# CIENCIA DE DATOS

## Un enfoque practico de Tecnologías, Herramientas y Aplicaciones

Luis Joyanes Aguilar

# CAPÍTULO 3
# EL PROCESO CICLO DE VIDA DE LA CIENCIA DE DATOS Y LOS CIENTÍFICOS DE DATOS

## APENDICE I. HABILIDADES TECNOLÓGICAS DEL CIENTÍFICO DE DATOS

1. **Matemáticas y Estadística**. Cálculo, Algebra, Probabilidad.

**2. Programación de computadoras**. Lenguajes de programación: Python, R, SAS, Perl, SQL, Java, Scala, Julia. Librerías de Python (TensorFlow, Keras) y de R.

**3. Negocios y comunicación**. Habilidades de comunicación. Comprensión de empresas/negocios, pensamiento crítico, habilidades de comunicación. *Visualización de datos* (Tableau, PowerBI, Qlik, Excel).

**4. *Machine Learning***. Técnicas de aprendizaje máquina (automático), procesamiento de texto, aprendizaje de refuerzo.

**5. Desarrollo de software**. Ingeniería de software.

Software moderno, Docker, DevOps, Kubernetes, microservicios, contenedores, computación sin servidores. Herramientas y plataformas para desarrollo de software: GitHub, Kaggle, Metodologías Agile, DevOps, Anaconda. Librerías de Python y R; *Plataformas especializadas para Minería de Datos, Machine Learning*, KNIME, DataRobot, RapidMiner.

**6. Base de datos**. SQL, Bases de datos NoSQL, Bases de datos en memoria (SAP HANA).

**7. Big Data/*Cloud***. AWS, Apache Hadoop, Apache Spark, Microsoft Azure, Google Cloud.

**8. *Deep Learning***. Algoritmos DL, NLP, Visión por computadora, KEERAS, PyTurch.

**9. Visualización de datos**. Tableau, Qlik, Microsoft.

**10. Minería DE Datos**. (Metodologías CRISP-DM, KDD, SEMMA, Microsoft TDSP).

**11. Analítica aumentada y analítica cognitiva**.

**12. Inteligencia Artificial.** Aprendizaje automático, aprendizaje profundo, lenguaje de procesamiento natural, redes neuronales.

**13. Aprendizaje Automático**. Aprendizaje supervisado, aprendizaje semisupervisado, aprendizaje no supervisado, aprendizaje reforzado.

**14. Aprendizaje Profundo** (*Deep Learning*). Redes neuronales, inteligencia conversacional, *bots, chatbots*, asistentes virtuales, altavoces inteligentes.

**15. Inteligencia Conversacional**. *chatbots*, asistentes virtuales, altavoces inteligentes.

**16. Objetos autónomos inteligentes**. autos autónomos, drones, ropa inteligente.

# APENDICE II. DESCRIPCIÓN DE ETAPAS DEL CICLO DE VIDA DEL PROCESO DE CIENCIA DE DATOS MICROSOFT

**Fase de descripción de negocio del ciclo de vida del proceso de Ciencia de Datos en equipo:**

<https://docs.microsoft.com/es-es/azure/architecture/data-science-process/lifecycle-business-understanding>.

**Fase de adquisición y comprensión de los datos del proceso de Ciencia de Datos en equipos**

<https://docs.microsoft.com/es-es/azure/architecture/data-science-process/lifecycle-business-understanding>.

**Objetivos**

- Genere un conjunto de datos limpio y de alta calidad cuya relación con las variables de destino se entienda. Busque el conjunto de datos en el entorno de análisis de adecuado para prepararse para el modelado.
- Desarrolle una arquitectura de solución de la canalización de datos que actualice y puntúe los datos con regularidad.

**Fase de modelado del ciclo de vida del proceso de Ciencia de Datos en equipos**

**Objetivo**

Implemente modelos con canalización de datos en un entorno de producción o similar para que el usuario final los acepte.

### Modo de hacerlo

La tarea principal que se aborda en esta fase es la siguiente:

**Uso del modelo**: implemente el modelo y la canalización en un entorno de producción o semejante para el consumo de aplicaciones.

### Uso de modelos

Cuando ya disponga de un conjunto de modelos que funcionan bien, los puede hacer operativos para que los consuman otras aplicaciones. Dependiendo de los requisitos empresariales, se realizan predicciones en tiempo real o por lotes. Para implementar modelos, los expone con una interfaz de API abierta. La interfaz permite que el modelo se utilice fácilmente por diferentes aplicaciones, como las siguientes:

<https://docs.microsoft.com/es-es/azure/architecture/data-science-process/lifecycle-modeling>.

## Fase de implementación del ciclo de vida del proceso de Ciencia de Datos en equipos

### Objetivo

Implementar modelos con canalización de datos en un entorno de producción o similar para que el usuario final los acepte. La tarea principal que se aborda en esta fase es la siguiente:

### Uso de modelos

Cuando ya disponga de un conjunto de modelos que funcionan bien, los puede hacer operativos para que los consuman otras aplicaciones. Dependiendo de los requisitos empresariales, se realizan predicciones en tiempo real o por lotes. Para implementar modelos, los expone con una interfaz de API abierta. La interfaz permite que el modelo se utilice fácilmente por diferentes aplicaciones, como las siguientes:

- Sitios web en línea
- Hojas de cálculo
- Paneles
- Aplicaciones de línea de negocio
- Aplicaciones de *back-end*

<https://docs.microsoft.com/es-es/azure/architecture/data-science-process/lifecycle-deployment>.

## Fase de aceptación del cliente

### Objetivo

*Finalización de los resultados del proyecto*: confirme que la canalización, el modelo y su implementación en un entorno de producción cumplen los objetivos del cliente.

### Modo de hacerlo

En esta fase se abordan dos tareas principales:

- *Validación del sistema:* confirme que el modelo implementado y la canalización cumplen las necesidades del cliente.
- *Entrega del proyecto*: entregue el proyecto a la entidad que va a ejecutar el sistema en producción.

El cliente debe validar que el sistema satisface sus necesidades empresariales y responde a las preguntas con una precisión aceptable para implementarlo en el entorno de producción y usarlo con la aplicación cliente. Se finaliza y revisa toda la documentación. El proyecto se entrega a la entidad responsable de las operaciones. Esta entidad podría ser, por ejemplo, un equipo de Ciencia de Datos de clientes o de TI o un agente del cliente responsable del funcionamiento del sistema en producción.

<https://docs.microsoft.com/es-es/azure/architecture/data-science-process/lifecycle-acceptance>.

# CAPÍTULO 5
# MINERIA DE DATOS

## INTRODUCCIÓN

Minería de Datos es un conjunto de técnicas que se utilizan para optimizar el desarrollo de la Inteligencia de Negocios a partir de los datos que una organización recolecta, organiza y almacena para, posteriormente, realizar el análisis de datos y contribuir del modo más eficiente posible en la toma de decisiones.

Existen grandes volúmenes de datos almacenados en las bases de datos, *data warehouses* y otros tipos de almacenes de datos como las bases de datos NoSQL, "en memoria" y MPP (bases de datos de procesamiento paralelo masivo) y los modernos repositorios de datos conocidos como lagos de datos (*data lakes*). Esta ingente cantidad de datos son cada día más ricos y rentables para las organizaciones, pero requieren un análisis eficiente, no siempre fácil, para su conversión en conocimiento para la toma de decisiones.

