això, suposant que es té accés al procés, es decideix instal·lar pluviòmetres a la zona que s'analitza, els quals envien la informació a un servidor que emmagatzema les diferents mesures que aquests aporten durant el dia. Com es pot apreciar en aquest exemple, s'ha afegit una nova rutina (implementació dels pluviòmetres) a un procés preexistent (estació meteorològica) al qual es té accés. D'aquesta manera, és possible capturar dades completament noves que han estat creades per intentar respondre a un problema específic.

Hi ha altres tècniques molt utilitzades en la creació de dades com:

**a) Creació de formularis:** és una tècnica que consisteix a elaborar una sèrie de preguntes que van dirigides als usuaris amb la finalitat d'obtenir informació (dades) sobre un sistema, procés o servei. Tot i ser una tècnica relativament senzilla i eficaç, es considera lenta i costosa, a més de presentar limitacions temporals. No obstant això, amb l'aparició de les xarxes socials i les diverses eines que permeten la creació de formularis en línia, és possible arribar a una gran quantitat d'usuaris augmentant, així, la popularitat d'aquesta tècnica.

**b) Crowdsourcing:** es tracta d'un concepte que va ser creat per a la resolució de problemes molt complexos que resulten difícils de resoldre per mètodes convencionals com la mineria de dades o la intel·ligència artificial. Es basa en aprofitar l'intel·lecte humà i la seva capacitat de raonament, superior als mètodes actuals, per resoldre un problema donat preguntant directament als usuaris sobre la possible solució.

**Exemple**

Un exemple d'aquesta tècnica seria en la classificació d'imatges que poden ser posteriorment utilitzades per entrenar un algoritme de *machine learning*. Aquesta tècnica s'observa,

típicament, en els sistemes de control basats en CAPTCHA (que és l'acrònim en anglès de *completely automated public Turing test to tell computers and humans apart*), que eviten que una màquina (bot) es registri de manera automàtica a un servei. Com es pot veure a la figura 4, el sistema li pregunta a l'usuari quines imatges corresponen a senyals de trànsit. En la imatge es tracta de la sortida d'un sistema de reconeixement de vídeo en el qual certs fragments no van ser processats correctament, per la qual cosa, a mesura que els usuaris seleccionen els fragments de la imatge on es troben aquests senyals, l'algoritme aprèn i li permet optimitzar el seu rendiment.

Figura 4. Exemple d'un CAPTCHA


**c) Dades qualitatives:** s'implementa quan les dades que es volen obtenir són difícilment quantificables, com sentiments, motivacions o emocions, de manera que es recorre a entrevistes on, per mitjà de preguntes, s'obté la informació directament de l'usuari. Les entrevistes normalment són gravades per després transcriure-les segons criteris preestablerts i poder identificar la informació rellevant que es vol obtenir.

### 2.1.2. Extracció

Aquest mecanisme s'utilitza quan no es té accés al sistema o no es pot intervenir en el procés de generació de les dades. Consisteix a capturar les dades que ja van ser generades per un procés per mitjà d'una recerca de les dades d'interès. Idealment, s'intenta que aquestes dades trobades s'emmagatzemin immediatament després de ser generades.

**Exemple**

Es pot dissenyar un algoritme que capturi els títols de les diferents notícies que apareixen a una certa pàgina web informativa. D'aquesta manera, en no tenir accés al procés de generació de dades, podem extreure-les de la mateixa manera segons el seu ordre d'aparició.

En la majoria dels casos, intervenir en el procés de generació de les dades no és possible, de manera que les tècniques d'extracció de dades són els mecanismes més habituals. Hi ha diverses tècniques d'extracció, essent les més utilitzades les següents:

**a) Accés mitjançant repositoris:** consisteix en l'adquisició de les dades per mitjà d'una descàrrega d'un repositori digital publicat en obert. Moltes institucions o pàgines web publiquen les seves dades i en permeten l'accés a tots els usuaris que vulguin utilitzar-les. Aquestes dades venen normalment organitzades en fitxers segons una classificació preestablerta.

**Exemple**

El World Bank Group és una institució que realitza anàlisis econòmiques i estadístiques dels diferents països. Aquesta organització permet l'accés i descàrrega de les seves dades en fitxers presentats en diferents formats, la qual cosa permet als usuaris la realització d'anàlisis independents.

**b) Extracció mitjançant una *application programming interface* (API):** en certs casos, les institucions o pàgines web ofereixen als usuaris més que la simple possibilitat de la descàrrega i accés de les seves dades per mitjà d'un fitxer, i permeten la realització de consultes específiques dins d'aquestes dades, de manera que s'obtingui només un conjunt que compleixi amb un cert paràmetre de consulta. Aquestes interfícies de consulta resulten interessants per a la programació d'algoritmes que recopilin dades de manera automatitzada.

**c) Manipulació de paràmetres de cerca:** s'utilitza quan els llocs web no disposen d'una API per realitzar consultes específiques, però es pot accedir als fitxers de dades mitjançant el seu URL. A vegades, els paràmetres del URL indiquen l'arquitectura del web. Aquesta característica permet que es puguin dissenyar *scripts* que automatitzin un procés de recerca, atorgant a l'usuari la capacitat de seleccionar subconjunts de dades d'interès.