La Minería de Datos busca descubrir el conocimiento de los datos recopilados y almacenados, de manera oportuna y en una forma escalable (para pasar de modo eficiente de gigabytes – terabytes – petabytes - exabytes y, en un futuro no lejano, a zettabytes).

Se desarrollan los conceptos fundamentales de Minería de Datos y las técnicas para el descubrimiento de patrones de datos de interés a partir de datos de aplicaciones diversas. Analizaremos los fundamentos de la Minería de Datos, sus aplicaciones y campos de implantación, así como las técnicas más utilizadas en el desarrollo de herramientas de Minería de Datos eficaces y eficientes. Asimismo, realizaremos un examen práctico de las herramientas de software más empleadas, tanto de software propietario y comercial como de código abierto y gratuito.

Las técnicas de Minería de Datos están evolucionando y se están integrando con las técnicas de aprendizaje automático dentro de la Inteligencia Artificial y la Ciencia de Datos, disciplina que tiene como espina dorsal la Minería de Datos y, de un modo más amplio, el Aprendizaje Automático, que se verá a lo largo de los capítulos 10 y 11.

## 5.1 MINERÍA DE DATOS: CONCEPTOS Y DEFINICIONES

El término Minería de Datos (*Data Mining)* se utiliza para describir el descubrimiento o "minado" (*mining*) del conocimiento a partir de grandes cantidades de datos; nació como una analogía de la minería tradicional (minería de oro o de cobre). Por esta razón, como se trataba de minar, buscar conocimiento significativo en ingentes volúmenes de datos, también se la ha denominado a veces como "minería de conocimiento" (*knowledge mining*), "descubrimiento del conocimiento" (*knowledge discovery*), incluso, "extracción del conocimiento".

Asimismo, en muchas ocasiones, dado que la Minería de Datos, en la práctica, hace análisis de datos a través de patrones previamente seleccionados, también se la ha denominado "análisis de datos", "análisis de patrones de datos" y "arqueología de datos" (Turban, 2011; Sharda, 2014).

En este sentido, deseamos destacar un término sinónimo y muy utilizado en los orígenes de la Minería de Datos: KDD (*Knowledge Discovery from Data*) o descubrimiento de conocimiento en bases de datos; sin embargo, como veremos posteriormente, el proceso KDD es un término más global y la Minería de Datos es una etapa dentro del proceso global KDD.

Desde un punto de vista técnico, Minería de Datos es el proceso que utiliza técnicas matemáticas, estadísticas y de Inteligencia Artificial para extraer datos, identificar información útil y conocimientos posteriores en forma de patrones. Estos patrones pueden estar en diferentes formatos: reglas de negocio, afinidades, correlaciones, asociación, tendencias o "modelos de predicciones". Una definición muy aceptada de Minería de Datos es (Fayyad et al., 1996, citado por Turban, 2011):

> *El proceso no trivial de identificación de patrones de datos, almacenados en bases de datos estructuradas y que tiene las propiedades de no triviales, válidos, nuevos (noveles), posteriormente útiles y en última instancia comprensibles, donde los datos se organizan en registros estructurados por variables categóricas, ordinales y continuas.*

El proceso de Minería de Datos implica una serie de etapas iterativas.

La Minería de Datos es una disciplina en la que confluyen muchas otras disciplinas clásicas en el mundo científico y de negocios: estadística, Inteligencia Artificial, Aprendizaje Automático, Aprendizaje de Máquina, sistemas de

información, algoritmos, bases de datos y, recientemente, Ciencia de Datos o Big Data.

El incremento de volumen y variedad de datos ha ido aumentando en las últimas décadas, lo que ha aumentado las necesidades de almacenamiento de datos y el control de la potencia de procesamiento en todo tipo de empresas. Estas necesidades han conducido a las empresas a potenciar sus capacidades de analítica de datos. Por estas razones, se ha hecho necesario "minar" los datos corporativos para descubrir nuevos soportes de conocimiento y mejorar los procesos y prácticas comerciales, esto significa que se han ido consolidando las técnicas de Minería de Datos junto con las nuevas técnicas de minería de textos, minería web o minería de sentimientos.

También en la década del 2000, se ha ido consolidando el desplazamiento a otro paradigma, el medio en el que los datos se capturaban y han emergido numerosas fuentes de datos (la mayoría, datos no estructurados que necesitan herramientas específicos).

La Minería de Datos se utilizó originalmente para describir los procesos a través de los cuales se iban descubriendo los primeros patrones desconocidos de los datos. Desde entonces, la definición se ha ido ampliando más allá de los límites por los proveedores de software para incluir la mayoría de las funcionalidades del análisis de datos, que vimos en el capítulo 4, con el fin de aumentar las ventas con la popularidad de la etiqueta Minería de Datos.

El minado de datos es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.

La Minería de Datos surgió con la intención o el objetivo de ayudar a comprender una enorme cantidad de datos y que estos pudieran ser utilizados para extraer conclusiones que contribuyan a la mejora y el crecimiento de las empresas. Sobre todo, en lo relacionado con las ventas o la fidelización de clientes.

Su principal finalidad es explorar, mediante la utilización de distintas técnicas y tecnologías, bases de datos enormes de manera automática. El objetivo es encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos que se han ido recopilando con el tiempo. Estos patrones pueden encontrarse utilizando estadísticas o algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

El término Minería de Datos se ha ido consolidando y muchas de sus técnicas tienen sus raíces en el análisis estadístico tradicional y también las técnicas de Inteligencia Artificial, en especial las técnicas de algoritmos del Aprendizaje Automático. En su acepción más sencilla, Minería de Datos es un término utilizado para el descubrimiento o "minado del conocimiento de grandes cantidades de

datos". Por eso, en ocasiones se utilizan términos asociados a la Minería de Datos como "extracción del conocimiento" o "análisis de patrones".

**Definiciones de Minería de Datos**

Así la Minería de Datos como refleja Sharda (2020) no es una nueva disciplina, pero sí utiliza muchas otras disciplinas, incluyendo estadística, Inteligencia Artificial, Aprendizaje Automático, sistemas de información, gestión y almacenamiento de bases de datos en sus diversa categorías, visualización de datos, así como depósitos de almacenamiento.

Las empresas se benefician de la Minería de Datos de muchas maneras: anticipando la demanda de productos, determinando las mejores formas de incentivar las compras de los clientes, evaluando el riesgo, protegiendo su negocio del fraude y mejorando sus esfuerzos de marketing.

Una definición muy ajustada a sus objetivos es de la empresa multinacional, experta en la disciplina, SAS[1]:

> *La Minería de Datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Empleando una amplia variedad de técnicas, se puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más.*

Gartner define Minería de Datos en su *IT Glossary*[2] como:

> *El proceso de descubrir correlaciones, patrones y tendencias significativos al filtrar grandes cantidades de datos almacenados en repositorios. La Minería de Datos emplea tecnologías de reconocimiento de patrones, así como técnicas estadísticas y matemáticas.*

Como síntesis, una definición muy extendida de Minería de Datos es:

> *El proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes volúmenes de datos. En esencia, la Minería de Datos busca conocimiento (patrones de interés) en sus datos.*

[1] SAS. *Data Mining: What it is and why it matters*. Disponible en: <https://www.sas.com/en_us/insights/analytics/data-mining.html>.

[2] *Definition of Data Mining. Gartner Information Technology Glossary*. Disponible en: < https://www.gartner.com/en/information-technology/glossary/data-mining>.

## 5.2 APLICACIONES DE LA MINERÍA DE DATOS

La Minería de Datos es una de las herramientas más utilizadas en la actualidad en un gran número de áreas, que va desde aplicaciones empresariales, de negocios, industriales hasta la investigación científica en numerosos campos como la medicina, la bioquímica, meteorología, astronomía, genética, bioinformática, física. Un área muy destacada de aplicaciones de Minería de Datos es la web y los textos que han dado lugar al nacimiento de minerías de datos específicas, conocidas como *Minería Web* y *Minería de Textos.* Es muy difícil, por no decir imposible, enumerar todas las aplicaciones en las que la Minería de Datos juega un papel fundamental. En esta sección tratamos de describir aplicaciones sobresalientes en diversos y variados campos.

**Medicina**

La Minería de Datos en medicina es una de las aplicaciones más prácticas, debido a que complementa la investigación médica en análisis clínicos y en el trascendental campo de los diagnósticos, entre otras especialidades:

- Identificación de patrones novedosos para mejorar la supervivencia de pacientes con cáncer.
- Predicción de tasas de éxito en trasplantes de órganos a pacientes para desarrollar políticas de donantes/receptores en el tratamiento clínico.
- Genómica. Identificación de los diferentes genes del cromosoma humano.
- Selección de embriones en reproducción artificial.
- Genética. Estudio del ADN.
- Descubrimiento de las relaciones entre síntomas y enfermedades, así como entre enfermedades y tratamientos con éxito.
- Identificación de terapias para diferentes enfermedades.
- Estudio de factores de riesgo en diferentes patologías.
- Segmentación de pacientes por grupos afines.
- Gestión hospitalaria y clínica para planificación temporal de habitaciones, quirófanos, salas de consulta, etcétera.

**Industria**

En el sector industrial son numerosas las aplicaciones y, también, en numerosos campos.

- Fabricación y producción. Predecir fallos de máquinas antes de que ocurran a través del uso de datos de sensores.

- Procesos industriales. Automatización del control de procesos: optimización del rendimiento de forma adaptativa; implementación de programas de mantenimiento predictivo.
- Control de calidad. Identificación de posibles causas de fallos; detección y localización precoz de defectos industriales; detección precisa de productos defectuosos; descubrimiento de patrones noveles para identificar y mejorar la calidad de los productos.

**Banca**

La Minería de Datos ayuda al sector bancario en numerosas aplicaciones:

- Detección de patrones de uso fraudulento de tarjetas y transacciones de banca en línea.
- Automatización de los procesos de concesión de préstamos para predecir, con la mayor precisión posible, los morosos más probables.
- Estudio de concesión de tarjetas de crédito.
- Determinación del gasto en tarjetas por segmentación de grupos.
- Identificación de reglas de comportamiento del mercado de valores a partir de los registros históricos de dichos mercados.
- Predicciones de hábitos y patrones de compra en grandes almacenes y en mercados en línea.
- Detección de segmentos de clientes predispuestos a la compra de determinados artículos, bien en el lanzamiento o cuando ya están en el mercado.
- Identificación de clientes fieles y también de fuga de clientes.

**Seguros**

La industria de seguros también utiliza técnicas de Minería de Datos en numerosas aplicaciones:

- Predicción de clientes propensos a suscribir nuevas pólizas con características especiales.
- Identificación y prevención de pagos de reclamaciones incorrectas y actividades y comportamientos fraudulentos.
- Identificación de grupos de clientes y patrones de riesgo.
- Previsión de gastos médicos de sectores de clientes específicos.
- Identificación de fidelidad de clientes.
- Identificación de fugas de clientes.

### Hardware y software de computadoras

Un equipo de computación con un hardware adecuado y un software eficaz aumentará el desempeño del proceso de buscar y analizar información; la Minería de Datos puede ser muy provechosa en el caso de poseer datos sobre sus procesos productivos, datos de seguimiento de clientes, datos externos del mercado, datos sobre la actividad de competidores, etcétera. Aplicaciones concretas pueden ser:

- Predicción de fallos de unidades de disco o memorias antes de que ocurran realmente.
- Identificación y filtrado de contenido web no deseado y mensajes de correo electrónico.
- Detección y prevención de puentes de seguridad de redes de computadoras.
- Identificación de productos de software potencialmente inseguros.

### Turismo

En la industria del turismo existe una gran variedad de aplicaciones para hoteles, líneas aéreas, viajes, alquiler de automóviles, trenes, etcétera:

- Predicciones de ventas de diferentes servicios (reserva de asientos en diferentes clases, reserva de habitaciones en hoteles, reserva de autos en compañías de alquiler, etcétera).
- Identificación de los clientes más rentables para proporcionarles mejores servicios (por ejemplo, las tarjetas de fidelización, "millas" de los clientes viajeros frecuentes, a los que se ofrecen beneficios como prioridad en salas VIP, subida de categoría, ofertas especiales en función de la tarjeta de fidelización, etcétera).
- Predicción de ocupación en aviones, trenes, dependiendo de rutas viajeras, épocas del año, entre otras.

### Gestión de relaciones con los clientes (CRM)

La gestión y administración de relaciones con los clientes es una de las actividades de las empresas de más aplicaciones de la Minería de Datos. El objetivo del CRM es crear relaciones personales con los clientes para facilitar los servicios que se han de proporcionar. El CRM tiene dos objetivos principales:

- Retención de clientes mediante la satisfacción del cliente.
- Desarrollo de relaciones con los clientes mediante el conocimiento de estos.

Las técnicas de Minería de Datos tienen una gran aplicación, sobre todo en la segmentación de clientes, campañas de marketing directo, campañas de marketing de contenidos, etcétera.

Así, aplicaciones globales de Minería de Datos clásicas son:

- Segmentación de clientes.
- Campañas de marketing directo.
- Análisis de secuencia y la bolsa de la compra.

También existen aplicaciones más específicas:

- Identificar compradores/demandantes de nuevos servicios/productos.
- Comprender las razones para mejorar la retención de clientes.
- Descubrir relaciones en el tiempo entre productos y servicios para maximizar ventas y valores del cliente.
- Identificar los clientes más rentables y sus preferencias para fortalecer las relaciones con ellos y maximizar las ventas.

**Deportes**

La Minería de Datos también tiene, curiosamente, numerosas y prácticas aplicaciones en el mundo del deporte. En casi todos los deportes se encuentran aplicaciones de minería. Una de las más conocidas en baloncesto, en la NBA de los Estados Unidos, que desarrolló una aplicación de Minería de Datos para PC (Advanced Scout) que permite descubrir patrones de comportamiento de interés para el desarrollo de los partidos.