**Exemple**

L'Institut Nacional d'Estadística i Cens (INEC) de Panamà permet l'accés a les seves dades directament des de la seva pàgina web, de manera que es poden manipular els paràmetres de l'URL per accedir a les dades desitjades. La secció de dades es troba estructurada de la manera següent: `http://www.contraloria.gob.pa/INEC/Avance/Avance.aspx?ID_CATEGORIA=<CAT>&ID_CIFRAS=<CIF>&ID_IDIOMA=<IDI>`. On:

- <CAT>
  - Indicadors de conjuntura = 1
  - Sector real = 2
  - Sectors fiscal i financer = 3
  - Etc.
- <CIF>
  - Índex Mensual d'Activitat Econòmica = 1
  - Indicadors de comerç exterior = 2
  - Etc.
- <IDI>
  - Castellà = 1

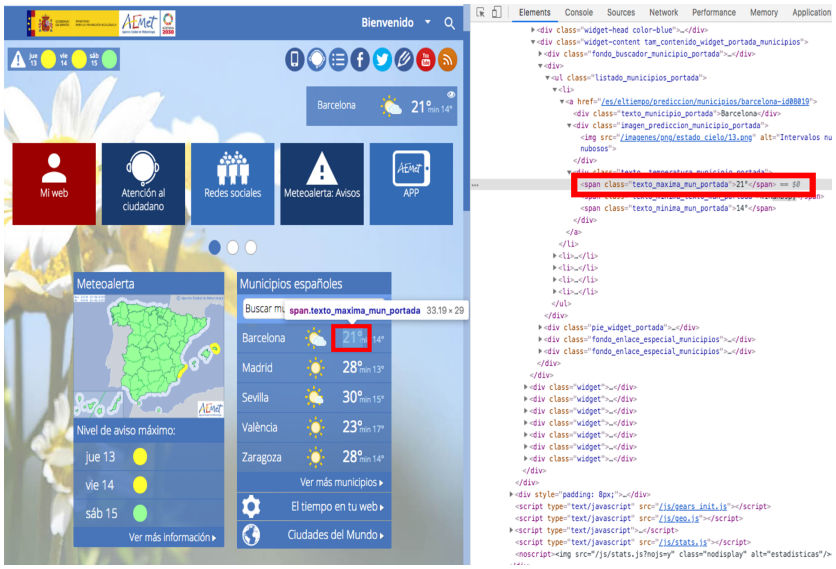
Tots són paràmetres que poden ser manipulats. D'aquesta manera, si es vol accedir a les dades del sector fiscal, es col·loca: `http://www.contraloria.gob.pa/INEC/Avance/Avance.aspx?ID_CATEGORIA=3&ID_CIFRAS=16&ID_IDIOMA=1`

**d) *Web scraping*:** en realitat, els llocs web no disposen d'una API per a la realització de consultes ni la possibilitat de descàrrega i accés a fitxers. Per aquesta raó, la tècnica més utilitzada d'extracció de dades de manera semiautomàtica és la coneguda com a *web scraping*. Aquesta tècnica consisteix en la implementació d'eines conegudes com bots que simulen la navegació d'un usuari real dins d'una pàgina web. Aquestes eines, a més de navegar dins del lloc, n'extreuen el contingut i la informació per poder ser utilitzada posteriorment. Aquest mecanisme és possible gràcies a l'estructura coherent de les pàgines web i la possibilitat de la seva inspecció per determinar el format d'aquesta estructura interna. D'aquesta manera, quan coneguem l'arquitectura de la pàgina, és possible programar un bot capaç de buscar i accedir a dades específiques dins del contingut.

**Exemple**

A la figura 5 podem observar com la pàgina de l'Agència Estatal de Meteorologia d'Espanya (AEMET) està estructurada en un format HTML, en el qual es pot accedir a dades específiques dins del contingut de la pàgina. En aquest cas, podem veure com es podria accedir a la temperatura actual a Barcelona movent-se a través de l'estructura del lloc. Un bot pot realitzar aquesta tasca de manera automàtica, amb la qual cosa és possible obtenir aquesta dada cada vegada que es requereixi.

Figura 5. Pàgina HTML de l'Agència Estatal de Meteorologia d'Espanya




## 2.2. Emmagatzematge

La següent fase del cicle de vida de les dades, després de la captura, és l'emmagatzematge de les dades en un format o representació adequada que, segons la seva tipologia, en permeti la utilització posterior de la manera més simple possible. A grans trets, hi ha dues maneres típiques d'emmagatzemar les dades, en fitxers simples o en bases de dades.

Els fitxers simples són estructures on les dades s'emmagatzemen segons uns criteris preestablerts. Per exemple, un fitxer pla que ens doni informació de la consola del computador sobre tots els processos realitzats durant un termini definit.

Per altra banda, les bases de dades consisteixen en una col·lecció de dades interrelacionades que solen manipular-se mitjançant un sistema de manipulació de bases de dades, o *database management system* (DBMS), que consisteix en la col·lecció de dades, i un conjunt de programes de programari per gestionar i accedir a aquestes dades.

Dins del concepte de bases de dades, se sol parlar típicament de dos tipus:

**a) Bases de dades relacionals:** es tracta d'una base de dades que consisteix a organitzar la informació mitjançant una col·lecció de taules, en la que a cadascuna se li assigna un nom únic. Cada taula consisteix en un conjunt d'atributs (columnes o camps) i normalment emmagatzema un gran conjunt de tuples (registres o files). Cada tupla d'una taula relacional representa un objecte identificat per una clau única i descrit per un conjunt de valors d'atribut. En l'actualitat, les bases de dades relacionals més utilitzades són MySQL, Oracle, SQL Server i PostgreSQL.

**Exemple**

Suposem que hi ha una plataforma de cadenes de televisió, en la qual els clients se subscriuen a cadascuna i poden escollir la seva forma de pagament. Per fer-ho, comencem amb una primera taula que conté només la informació de cada client.

| Client_ID | Client_FirstName | Client_LastName |
|---|---|---|
| 1 | Juan | González |
| 2 | Luis | Pérez |
| 3 | José | Hernández |

Després, tenim una taula amb les diferents cadenes de televisió disponibles i una altra amb la forma de pagament.

| TV_ID | TV_Name | TV_Price |
|---|---|---|
| 1 | Informative TV | 20 |
| 2 | Cartoon TV | 30 |
| 3 | Music TV | 15 |

| Payment_ID | Payment_Type |
|---|---|
| 1 | Efectiu |
| 2 | Transferència bancària |
| 3 | Targeta de crèdit |

Finalment, hi ha una taula que relaciona totes les taules anteriors.

| ID | Client_ID | TV_ID | Payment_ID |
|---|---|---|---|
| 1 | 1 | 3 | 3 |
| 2 | 2 | 3 | 2 |
| 3 | 2 | 3 | 2 |
| 4 | 3 | 1 | 1 |

Com es pot observar, amb la informació de la última taula podem extreure qualsevol dada que es requereixi sobre cada client. Per exemple, per al client de nom Luis Pérez, sabem que el seu Client_ID és 2 i podem veure que apareix en dues files, ja que s'ha subscrit a dues cadenes de televisió (ID = 2 i ID = 3) corresponents a Cartoon TV i Music TV. A més, s'observa que ha triat pagar amb transferència bancària en els dos casos (Payment_ID = 2). És important fixar-se que no cal incloure la tarifa a la taula final, ja que si coneixem el TV_ID es pot calcular fàcilment el preu.