Otro caso muy conocido es el equipo AC de Milán que —desde hace bastantes años— utiliza un sistema inteligente para prevenir lesiones. El club posee aplicaciones de redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta, de modo que pueda ayudar a seleccionar el fichaje de un posible jugador o a alertar al médico del equipo de una posible lesión. El sistema de Minería de Datos fue creado por la compañía Computer Associates International y se alimenta de datos de cada jugador relacionados con su rendimiento, alimentación, respuesta a estímulos externos que se obtienen y analizan cada quince días. El sistema dispone de más de 800 casos registrados, que permiten predecir alguna posible lesión.

## 5.3 PROCESO DE MINERÍA DE DATOS

La Minería de Datos, como señalaban en los primeros años de despliegue de Ciencia de Datos, Provost y Fawcett (2013) es un arte que implica la aplicación de una cantidad sustancial de ciencia y tecnología, pero la aplicación propia en sí también es un arte. Existen diferentes procesos de minería de datos adoptados por organizaciones y empresas. Entre ellos destacan los tres más tradicionales: KDD, CRISP-DM y SEMMA.

El proceso de Minería de Datos más aceptado universalmente sigue el proceso experimentado y probado en la industria como CRISP-DM y que fue descrito en el capítulo 3 ya que se ha convertido también, en el proceso más referenciado en Ciencia de Datos.

Otra metodología que tiene también gran aplicación en el proceso de Minería de Datos es KDD (*Knowledge Discovery in Databases*, KDD), que fue creada por Fayyad *et al*. en 1996. Esta metodología se basa en la extracción de conocimiento principalmente relacionada con el proceso de descubrimiento, conocido como descubrimiento de conocimiento en bases de datos y que se refiere al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. No es un proceso automático, es iterativo que, exhaustivamente, explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad y que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos.

## 5.4 TÉCNICAS Y ALGORITMOS DE MINERÍA DE DATOS

Las técnicas de Minería de Datos se clasifican en dos grandes categorías (también denominadas modelos o métodos, tareas) en función de su propósito general: descriptivas y predictivas.

- **Modelos descriptivos**

Describen el comportamiento de los datos (conjunto de datos) de forma que sean interpretables por un usuario experto. Encuentra valores interpretables que describen los datos. Trata de proporcionar información entre las relaciones de los datos y sus características. En el modelo descriptivo se dispone de una variable (denominada respuesta) con valor desconocido y el objetivo fundamental del modelo es determinar ese valor.

- **Modelos predictivos**

  Describe los datos y se utilizan para predecir el valor de algún atributo desconocido, es decir, se utilizan algunas variables para predecir valores desconocidos de otras variables. Están orientadas a estimar valores de salida. Encuentran patrones interpretables que describen los datos.

  Un modelo predictivo intenta predecir o responder a preguntas futuras sobre la base de un estudio de comportamiento pasado. Preguntas que responden a este tipo de modelo de datos (o minería):

  ¿Cómo se venderá el próximo año el producto x?

  ¿Cuántos tipos de personas comprarán el producto?

  ¿Qué riesgo tiene una cierta persona de contraer una enfermedad determinada x sobre la base de sus características personales?

¿Qué clientes son más propensos a darse de baja de nuestra empresa?

El modelo predictivo requiere ser entrenado, utilizando un conjunto de datos cuyo valor de variable objetivo es conocido. En esencia, el modelo entrega resultados sobre la base de un aprendizaje que se va ajustando a la realidad conocida.

**Técnicas de Minería de Datos**

Las técnicas de Minería de Datos se clasifican de acuerdo con los modelos anteriores en predictivas (supervisadas) y descriptivas (no supervisadas)

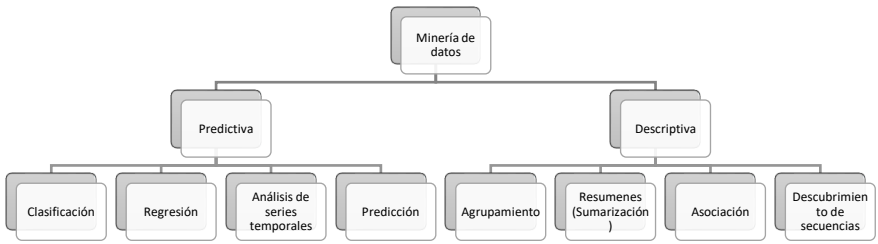



**Figura 5.1. Técnicas de Minería de Datos**

Las **técnicas predictivas** más empleadas son:

- Clasificación
- Regresión
- Series temporales
- Detección de desviaciones/anomalías (predicciones)

Las **técnicas descriptivas** más utilizadas son los algoritmos de:

- Asociación (reglas de asociación)
- Agrupamiento o segmentación (*clustering*)
- S*umarización (*resúmenes)
- Descubrimiento de secuencias de patrones

## 5.4.1 ALGORITMOS DE APRENDIZAJE

Los algoritmos de aprendizaje de los métodos de Minería de Datos se clasifican en función del modo en que los patrones son extraídos de los datos históricos, en las dos categorías ya mencionadas: *supervisados* y *no supervisados*.

> *En los algoritmos de aprendizaje supervisados, el entrenamiento de los datos incluye tanto los atributos descriptivos (por ejemplo: variables de salida o variables de decisión) y los atributos de clase (por ejemplo: variable de salida o variable de resultado); por el contrario, en el aprendizaje no supervisado, el entrenamiento incluye solo los atributos descriptivos (Sharda, 2020: 205)*

### Aprendizaje supervisado

Aprende, a partir de un conjunto de instancias pre-etiquetadas un método para predecir (ejemplo, clasificación: la clase a que pertenece una nueva instancia). El aprendizaje supervisado supone que partimos de un conjunto de datos etiquetado previamente, es decir, conocemos el valor del atributo objetivo para el conjunto de datos que disponemos.

Los algoritmos de aprendizaje supervisado basan su aprendizaje en un juego de datos de entrenamiento previamente etiquetados. Por etiquetado entendemos que para cada vez que ocurra el juego de datos de entrenamiento conocemos el valor de su atributo objetivo. Esto le permitirá al algoritmo poder "aprender" una función capaz de predecir el atributo objetivo para un juego de datos nuevo.

### Aprendizaje no supervisado

No hay conocimiento a priori sobre el problema, no hay instancias etiquetadas, no hay supervisión sobre el procedimiento. El aprendizaje no supervisado parte de datos no etiquetados previamente, por ejemplo, *clustering*, que encuentra un agrupamiento de instancias "natural" dado un conjunto de instancias no etiquetadas.

Los métodos no supervisados son algoritmos que basan su proceso de entrenamiento en un juego de datos sin etiquetas o clases previamente definidas. Es decir, a priori no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico. El aprendizaje no supervisado está dedicado a las tareas de agrupamiento, también llamadas *clustering* o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos.

La implementación de estos modelos requiere del uso de algoritmos de aprendizaje supervisado y aprendizaje no supervisado, que en la actualidad se centran en técnicas de Aprendizaje Automático de Inteligencia Artificial. La tendencia de mayor impacto en la actualidad en dicho ámbito es el Aprendizaje Profundo (*Deep learning*), una disciplina inmersa en el Aprendizaje Automático.

Los modelos predictivos se apoyan en modelos de aprendizaje supervisado, como aplicaciones para la detección de aquellos clientes de una empresa (por ejemplo, operadora telefónica) más propensos a abandonar la empresa como clientes. Los modelos descriptivos se apoyan en modelos de aprendizaje no supervisado, donde se evalúan de acuerdo con las observaciones o datos entregados, recurriendo, normalmente, a argumentos heurísticos que permiten evaluar la calidad de los resultados.

## 5.5 RELACIONES DE LA MINERÍA DE DATOS CON OTRAS DISCIPLINAS: BIG DATA, APRENDIZAJE AUTOMÁTICO Y CIENCIA DE DATOS

La Minería de Datos tiene un enfoque integrador de múltiples disciplinas (multidisciplinar) debido a que utiliza diferentes técnicas según el tipo de información que se ha de extraer. En este sentido, se consideran disciplinas y tecnologías los métodos analíticos, estadísticos, enfoques gráficos, visualización, algoritmos, reconocimiento de patrones, bases de datos, Aprendizaje de Máquina o Aprendizaje Automático y la Inteligencia Artificial.

La Minería de Datos no es una disciplina nueva, ya que se apoya en el proceso KDD que, como ya se ha comentado, tiene sus orígenes en 1996, pero en estos últimos años ha evolucionado considerablemente al incluir numerosas nuevas tendencias. Así, la Minería de Datos es una fusión de múltiples disciplinas, tales como:

- Estadística
- Bases de datos y sistemas de información
- Aprendizaje Automático (*Machine Learning*)
- Aprendizaje Profundo (*Deep Learning*)
- Visualización de datos
- Computación paralela y distribuida
- Interfaces de lenguaje natural con bases de datos
- Sistemas de toma de decisiones
- Inteligencia Artificial.
- Tecnologías de bases de datos y almacenamiento de datos
- Reconocimiento de patrones
- Modelos matemáticos
- Algoritmia y programación
- Otras disciplinas
- Modelos matemáticos

En síntesis, en la figura se señala la relación entre la Minería de Datos con otras disciplinas.


**Figura 5.2. Disciplinas relacionadas con la Minería de Datos**

## 5.6 MINERÍA DE DATOS VERSUS APRENDIZAJE AUTOMÁTICO

La Minería de Datos y el Aprendizaje Automático han ido evolucionando en la década pasada y en la actual; en muchos casos, se suelen considerar términos comunes y se utilizan indistintamente. Sin embargo, aunque tienen un gran número de características, funcionalidades y sus objetivos finales coinciden, existen diferencias y beneficios para cada disciplina. Pero es frecuente considerar la Minería de Datos como un subconjunto de la Analítica de Negocios y la Inteligencia Artificial, pero a medida que numerosas aplicaciones de Minería de Datos han ido incorporando técnicas y aplicaciones de Aprendizaje Automático, la confluencia de ambas disciplinas se va haciendo una realidad. De todos modos, consideramos que la Minería de Datos es una disciplina consolidada y con grandes aplicaciones, como se ha comentado anteriormente, en la Inteligencia de Negocios y su componente fundamental de analítica de negocios.

Para poder ver las diferencias entre ambas disciplinas y poder utilizarlas de modo integrado hemos recurrido a las definiciones de la consultora Gartner y algunos trabajos de reconocido prestigio como los del experto Bernard Marr, para tratar en publicaciones muy reconocidas.

**Definiciones de la consultora Gartner[3]**

La Minería de Datos es el proceso de descubrir correlaciones, patrones y tendencias significativos al filtrar grandes cantidades de datos almacenados en repositorios. La Minería de Datos emplea tecnologías de reconocimiento de patrones, así como técnicas estadísticas y matemáticas.

Los algoritmos avanzados de Aprendizaje Automático se componen de muchas tecnologías (como el Aprendizaje Profundo, las redes neuronales y el procesamiento del lenguaje natural), que se utilizan en el aprendizaje supervisado y no supervisado, que funcionan guiados por las lecciones de la información existente.

**Diferencias entre Minería de Datos y Aprendizaje Automático (Bernard Marr)**

Bernard Marr[4], uno de los investigadores más reconocidos en el campo de los datos, en sus diferentes categorías, autor con blog propio y cuyas publicaciones de mayor impacto se publican en la prestigiosa revista *Forbes*, publicó el artículo: What Is The Difference Between Data Mining And Machine Learning?

La Minería de Datos es un subconjunto del análisis empresarial y se refiere a la exploración de un gran conjunto de datos existente para descubrir patrones, relaciones y anomalías previamente desconocidos que están presentes en los datos. Nos da la capacidad de encontrar ideas completamente nuevas que no necesariamente estábamos buscando, incógnitas desconocidas.

El aprendizaje automático es un subconjunto de la Inteligencia Artificial (IA). Con el aprendizaje automático, las computadoras analizan grandes conjuntos de datos y luego "aprenden" patrones que las ayudarán a hacer predicciones sobre nuevos conjuntos de datos. Aparte de la programación inicial y tal vez algunos ajustes, la computadora no necesita interacción humana para aprender de los datos.

Así las diferencias clave según Marr son:

- Mientras que la Minería de Datos simplemente busca patrones que ya existen en los datos, el aprendizaje automático va más allá de lo que sucedió en el pasado para predecir resultados futuros en función de los datos preexistentes".
- En la Minería de Datos, las "reglas" o patrones se desconocen al inicio del proceso. Mientras que, con el aprendizaje automático, la máquina generalmente recibe algunas reglas o variables para comprender los datos y aprender.

[3] Disponible en: <https://www.gartner.com/en/information-technology/glossary/machine-learning>.

[4] Bernard Marr. *What Is The Difference Between Data Mining And Machine Learning?* Disponible en: <https://bernardmarr.com/what-is-the-difference-between-data-mining-and-machine-learning/>.

- La Minería de Datos es un proceso más manual que depende de la intervención humana y la toma de decisiones. Pero, con el aprendizaje automático, una vez que se establecen las reglas iniciales, el proceso de extracción de información y "aprendizaje" y refinamiento es automático y se lleva a cabo sin intervención humana. En otras palabras, la máquina se vuelve más inteligente por sí misma.
- La Minería de Datos se utiliza en un conjunto de datos existente (como un almacén de datos) para encontrar patrones. El aprendizaje automático, por otro lado, se entrena en un conjunto de datos de "entrenamiento", que le enseña a la computadora cómo dar sentido a los datos y luego hacer predicciones sobre nuevos conjuntos de datos.

## KDNUGGETS[5]: DIFERENCIAS ENTRE DATA MINING Y MACHINE LEARNING

En el prestigioso portal KDnuggets, especializado en Minería de Datos, Inteligencia Artificial y Ciencia de Datos, Sudeep Srivastava publicó en junio de 2022 el estudio "How is Data Mining Different from Machine Learning?"; y en el apartado "Data Mining Vs. Machine Learning" se hace una síntesis de sus conclusiones más relevantes.

La Minería de Datos utiliza dos componentes (bases de datos y aprendizaje automático) para gestión de datos y técnicas de análisis de datos. Ayuda a extraer datos valiosos que pueden proporcionar conocimientos excelentes de un producto o servicio. Sin embargo, el Aprendizaje Automático solo utiliza algoritmos y posee una capacidad de autoaprendizaje para cambiar las reglas por escenario para encontrar la solución.