**b) Bases de dades no relacionals:** són un tipus de base de dades que s'implementa quan la informació que s'ha d'emmagatzemar és molt complexa per poder ser expressada en una taula. A diferència de les bases de dades relacionals, no cal conèixer *a priori* què és el que es vol emmagatzemar, ja que les bases de dades no relacionals són més flexibles i poden emmagatzemar qualsevol tipus de dada sense importar la seva estructura. En aquestes bases de dades, no es té un identificador que serveixi per relacionar els conjunts de dades ni un esquema exacte del que s'emmagatzemarà, per la qual cosa s'utilitzen dades del tipus JSON. Actualment, les bases de dades no relacionals més utilitzades són MongoDB, Redis i Cassandra.

**Exemple**

Suposem que tenim diverses plataformes que incorporen diferents sensors i càmeres per donar informació sobre el trànsit en un veïnat i quin tipus de vehicle hi circula. Cada cert temps, una estació rep un informe amb els resultats de cada plataforma en fitxers de tipus JSON, ja que no se sap *a priori* quin tipus de sensors té cada plataforma ni com seran les dades que es rebran. Els informes tenen la forma següent:

```
{
"Plataforma_ID":0001,
"Ubicacion":"6.054, 987.0, 69.78",
"Vistas":[
"auto",
"bicicleta",
"desconocido",
"auto",
"auto",
],
},
"Sonido":{
"min":15,
"max":44
}
}
```

```
{
"Plataforma_ID":0002,
"Ubicacion":"8.022, 767.0, 29.87",
"Vistas":[
"bicicleta",
```

```
"bicicleta",
"desconocido",
"desconocido",
],
},
"Viento":{
"min":120,
"max":298
}
}
```

Com podem apreciar, aquestes plataformes ofereixen informes diferents a l'altra i es té molt poca o cap informació sobre les seves estructures, de manera que no és recomanable dissenyar una base de dades relacional per emmagatzemar aquestes dades. En aquest exemple, el millor és utilitzar una base de dades no relacional que emmagatzemi tota la informació tal qual se li presenti. És important destacar que és possible processar aquestes dades posteriorment per transformar-les en una base de dades relacional per a la seva futura anàlisi, si així es desitja. No obstant això, típicament, aquest pas no és necessari.

En alguns casos, quan es disposa de múltiples bases de dades pertanyents a una mateixa organització, se solen implementar magatzems de dades coneguts com a *data warehouses*, que tenen com a objectiu optimitzar les consultes i generar informes a partir d'un resum de les dades provinents de diferents fonts. En aquests magatzems es recopila la informació de totes les bases de dades d'una manera preestablerta, de manera que se'n faciliti les consultes. Cal assegurar la disponibilitat, la consistència i la preservació de les dades al llarg del temps.

En el cas de necessitar un subconjunt de dades provinents del magatzem, amb el propòsit de brindar suport a una àrea específica, es pot implementar un *datamart*. Aquest tipus de base de dades és simplement una capa d'accés al magatzem, que filtra i selecciona les dades necessàries per a l'anàlisi en una àrea específica, facilitant així la presa de decisions.

## 2.3. Preprocessat

És l'etapa en la que es preparen les dades per a la seva anàlisi posterior. S'hi aplica una sèrie de tècniques de manera que els analistes no hagin de preocupar-se per la qualitat ni procedència d'aquestes dades. A continuació, es presenten les tècniques habitualment utilitzades en l'etapa de preprocessat. No obstant això, la seva implementació dependrà del tipus de dades que es vulguin tractar i de l'anàlisi que es vulgui fer posteriorment.

**Lectura recomanada**

**Han, Jiawai**; **Kamber, Micheline**; **Pei, Jian** (2012). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

### 2.3.1. Integració

La integració és la combinació de dades provinents de fonts diferents, amb l'objectiu d'agrupar-les en una estructura única que faciliti les seves anàlisis i permeti realitzar inferències més profundes.

**Exemple**

Si es coneixen els identificadors dins d'una base de dades «A» d'un grup d'usuaris subscrits a un cert servei i també se sap, inequívocament, que el mateix grup d'usuaris està subscrit a altres serveis descrits a la base de dades «B ». Aleshores, podem fusionar aques-

ta informació en una sola taula «C» que ens permeti fer anàlisis més pertinents sobre aquest grup.

### 2.3.2. Selecció

Aquesta tècnica es basa en el filtrat de les dades que es troben dins d'un fitxer o base de dades amb l'objectiu de seleccionar només les dades d'interès. Per fer-ho, s'adopten criteris de cerca que ens permetin discernir, entre el grup de dades, aquelles que són realment necessàries per a l'anàlisi posterior.

**Exemple**

En una base de dades clínica, es vol analitzar els pacients amb risc d'infart. D'aquesta manera, es filtren les dades per mitjà d'una simple recerca dels pacients majors de 30 anys, ja que es considera que els pacients més joves no presenten un risc significatiu pel qual puguin esbiaixar les anàlisis.

### 2.3.3. Reducció de dades

Aquesta tècnica s'utilitza quan es pretenen analitzar grans volums de dades. Consisteix a aplicar certs mètodes sobre un conjunt de dades per obtenir-ne un volum menor o una representació reduïda, mantenint en tot moment la integritat de les dades originals. En altres paraules, les anàlisis realitzades sobre la representació reduïda de les dades han de produir els mateixos resultats (o molt aproximats) que si s'apliquen sobre el conjunt original de dades. A continuació, es presenten tres mètodes típicament utilitzats per reduir la dimensionalitat:

**a) Reducció de dimensionalitat:** són els mètodes que tenen com a objectiu reduir el nombre de variables aleatòries o atributs sota consideració. Entre els més utilitzats hi ha les transformacions *wavelet* o l'anàlisi de components principals (ACP), que transformen o projecten les dades originals en un espai més reduït. També és possible utilitzar el mètode de selecció de subconjunts d'atributs (*attribute subset selection*), amb el qual es detecten i s'eliminen atributs o dimensions irrellevants o redundants. Altres models paramètrics s'utilitzen per estimar les dades, de manera que s'emmagatzemen només els seus paràmetres. Per exemple, models de regressió o logarítmics.

**b) Reducció de quantitat:** consisteix a substituir el volum original de dades per una representació alternativa de menys volum. Entre els models més utilitzats, tenim els models no paramètrics com, per exemple, histogrames, agrupament (*clustering*) i mostreig (*sampling*).

**c) Compressió de dades:** són tècniques que consisteixen a aplicar transformacions a un conjunt de dades per obtenir una representació comprimida de les dades originals. Aquestes tècniques poden dividir-se en tècniques amb pèrdues o sense pèrdues, depenent de si es poden reconstruir les dades originals a

partir de les dades comprimides mantenint la integritat de la informació. Cal tenir en compte que els mètodes de reducció de dimensionalitat o reducció de quantitat es poden considerar com a formes de compressió de dades.

### 2.3.4. Conversió

També coneguda com a transformació, la conversió de dades implica una sèrie de tècniques que tenen com a objectiu convertir el format original de les dades en un format més pla i comprensible, capaç de facilitar l'anàlisi posterior o mineria de dades. Entre les tècniques més utilitzades hi ha:

**a) Suavitzat (*smoothing*):** té com a objectiu eliminar el soroll de les dades. Les tècniques inclouen filtrat, *binning*, regressions i *clustering*.

**b) Construcció d'atributs:** consisteix a afegir nous atributs resultants d'operacions sobre un altre conjunt d'atributs per facilitar el procés de mineria. En altres paraules, es creen nous atributs a partir de càlculs sobre les variables disponibles.

**Exemple**

A l'hora de calcular la ràtio entre dues variables o realitzar conversions en les unitats.

**c) Agregació de dades:** s'apliquen operacions de síntesi o d'addició a un conjunt de dades per permetre múltiples nivells d'abstracció en el procés de mineria de dades i resumir el conjunt d'atributs en un de sol.

**Exemple**

Sumar el nombre de subscripcions diàries a un servei determinat per obtenir les subscripcions mensuals.

**d) Normalització:** consisteix a modificar l'escala dels atributs de manera que es trobin dins d'un rang més petit, típicament, entre -1.0 i 1.0 o entre 0.0 i 1.0.

**e) Discretització:** es basa en substituir els valors numèrics dels atributs per etiquetes, la qual cosa dona com a resultat un nivell més alt d'abstracció. Aquestes etiquetes poden ser en intervals o conceptuals.

**Exemple**

A l'hora de parlar de les edats d'un grup de pacients, en comptes de col·locar dins de l'atribut el valor numèric de l'edat, podem col·locar una etiqueta amb intervals de 0-20, 21-40, 41-60, o col·locar etiquetes conceptuals com: infant, nen, adolescent, adult o ancià. Fixem-nos que aquestes etiquetes poden organitzar-se al seu torn de manera recursiva, la qual cosa dona com a resultat la generació d'una jerarquia de conceptes.

**f) Generació de jerarquia de conceptes per a dades nominals:** s'utilitza, a diferència del cas anterior, quan els atributs no posseeixen valors numèrics, sinó nominals.

**Exemple**

Una adreça es pot substituir per etiquetes conceptuals d'un nivell superior, com: ciutat, regió o país.

### 2.3.5. Neteja

La neteja de dades o *data cleaning* es considera un dels passos més importants del preprocessat de les dades, ja que la qualitat i la veracitat dels resultats dependran, en gran part, del desenvolupament correcte d'aquesta fase. La neteja de dades consisteix a eliminar-ne les inconsistències, ja sigui de manera automàtica a través d'una detecció, o bé manualment després d'una inspecció més detallada. Aquest procés es realitza ja que, normalment, la dada en «brut» ve contaminada amb soroll i inconsistències que no permeten l'obtenció de coneixement amb l'aplicació directa dels mètodes d'anàlisi.

Entre les inconsistències típicament trobades en les dades tenim: la presència de dades perdudes, dades no definides (zero, buit i nul) o l'aparició de valors extrems. Els mètodes aplicats per a la resolució d'aquests problemes dependran de l'origen i del tipus de dades que es vulguin tractar, com veurem a continuació.

**a) Anàlisi de dades perdudes:** la denominació de dades perdudes o *missing data* s'empra quan, per a una variable o observació, no es té cap dada. Aquest és un dels problemes més comuns que tenen lloc en el moment de la verificació de les dades, abans de realitzar-ne la neteja. Poden sorgir pel mal funcionament dels dispositius o processos de captura, errors (oblits) humans o per errors de transmissió de dades. Depenent de la seva naturalesa, hi ha diferents solucions per resoldre aquest inconvenient. Entre elles tenim:

- **Ignorar l'atribut:** s'aplica quan no es té cap referència sobre l'atribut ni la seva procedència. És una tècnica molt poc efectiva ja que, depenent de la seva importància en l'anàlisi, pot causar la pèrdua completa d'un individu o classe, la qual cosa comporta la pèrdua d'altres atributs que sí que podrien estar definits.

- **Compleció manual:** és una tècnica que consisteix a omplir manualment la dada faltant per una dada coneguda *a priori*. També es considera poc efectiva, ja que requereix molt temps de verificació i resulta inaplicable quan es tracta de grans volums de dades.

- **Compleció amb una constant global:** consisteix a substituir els valors perduts d'un atribut per una mateixa constant o etiqueta (per exemple, «Desconegut» o «∞»). Això no obstant, tampoc no es considera una tècnica infal·lible, ja que, si no es prenen les precaucions necessàries en el programa d'anàlisi, aquest pot interpretar aquestes dades com a atributs d'interès per culpa del seu valor en comú.