Otra diferencia de contraste reside en el efecto humano, ya que la Minería de Datos requiere intervención humana continua, pero el Aprendizaje Automático solo requiere a los humanos para definir el algoritmo. Es un proceso automatizado que trabaja para producir resultados más precisos comparados con los resultados de Minería de Datos.

La Minería de Datos está limitada en el modo en que se organizan y recolectan los datos y actúa como un medio para extraer conocimientos relevantes de conjuntos de datos complejos. Por su parte, el Aprendizaje Automático identifica las correlaciones entre todos los puntos de datos relevantes para entregar conclusiones precisas; en última instancia, configurar el comportamiento del modelo.

Por ejemplo, los sistemas CRM implementan procedimientos de Aprendizaje Automático para mejorar su inteligencia de relaciones para comprender mejor a

[5] Sudeep Srivastava (2022). *How is Data Mining Different from Machine Learning?* Disponible en: <https://www.kdnuggets.com/2022/06/data-mining-different-machine-learning.html>.

los clientes. Puede analizar acciones pasadas para impulsar las conversiones y mejorar los puntajes de satisfacción del cliente. El autor resalta en el artículo que incluye la siguiente tabla para ayudar a diferenciar mejor entre minería de datos y aprendizaje automático.

| *Característica* | *Minería de datos* | *Aprendizaje automático* |
|---|---|---|
| *Origen* | *Se originó a partir de bases de datos tradicionales que contienen datos no estructurados* | *Se originó a partir de datos y algoritmos existentes* |
| *Significado* | *Ayuda a extraer información de un conjunto de datos grande* | *Introduce un nuevo algoritmos a partir de datos* |
| *Responsabilidad* | *Se utiliza para obtener las reglas de datos existentes.* | *Enseña al computador a aprender y comprender las reglas dadas* |
| *Naturaleza* | *Es más, un proceso manual y requiere interferencia humana* | *Es un proceso automatizado y no requiere esfuerzo humano un vez que se ha implementado el diseño* |
| *Implementación* | *Los usuarios pueden desarrollar modelos para utilizar las técnicas de minería de datos* | *El usuario puede utilizarlo en árboles de decisión, redes neuronales y áreas específicas de inteligencia artificial* |
| *Ámbito* | *Se puede aplicar en un área limitada* | *Se puede utilizar en áreas muy extensas* |

El artículo publicado en KDnuggets concluye que:

> *Los procedimientos de Minería de Datos ayudan a predecir el resultado a partir de datos históricos o a encontrar una nueva solución a partir de los datos existentes y el aprendizaje automático supera los problemas de la Minería de Datos, lo que lo está ayudando a crecer de una manera mucho más rápida. Además, el aprendizaje automático es más preciso y menos propenso a errores, lo que lo hace capaz de tomar sus propias decisiones y resolver problemas.*

## 5.7 HERRAMIENTAS DE SOFTWARE DE MINERÍA DE DATOS

Muchos proveedores comerciales de software desarrollan y comercializan herramientas de Minería de Datos. Existe un gran número de proveedores de soluciones de software especializados en Minería de Datos como IBM, SAS, Megaputer y SAP. De igual forma, los proveedores de soluciones de software de Inteligencia de Negocios ofrecen soluciones específicas o integradas en sus

herramientas como IBM Cognos, Oracle Hyperion, SAP Business Object, Microsoft, Microstrategy, Teradata, Qlik, Tableau. Además de estas herramientas de software propietario, han emergido numerosas herramientas de software libre y gratuito y de código abierto (*open source*) como WEKA, RapidMiner, Orange, Keel, KNIME, que ofrecen sus soluciones de software abierto y soporte para usuarios y desarrolladores. Asimismo, plataformas de software abierto de Inteligencia de Negocio ofrecen soluciones de código abierto e infraestructuras para soportar los desarrollos de Minería de Datos. Las plataformas más populares son Pentaho, Jaspersoft y Birst.

### 5.7.1 HERRAMIENTAS DE CÓDIGO ABIERTO

Existe un gran número de soluciones de código abierto, y también gratuitas, que pueden ser utilizadas en ámbitos académicos y de investigación, así como en ámbitos profesionales y empresariales.

#### WEKA

Es una de las herramientas gratuitas y de código abierto más populares. Fue desarrollada por un grupo de investigadores de la Universidad de Waikato de Nueva Zelanda. La herramienta se puede descargar gratuitamente del sitio (cs.waikato.ac.nz/ml/weka). Weka es un software de código abierto bajo la licencia GNU GPL (General Public License), escrito en el lenguaje de programación Java, y es una colección de algoritmos de Aprendizaje Automático para tareas de Minería de Datos. Los algoritmos se pueden aplicar directamente a un conjunto de datos o bien llamados desde su propio código Java. Weka contiene herramientas para preprocesamiento de datos, clasificación, regresión, agrupamiento (clustering) y visualización. Weka proporciona acceso a bases de datos SQL, utilizando conectividad de bases de datos en Java y puede procesar los resultados devueltos como consultas de bases de datos.

Existen dos versiones de Weka[6], según el sitio web oficial: Weka 3.8 es la última versión estable y Weka 3.9 es la versión de desarrollo. Nuevas actualizaciones (releases) de estas dos versiones se hacen normalmente una o dos veces al año. Todas ellas corren en los sistemas operativos Windows, Mac OS X y Linux.

#### KNIME (Konstanz Information Miner).

Es una plataforma de Minería de Datos (www.knime.org) que permite el desarrollo de modelos en un entorno visual. Está desarrollada sobre la plataforma Eclipse y programada, fundamentalmente, en Java. Es una plataforma de análisis de datos de código abierto, de fácil uso y comprensible, para integración de datos, procesamiento, análisis y exploración. A través de *plugins*, los usuarios pueden

[6] Las últimas actualizaciones son: 3.8.3 versión más estable y 3.9.6, la versión para desarrolladores. Se define como Weka 3; *Data Mining software in Java.* Disponible en: <www.cs.waikato.ac.nz/ml/weka/index.html>.

añadir módulos de texto, imagen, procesamiento de series temporales y la integración de varios proyectos de código abierto como R y WEKA.

Se desarrolló originalmente en el Departamento de Bioinformática y Minería de Datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com Gmbh es la propietaria de KNIME y está radicada en Zurich, Suiza, donde realiza labores de consultoría y formación (knime.org).

**RapidMiner (rapid-i.com)**

Es una herramienta de software comercial desarrollada por Rapid-I, que se puede descargar del sitio (rapid-i.com) y que fue creada como una herramienta de código abierto. Es una herramienta especializada en Minería de Datos y Aprendizaje Automático. Dispone de una interfaz gráfica de usuario muy mejorada, que utiliza un gran número de algoritmos e incorpora un conjunto de características muy potentes de visualización de datos. Se utiliza tanto en el mundo profesional como en investigación. Ha evolucionado considerablemente desde su creación y, en la actualidad, se ofrece también como una plataforma unificada de Ciencia de Datos.

El sitio web de RapidMiner ofrece una gran cantidad de recursos, guía de referencia de algoritmos, un blog de Ciencia de Datos, reportes, casos de estudio y un buen catálogo de libros electrónicos y artículos. RapidMiner ofrece diferentes versiones, gratuitas (con comunidad de soporte) y de empresa. La versión RapidMiner Studio es la más utilizada para predicciones, validaciones y prototipados. Dispone de un programa con licencias gratuitas para uso académico, tanto para estudiantes como para profesores e investigadores.

**KEEL (keel.es)**

KEEL (*Knowledge Extraction for Evolutionary Learning)* es un software de código abierto de Aprendizaje Automático, desarrollado e implementado en Java, en la Universidad de Granada, distribuido bajo la licencia GPLv3. Está potenciado por una GUI bien organizada que permite administrar (importar, exportar, editar y visualizar) datos en diferentes formatos de archivos y experimentar con ellos (a través de sus sistemas de preprocesamiento de datos, bibliotecas estadísticas y algoritmos de Minería de Datos y Aprendizaje Evolutivo). Dado que KEEL está basado en Java, el entorno JVM ha de estar instalado en el sistema para ejecutar sus propios GUI y realizar proyectos de Minería de Datos. En el sitio (keel.es) se puede consultar la larga lista de algoritmos soportados. KEEL es ideal para fines académicos y de investigación; al ser un producto desarrollado en un grupo de investigación de la Universidad de Granada, es ideal para soporte y apoyo a profesores de materias como Minería de Datos, Ciencia de Datos o Inteligencia de Negocios en general.

**Orange (orange.biolab.si).**

Es una solución integrada (*suite*) para Minería de Datos y Aprendizaje Automático, desarrollada por la Facultad de Informática de la Universidad de Ljubljana (Eslovenia). Es una herramienta que permite programación visual para el análisis exploratorio de datos y visualización, y está integrada con la biblioteca de

algoritmos y programas de Python. Es una aplicación multiplataforma de software libre y de código abierto, que se distribuye bajo licencia GPL.

**Python**

En la actualidad es el lenguaje de programación líder para aplicaciones de Minería de Datos y de Aprendizaje Automático. Su gran número de bibliotecas de funciones y aplicaciones, y su facilidad para el aprendizaje lo hacen ideal para el desarrollo de todo tipo de aplicaciones, en particular, para todas las relacionadas con Minería de Datos.

**R**

Es un lenguaje de programación y entorno de software de código abierto y gratuito para computación, estadística y gráficos. Proporciona múltiples técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásicas, clasificación, agrupación en clústeres. Compila y ejecuta una amplia variedad de sistemas operativos (Unix, Windows y Mac OS). El lenguaje R se utiliza ampliamente entre mineros de datos para desarrollo de software estadístico y análisis de datos.

**NLTK**

Es un *pool* (grupo) de herramientas de procesamiento de lenguajes, incluyendo Minería de Datos, Aprendizaje Automático, *data scraping* (raspado de datos), análisis de sentimientos y otras tareas diversas de procesamiento de lenguajes. Está escrito en Python y está disponible en Windows, Mac OS X y Linux. Es un proyecto de código abierto y está controlado por una comunidad.

## 5.7.2 HERRAMIENTAS DE SOFTWARE PROPIETARIO

La mayoría de los proveedores comerciales de software de soluciones de Inteligencia de Negocios disponen de herramientas propias de Minería de Datos o integradas en *suites* (paquetes integrados) comerciales. Algunas de las herramientas más acreditadas del mercado son comercializadas por IBM, SAS, Oracle, SAP y Microsoft.

**IBM SPSS Modeler**
**https://www.ibm.com/products/spss-modeler; https://www.ibm.com/spss**

Es una herramienta de Minería de Datos de IBM (antes conocida por el popular nombre de Clementine) muy eficaz, que funciona con diferentes fuentes de datos (ASCII, ODBC), con una interfaz visual basada en procesos de datos y que manipula la mayoría de las diferentes técnicas de Minería de Datos más populares —clasificación, regresión, agrupamiento (*clustering*), redes neuronales, reglas temporales, árboles de decisión—. La interfaz gráfica facilita el descubrimiento de patrones y tendencias de grandes volúmenes de datos (estructurados y no estructurados).

Es una herramienta muy popular en análisis predictivo, se la considera como uno de los mejores proveedores de software de Minería de Datos, y se adapta para el análisis de Minería de Datos con Big Data. La herramienta SPSS Modeler es un programa comercial con licencia de pago y será necesario que las empresas analicen previamente la puesta en marcha de la aplicación, aunque la experiencia y eficacia de la herramienta y su gran implantación en el mercado en los últimos años, tanto en grandes empresas como en pymes, hacen que SPSS Modeler sea una opción muy idónea como herramienta de Minería de Datos. El sistema se integra con IBM Cognos e InfoSphere. La combinación de estas tres herramientas ayuda a realizar predicciones precisas sobre cómo se desarrollará el estado de su negocio. También ayuda a mejorar los resultados comerciales en CRM, marketing, planificación de recursos, mitigación de riesgos y otras áreas.

**SAS Enterprise Miner (www.sas.com/technologies/analytics/datamining/miner/)**

Es una herramienta de SAS (uno de los proveedores líder a nivel mundial de herramientas de analítica de datos) que realiza el proceso de Minería de Datos y facilita la creación de modelos predictivos y descriptivos de alta precisión para grandes volúmenes de datos. La herramienta dispone de una interfaz gráfica que integra un conjunto de herramientas necesarias para la toma de decisiones y otras herramientas de análisis de datos pioneras y líderes del mercado de analítica.

SAS Enterprise Miner está basada en la metodología de Minería de Datos SEMMA, desarrollada por SAS Institute. Es una de las herramientas con mayor implantación en el mercado para soluciones de grandes bases de datos y de Big Data. Ofrece un amplio conjunto de algoritmos avanzados muy completo para modelados predictivos y descriptivos, incluyendo regresión, redes neuronales, árboles de decisión. SAS es uno de los grandes proveedores de software analítico a nivel mundial; por lo que su adopción como herramienta de Minería de Datos entraña también un estudio profundo sobre su coste y funcionalidades, especialmente, para aquellas empresas que no sean clientes ya de otras soluciones de SAS.

**Oracle Data Mining (www.oracle.com/products/database/options/advanced-analytics/index.html)**

Oracle Data Mining (ODM) es una herramienta de software de Minería de Datos desarrollada por la empresa Oracle —líder mundial en software de bases de datos— con soporte en las técnicas más avanzadas de Minería de Datos y aplicada a grandes volúmenes de datos. Es una herramienta que realiza el ciclo completo de integración de datos de la importación de datos, preparación, desarrollo al despliegue del modelo.

La herramienta ODM integra todas las etapas del proceso de Minería de Datos y permite integrar los modelos con las bases de datos comerciales de Oracle, por lo que no necesita exportar los archivos de usuario a paquetes de software externos.

Una funcionalidad muy sobresaliente de la herramienta de Minería de Datos de Oracle es la oferta de las dos versiones comerciales que ofrece ODM:

1. Herramientas tradicionales de Minería de Datos que, mediante una interfaz gráfica muy potente, les permite a los usuarios aplicar las técnicas que consideren necesarias para poder tomar las mejores decisiones en los procesos de negocios.
2. La herramienta ODM permite a los desarrolladores utilizar las API de Oracle, por lo que podrán realizar sus propias aplicaciones específicas para la empresa. Esta propiedad es una excelente herramienta para las empresas clientes de Oracle, sobre todo, si utilizan otras herramientas del catálogo empresarial.