- **Compleció a partir d'una mesura de tendència central:** se substitueixen les dades que faltaven per un mateix valor, que és el resultat d'una mesura de tendència central (mitjana o mediana), depenent del tipus de distribució de les dades.

- **Compleció amb el valor més probable:** es tracta de la implementació de mètodes probabilistes per determinar o predir el valor que podria adoptar aquest atribut donant un conjunt de dades. Entre les eines més utilitzades, en aquest cas, tenim: regressions, inferències basades en models bayesians o a partir d'arbres de decisió.

**b) Anàlisi de dades no definides:** típicament, quan es tracta de dades no definides, ens referim als zeros, buits i nuls. No obstant això, l'enteniment i la definició d'aquestes dades dependran del context i l'origen de les dades.

Per poder entendre quan s'ha d'utilitzar aquest tipus de dades, cal entendre les diferències entre elles. En primer lloc, els zeros es consideren valors legítims i vàlids per a una mesura numèrica, sempre i quan el zero sigui part del domini (per exemple, resultat final en un partit de futbol: 0 a 1).

Per altra banda, una dada buida representa l'existència d'un atribut del qual es desconeix el valor. En altres paraules, aquest valor es pot trobar. Per exemple, una data que es va oblidar col·locar en un camp d'una taula pot generar un valor buit, ja que no se'n sap el valor, però aquest valor se sap que existeix. Quan es treballa amb cadenes de caràcters o *strings* se solen utilitzar les dades buides com a equivalents a un valor zero, és a dir, vol dir que el valor d'aquest camp per a aquesta instància és «res», la qual cosa és completament legítim.

Finalment, les dades nul·les representen la no existència de valor, ni tan sols de «res». Per exemple, un camp d'una taula que indiqui si un client va contractar cert servei pot col·locar-se com a nul, si aquest client mai no ha contractat cap servei a la companyia (la dada no existeix).

Com es pot apreciar, l'ús d'aquest tipus de dades dependrà del missatge o significat que es vulgui transmetre dins del conjunt de dades. Cal tenir en compte que les dades perdudes es poden substituir per aquest tipus de valors, si convé. A més, durant aquesta etapa de preprocessat, les dades d'aquest tipus es poden substituir entre elles en funció de l'anàlisi que es realitzarà posteriorment. Per exemple, substituir els zeros per dades nul·les per evitar que interfereixin en certes proves estadístiques.

**Bibliografia recomanada**

**Squire, Megan** (2015). «Data Mining». Birmingham: Packt Publishing.

**c) Anàlisi de valors extrems:** es consideren valors extrems aquells valors que estan molt per sobre de la distribució normal per a una variable o població. També es pot entendre com a observacions que es desvien tant de la resta que poden despertar sospites de si van ser generades per un mecanisme diferent.

En general, es consideren valors extrems els que superen ± 3 desviacions estàndard, quan la mostra és prou gran. L'aparició d'aquests valors extrems pot afectar els resultats adversament de diferents maneres:

- Incrementant l'error en la variància i reduint el poder en tests estadístics, alterant així els seus resultats.
- Si aquests valors no són generats de manera aleatòria, poden alterar significativament les probabilitats d'errors de tipus I i tipus II (veure la definició a l'apartat del glossari).
- Poden afectar els resultats sobre correlacions entre variables o regressions.
- Esbiaixen els càlculs i estimacions sobre un conjunt de dades per no pertànyer a la població d'interès.

Aquests valors poden aparèixer per diferents motius i s'han de tractar en funció de cada cas. Per exemple, una possible raó de la seva aparició és per errors en la captura de les dades, com ara errors humans de transcripció (en una enquesta, l'entrevistador s'oblida de posar un punt decimal en el camp, la qual cosa dona com a resultat un atribut amb un valor de 100 en lloc d'1,00). Aquests errors poden solucionar-se fàcilment tornant als registres originals o tornant a realitzar el càlcul.

**Bibliografia recomanada**

**Osborne, Jason W.** (2010, març). «Data cleaning basics: Best practices in dealing with extreme scores». *Newborn and Infant Nursing Reviews* (vol. 10, núm. 1, pàg. 37-43).

Un altre exemple d'aparició d'aquest tipus de valors és per errors o biaix en el mostreig, és a dir, quan els valors són produïts per la captura de dades que no pertanyen a la mateixa població que es vol analitzar. En una anàlisi dels salaris dels treballadors d'una companyia, observem que alguns individus presenten valors extrems; quan els verifiquem, comprenem que corresponen als salaris dels accionistes (propietaris) de la companyia. En aquest cas, les dades no es poden considerar legítimes, ja que no pertanyen a la mateixa població de treballadors. En aquests casos, s'han d'eliminar les dades que no pertanyen a la població en qüestió.

Cal tenir en compte que no tots els valors extrems són il·legítims o s'han d'eliminar, per això és important saber-ne l'origen per poder prendre una decisió sobre quin procediment realitzar al respecte.

### 2.4. Anàlisi

Aquesta etapa, també coneguda com a fase de mineria de dades, té com a objectiu crear diferents models que permetin explicar com són les dades i quines són les seves característiques principals. D'aquesta manera, a partir d'aquestes respostes es pretén respondre a les preguntes que es plantegen en el marc d'un projecte d'on es van extreure aquestes dades.

En funció de la naturalesa de les dades i dels objectius del projecte, es poden aplicar diferents tipus d'anàlisis. Les més utilitzades són les que es presenten a continuació:

**Bibliografia recomanada**

**Minguillón, Julià** (2016). «Fundamentos de Data Science». Editorial UOC.

- **Anàlisi estadística descriptiva**. Consisteix en modelar les dades a partir d'un conjunt reduït de valors, d'acord amb alguna distribució coneguda. L'objectiu és descriure adequadament les característiques intrínseques d'aquest conjunt. En aquest cas, no es té en compte si el conjunt en tractament forma part o no d'un conjunt de dades superior. Per exemple, càlcul de mitjanes, medianes, desviació estàndard, etc.

- **Anàlisi estadística inferencial**. A diferència del cas anterior, en aquest tipus d'anàlisi s'intenta modelar les dades a través d'una distribució desconeguda. La raó és que s'assumeix que el conjunt de dades a analitzar representa només una fracció de la totalitat d'una població. Per conseqüent, en aquest cas l'objectiu és inferir com és la població. Per a això, s'assumeix un grau d'error en les estimacions, ja que no es disposa de totes les dades. Exemples d'aquest tipus d'anàlisi són: contrastos d'hipòtesis, regressions, correlacions, etc.

- **Extracció de característiques**. Se serveix de tots els atributs disponibles dins d'un conjunt de dades per obtenir, a través de diferents operacions, atributs nous que, mantenint la integritat de les dades originals, permetin representar d'una millor manera els aspectes que es volen analitzar dins del projecte. Aquest nou grup d'atributs també es coneix com a característiques, o *features*, en anglès.

**Exemple**

Si es vol determinar el consum elèctric d'una sèrie de circuits electrònics, però només es disposa de dades dels corrents «I» i dels components resistius «R». Aleshores, per mitjà de la Llei d'Ohm, és possible extreure una característica nova crida de potència elèctrica ($P = I^2 \cdot R$) que es troba més relacionada amb el concepte de consum, de manera que es facilita la interpretació i anàlisi d'aquestes dades.

D'altra banda, també existeixen altres mètodes que permeten extreure característiques d'un conjunt de dades a partir d'anàlisis específiques (per exemple, anàlisi de la variància) sobre les dades disponibles. Entre els exemples més comuns d'aquest tipus d'anàlisi tenim: l'anàlisi de components principals (ACP), xarxes neuronals simples i *deep learning*, que, per mitjà d'una combinació de xarxes neuronals, busca extreure un coneixement a partir del conjunt donat de dades.

- **Models supervisats**. Són mètodes que s'apliquen quan es disposa de dades en les quals un o diversos atributs representen l'objectiu del problema que es pretén resoldre. D'aquesta manera, es dissenyen models que busquen predir els valors de noves variables corresponents, d'entrada, a aquests atributs, a partir de les altres variables del conjunt de dades. De manera general, les variables que es busca predir són conegudes com a variables depen-

dents, mentre que les altres són conegudes com a variables independents. Entre els mètodes basats en models supervisats més utilitzats tenim: classificadors, arbres de decisió, xarxes neuronals i models lineals.

- **Models no supervisats**. S'utilitzen quan no es disposa d'una variable o atribut objectiu. Per aquesta raó, els models intenten comparar les dades entre elles amb l'objectiu de trobar diferències o semblances que permetin detectar algun tipus d'estructura interna que pugui formar agrupacions en funció d'un criteri d'avaluació. Aquests models poden ser utilitzats per extreure noves característiques que, al seu torn, puguin ser utilitzades en un altre tipus de models com, per exemple, models supervisats. Exemples de mètodes utilitzats per al disseny d'aquests models són: algoritmes de *clustering*, xarxes neuronals i mapes autoorganitzatius.

### 2.5. Visualització

També coneguda com a fase de representació, la visualització consisteix a aprofitar la capacitat del sistema visual humà per a la detecció de patrons, tendències i extracció d'un coneixement a partir de gràfics, models i qualsevol altra eina que permeti interactuar amb els resultats de la fase d'anàlisi.

Per a una implementació més eficient d'aquesta fase, cal que les representacions de les dades posseeixin una interfície de navegació sobre elles, que permetin operacions com, per exemple, selecció, comparació, agregació, marcatge, etc. D'aquesta manera es facilita l'extracció de coneixement a través de la combinació d'un conjunt de dades, sense sobrecarregar d'informació l'usuari.

Hi ha set tasques bàsiques que permeten un nivell més alt d'abstracció per a la visualització de dades:

- **Panorama general** (*overview*): permetre una visió general de la col·lecció de dades.
- **Acostament** (*zoom*): permetre l'acostament a un punt d'interès.
- **Filtratge** (*filter*): permetre el filtratge dels elements no interessants segons un cert criteri.
- **Detalls a petició** (*details on demand*): obtenir detalls sobre un element o grup de ser necessari.
- **Relacions** (*relate*): permetre la visualització de tendències entre els elements.

**Bibliografia recomanada**

**Shneidermann, Ben** (1996). «The eyes have it: a task by data type taxonomy for information visualization». *Proceedings of the 1996 IEEE Symposium on Visual Languages* (vol. 96, pàg. 336-343).

- **Historial** (*history*): mantenir un historial d'accions per poder revenir a passos anteriors o repetir accions.

- **Extracció** (*extract*): permetre l'extracció d'un subconjunt i els seus detalls corresponents.

D'aquesta manera es combina la capacitat humana amb la potència d'un sistema informàtic que permeti aquest tipus de visualitzacions sobre un conjunt de dades i sobre els resultats obtinguts després de l'anàlisi, per així extreure un veritable coneixement del conjunt de dades original i respondre a les preguntes del projecte en qüestió.

### 2.6. Publicació

Aquesta fase final del cicle de vida de les dades té com a objectiu documentar els resultats obtinguts, de manera que sigui possible la seva reutilització per a la resolució de nous projectes.

Per a una publicació correcta d'aquestes dades, cal utilitzar espais que estiguin optimitzats per a tal fi. Per exemple, els repositoris digitals. La raó d'això és que aquest tipus d'espais garanteixen la disponibilitat d'aquestes dades al llarg del temps i, a més, són capaços d'adaptar el format i l'estructura interna de les dades per a què sigui possible el seu accés a través d'altres eines. Aquesta característica es coneix com preservació i és un aspecte que es busca complir a l'hora d'una publicació de dades.

Una altra característica que cal tenir en compte per a una publicació correcta és la disseminació. Aquesta té com a objectiu assegurar la reutilització de les dades, facilitant la seva recerca, accés i enteniment. Per aconseguir-ho, és important que es disposi de metadades adequades que permetin una descripció correcta de les dades, indicant les seves condicions i limitacions d'ús per, així, permetre a tercers l'enteniment del seu origen i naturalesa.

**Bibliografia recomanada**

**Minguillón, Julià** (2016). «Fundamentos de Data Science». Editorial UOC.

Com s'ha explicat anteriorment, en la fase de captura, idealment, un repositori digital hauria d'incorporar una API que facilités l'accés a les dades. A més, és recomanable que també es pugui seleccionar el tipus de format en el qual es descarreguen les dades, com per exemple *comma-separated values* (CSV), *JavaScript Object Notation* (JSON), *Extensible Markup Language* (XML), etc. D'aquesta manera es garanteix que les dades i els seus resultats puguin ser utilitzats en nous projectes, i que en permetin extreure un valor afegit.

# Resum

En aquest mòdul didàctic s'han revisat els conceptes bàsics de la societat de la informació. En primer lloc, en la introducció, s'han descrit els perfils principals de la ciència de dades (científic de dades, enginyer de dades, arquitecte de dades, analista de dades, estadístic, administrador de base de dades, analista de negoci i líder de ciència de dades).

A l'apartat "Què són les dades?" s'han descrit les diferents classificacions de dades (segons estructura, nivell d'accés i tipus d'informació), i els diferents paràmetres per avaluar la qualitat de la informació en les dades, mostrant exemples reals per entendre les diferents dimensions (exactitud , completesa, consistència, atemporalitat, unicitat i validesa).

Finalment, a l'apartat "Cicle de vida de les dades" s'han descrit les principals fases del cicle de vida, que són captura (que inclou creació i extracció), emmagatzematge, anàlisi preprocessada (que inclou integració, selecció, reducció de dades, conversió i neteja), anàlisi (que pot ser descriptiva, inferencial, d'extracció de característiques, models supervisats i no supervisats), representació i publicació.

# Exercicis d'autoavaluació

**1.** Quines diferències hi ha entre un *data scientist* i un *data engineer*?

**2.** Imagina't que crees una empresa. Posa un exemple de cada un dels tres tipus de dades que la teva empresa podria gestionar. Agafa un exemple dels tres que has esmentat anteriorment i explica cada fase del cicle de vida de les dades.

**3.** Explica amb les teves paraules quan és útil fer *web scraping*. Imagina que tens un negoci, explica quan podria ser útil aplicar *web scraping*.

# Solucionari

**1.** Les diferències entre un *data scientist* i un *data engineer* són:

- Un *data scientist* és una persona capaç de plantejar-se les preguntes adequades a partir d'un conjunt de dades relatiu a un domini, i establir quins mètodes i tècniques són els més adequats per extreure el coneixement necessari per respondre a aquestes preguntes, per posteriorment realitzar aquesta tasca. Aquest perfil està orientat principalment a resoldre el «què?».
- Un *data engineer* és una persona capaç de preparar un conjunt de dades de manera que tingui l'estructura i informació adequades per a la seva posterior anàlisi, així com de dur a la pràctica una solució o prova de concepte i convertir-la en una implementació que pugui utilitzar-se en un entorn productiu real. Aquest perfil està més orientat a resoldre el «com?».

**2.** L'empresa que s'estudia és una empresa que prediu possibles complicacions en les revisions de salut anuals que es realitzen als empleats d'una empresa. Un exemple de cada un dels 3 tipus de dades que aquesta empresa podria gestionar és:

- Dades simples: el pes corporal d'un empleat en quilograms.
- Dades compostes o estructurades: una imatge mèdica com, per exemple, una radiografia de fèmur.
- Dades semiestructurades o no estructurades: el document d'anàlisi de la revisió anual periòdica.

Les etapes del cicle de vida de les dades són les següents (s'agafa com a exemple el tipus de dada simple «pes corporal»):

- **Captura:** una persona es pesa en una bàscula i indica el seu pes.
- **Emmagatzematge:** la persona emmagatzema el valor en una aplicació (app) que controla el seguiment del seu pes.
- **Preprocessat:** fusió (es fusiona amb altres dades ja emmagatzemades a l'aplicació), conversió (es converteix en quilograms, si cal), neteja (es detecten inconsistències, per exemple, pesos que no són raonables, que són *outliers*), agregació (s'agreguen els diferents valors del pes per poder conèixer l'evolució, per exemple, es podria fer una agregació mensual), creació de noves variables (si es disposa de l'altura es podria calcular, per exemple, l'índex de massa corporal).
- **Anàlisi:** descriptiva (descriure la persona que s'està pesant), estadística inferencial (qui està realitzant la mesura), extracció de característiques (càlcul de l'índex de massa corporal), reducció de dimensionalitat (si hi ha diverses mesures es pot realitzar un *principal component analysis* per reduir la dimensionalitat), models supervisats (s'intenta predir quin serà el pes del proper mes a partir de dades obtingudes), models no supervisats (es realitza *clustering* per agrupar les dades), visualització (visualització dels patrons, irregularitats o *outliers*).
- **Visualització:** l'aplicació pot proporcionar una visualització temporal diferent dels resultats i, per què no, també geoespacial.
- **Publicació**: les dades es poden publicar per a què altres persones les puguin analitzar o utilitzar.

**3.** És útil realitzar *web scraping* quan no disposem d'API per accedir a les dades web. Com a exemple d'un negoci, suposem que tinc una botiga que ven sabates i vull fer un seguiment dels preus de la meva competència. Podria anar al lloc web del meu competidor cada dia per comparar el preu de cada sabata amb el meu, però, això requeriria molt temps i no escalaria si vengués milers de sabates o si necessités controlar els canvis de preus amb més freqüència. O potser només vull comprar una sabata quan estigui a la venda. Podria tornar i revisar el lloc web de la sabata cada dia fins que tingui sort, però la sabata que vull podria no estar en oferta durant mesos. Aquests dos processos manuals repetitius podrien substituir-se per una solució automatitzada utilitzant les tècniques de *web scraping*.

# Glossari

**ACP** *m* Vegeu ***principal component analysis***.

**API** *f* Vegeu ***application programming interface***.

***application programming interface*** *f* Vegeu **interfície de programació d'aplicacions**.

**CAPTCHA** *m* Vegeu ***completely automated public turing test to tell computers and humans apart***

***comma-separated values*** *m* Format d'arxiu de text on s'utilitzen comes per separar els camps.
sigla **CSV**

***completely automated public Turing test to tell computers and humans apart*** *m* Test de Turing completament automàtic i públic per diferenciar ordinadors d'humans.
sigla **CAPTCHA**

**CSV** *m* Vegeu ***comma-separated values***.

***data, information, knowledge and wisdom*** Vegeu **dades, informació, coneixement i saviesa**.

***database management system*** *m* Sistema per crear i gestionar bases de dades. Aquests sistemes proporcionen als usuaris i programadors una manera sistemàtica de crear, recuperar, actualitzar i gestionar dades.
sigla **DBMS**

**dades, informació, coneixement i saviesa** Piràmide que permet obtenir valor de les dades fins a arribar a la saviesa.
sigla **DICS**

**DBMS** *m* Vegeu ***database management system***.

**Error tipus I** *m* L'error tipus I, també anomenat error de tipus alfa ($\alpha$) o fals positiu, és l'error que es comet quan l'investigador rebutja la hipòtesi nul·la essent aquesta verdadera en la població.

**Error tipus II** *m* L'error tipus II, també anomenat error de tipus beta ($\beta$) ($\beta$ és la probabilitat que existeixi aquest error) o fals negatiu, es comet quan l'investigador no rebutja la hipòtesi nul·la essent aquesta falsa en la població.

***extensible markup language*** *m* Vegeu **llenguatge de marcatge extensible**.

**HTML** *m* Vegeu ***HyperText Markup Language***.

***HyperText Markup Language*** *m* Vegeu **llenguatge de marques d'hipertext**.

**interfície de programació d'aplicacions** *f* Conjunt de rutines que permeten accedir a funcions d'un determinat programari; a Internet, les API permeten accedir al contingut d'un lloc web.
sigla **API**

***JavaScript object notation*** Format d'arxiu d'estàndard obert que utilitza text llegible per l'ésser humà per transmetre objectes de dades que consisteixen en parells d'atributs-valor i arranjaments de dades. És un format de dades molt comú utilitzat per a la comunicació asíncrona navegador-servidor.
sigla **JSON**

**JSON** Vegeu ***JavaScript object notation***.

**llenguatge de marcatge extensible** *m* Metallenguatge extensible d'etiquetes, desenvolupat pel World Wide Web Consortium (W3C) i adaptat del SGML (*Standard Generalized Markup Language*).
sigla **XML**

**llenguatge de marques d'hipertext** *m* Llenguatge de marcatge utilitzat per a l'elaboració de pàgines web.
sigla **HTML**

**llenguatge estructurat de consultes** *m* Llenguatge d'accés a una base de dades.

sigla **SQL**

**mineria de dades** *f* Procés d'anàlisi per descobrir patrons en conjunts de dades. S'apliquen mètodes d'aprenentatge automàtic, estadístiques, entre d'altres.

**paquet estadístic per a les ciències socials** *m* Paquet estadístic utilitzat per analitzar dades.
sigla **SPSS**

***principal component analysis*** *m* Procediment estadístic que utilitza una transformació ortogonal per convertir un conjunt d'observacions o variables possiblement correlacionades en un conjunt de valors de variables linealment no correlacionades, anomenades components principals.
sigla **ACP**

**recovery time objective** *m* És la durada específica del temps i un nivell de servei dins del qual s'ha de restaurar un procés comercial després d'un desastre (o interrupció) per evitar conseqüències inacceptables associades a una interrupció en la continuïtat del negoci.
sigles **RTO**

**RTO** *m* Vegeu ***recovery time objective***.

**SAS** *m* Vegeu ***statistical analysis system***.

**sistema estadístic d'anàlisi** *m* Llenguatge de programació utilitzat per analitzar dades.
sigla **SAS**

**SPSS** *m* Vegeu ***statistical package for the social sciences***.

**SQL** *m* Vegeu ***structured query language***.

***statistical analysis system*** *m* Vegeu **sistema estadístic d'anlàlisi**.

***statistical package for the social sciences*** *m* Vegeu **paquet estadístic per a les ciències socials**.

***structured query language*** *m* Vegeu **llenguatge estructurat de consultes**.

**URL** Vegeu ***uniform resource locator***.

***uniform resource locator*** Referència a un recurs web que especifica la seva ubicació en una xarxa informàtica i un mecanisme per recuperar-la.
sigles **URL**

**XML** *m* Vegeu ***extensible markup language***.

# Bibliografia

**Fayyad, Usama**; **Piatetsky-Shapiro, Gregory**; **Smyth, Padhraic** (1996, març). *«From data mining to knowledge discovery in databases»*. AI Magazine (vol. 17, núm. 3, pàg. 37-54).

**Han, Jiawai**; **Kamber, Micheline**; **Pei, Jian** (2012). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

**Jarman, Kristin H.** (2013). *The art of data analysis. How to answer almost any question using basic statistics*. Hoboken, NJ: Wiley.

**Minguillón, Julià** (2016). *«Fundamentos de Data Science»*. Editorial UOC.

**Osborne, Jason W.** (2010, març). *«Data cleaning basics: Best practices in dealing with extreme scores»*. Newborn and Infant Nursing Reviews (vol. 10, núm. 1, pàg. 37-43).

**Riquelme, José Cristóbal**; **Ruiz, Roberto**; **Gilbert, Karina** (2006). *«Minería de datos: conceptos y tendencias»*. Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial (vol. 10, núm. 29, pàg. 11-18).

**Shneidermann, Ben** (1996). *«The eyes have it: a task by data type taxonomy for information visualization»*. Proceedings of the 1996 IEEE Symposium on Visual Languages (vol. 96, pàg. 336-343).

**Squire, Megan** (2015). *«Data Mining»*. Birmingham: Packt Publishing.