**SAP Business Object**
**(**`www.sap.com/spain/products/analytics/business-intelligence-bi.html`**)**

Es una herramienta integrada que, además de Minería de Datos, abarca todo el ciclo de vida de un sistema de Inteligencia de Negocios. Está disponible como una solución integrada en la empresa cliente (*on premise*) o en la nube. Sus funcionalidades más importantes son plataforma de Inteligencia de Negocios, visualización y análisis de datos, tableros de datos (*dashboards*) e informes. Se integra fácilmente con el sistema de bases de datos "en memoria" (SAP HANA).

**Microsoft. SQL Server Data Mining**
**(https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-ssas)**

SQL Server es una herramienta para análisis predictivo, desde la versión en el 2000, y facilita la Minería de Datos en Analysis Services. La combinación de la Minería de Datos con *Integration Services, Reporting Services y SQL Server* proporciona una plataforma integrada para el análisis predictivo, en la que se incluye la limpieza de los datos, la preparación, el aprendizaje automático y la generación de informes. SQL Server incluye varios algoritmos estándar, como los modelos de clústeres EM y mediana-K, redes neuronales, regresión logística y regresión lineal, árboles de decisión y clasificadores de Bayes. Todos los modelos tienen visualizaciones integradas para desarrollar, restringir y evaluar los modelos. Integrar la Minería de Datos en una solución de inteligencia empresarial puede ayudar a tomar decisiones inteligentes sobre problemas complejos.

En SQL Server 2017, la Minería de Datos es eficaz y accesible, y está integrada con las herramientas preferidas de los usuarios para el análisis y la creación de informes.

**Otros proveedores**

Los proveedores de Minería de Datos son muy numerosos. Muchos ofrecen soluciones tanto de Minería de Datos como de Aprendizaje Automático, y otros integran ambos tipos de soluciones. Así otros proveedores de productos de software son:

- Altair
- Alterix

- DataiKu
- Tableau
- Hadoop
- Apache Spark
- Sisense
- Teradata

### 5.7.3 CASO DE ESTUDIO: KAGGLE

Kaggle es un portal de referencia a nivel mundial (www.kaggle.com) en Ciencia de Datos y, por consiguiente, en Minería de Datos y herramientas de software de Ciencia de Datos. Es la comunidad internacional de ciencia de datos y de científicos de datos más grande del mundo. Es uno de los sitios más recomendados para aprendizaje, formación e investigación en disciplinas como Ciencia de Datos y, en consecuencia, Minería de Datos, Aprendizaje Automático, Aprendizaje Profundo y algoritmos de Inteligencia Artificial. La comunidad de Kaggle es muy numerosa y ayuda a mejorar las habilidades de los científicos de datos y, por sus características, es un portal excelente para el aprendizaje de minería de datos y aprendizaje automático y, por extensión, ciencia de datos.

En el **apartado 3.8.1 del capítulo 3** se describe más en detalle, las características y las grandes ventajas de usar y manejar con eficiencia el portal Kaggle. Cifras de octubre de 2022 consideran que Kaggle ha aumentado un 128 % en 5 años y la plataforma de ciencia de datos tiene más de 5 millones de usuarios en 194 países.

## RESUMEN

Minería de Datos es un conjunto de técnicas que se utiliza para optimizar el desarrollo de la Inteligencia de Negocios a partir de los datos que una organización recolecta, organiza y almacena para, posteriormente, realizar el análisis de datos y contribuir del modo más eficiente posible en la toma de decisiones.

El término Minería de Datos (*Data Mining)* se utiliza para describir el descubrimiento o "minado" (*mining*) del conocimiento a partir de grandes cantidades de datos. La Minería de Datos es una etapa dentro del proceso KDD (*Knowledge Discovery from Data*), descubrimiento del conocimiento a partir de datos, aunque el proceso KDD es un término más global, es decir, la Minería de Datos es una etapa dentro del proceso global KDD.

La Minería de Datos es una disciplina en la que confluyen muchas otras disciplinas clásicas en el mundo científico y de negocios: estadística, Inteligencia Artificial, Aprendizaje Automático, Aprendizaje de Máquina, sistemas de información, algoritmos, bases de datos y, recientemente, Ciencia de Datos o Big Data.

Las aplicaciones de la Minería de Datos son muy numerosas y de impacto en una gran cantidad de sectores de todo tipo: medicina, industria, banca, seguros, hardware y software de computadoras, gestión de relaciones con los clientes, deportes.

La extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento de conocimiento en bases de datos (KDD), que se refiere al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información: Las etapas del proceso KDD son:

1. Selección de datos
2. Preprocesamiento
3. Transformación
4. Minería de Datos
5. Interpretación y evaluación

La metodología o proceso CRISP-DM tiene seis etapas:

1. Comprensión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Despliegue/Distribución o desarrollo (implantación)

Los modelos de Minería de Datos más empleados, en función de su propósito general, son: descriptivos y predictivos.

La Minería de Datos tiene relación con otras disciplinas importantes en la Inteligencia de Negocios:

- Estadística
- Computación paralela y distribuida
- Aprendizaje automático (*Machine Learning*).
- Visualización
- Sistemas de toma de decisiones
- Tecnologías de bases de datos
- Reconocimiento de patrones.
- Algoritmia y programación
- Otras disciplinas

Existen numerosas herramientas de Minería de Datos bien específicas o integradas en paquetes o *suites*, con otras herramientas de Inteligencia de Negocios. Las herramientas de código abierto y gratuitas más populares son: WEKA, KNIME, RapidMiner, KEEL, Orange y proveedores comerciales de software propietario como IBM SPSS Modeler (Clementine), SAS Enterprise Miner, Oracle Data Mining.

## BIBLIOGRAFIA

BURK, Scott y Gary D. MINER (2020). It's All Analytics!. The Foundations of AI, Big Data, and Data Science Landscape for Professionals in Healthcare, Business, and Government. CRC Press.

CIELEN, Davy; Arno D. B. MEYSMAN y Mohamed ALI (2016). Introducing Data Science.

JOYANES, Luis (2019). Inteligencia de Negocios y Analítica de Datos. Una visión global de Business Intelligence & Analytics. Editorial Alfaomega CDMX, México y Marcombo (Barcelona, España).

JOYANES, Luis (2016). Sistemas de Información. Un enfoque dirigido a la empresa. Barcelona: Marcombo; México DF: Alfaomega.

JOYANES, Luis (2014). Big Data. El análisis de los grandes volúmenes de datos. Barcelona: Marcombo; México DF: Alfaomega.

KOTU, Vijay y Bala DESHPANDE (2019). Data Science. Concepts and Practice. Elsevier.

PROVOST, Foster y FAWCETT, Tom (2013). Data Science for Business. O`Reilly

SHAH, Chirag (2020). A Hands-On Introduction to Data Science. Cambrige University Press.

SHARDA, Ramesh; Delen DURSUN y Efrain TURBAN (2017). Business Intelligence and Data Science: A Managerial Perspective. New Jersey, USA: Pearson.

SHARDA, Ramesh; Delen DURSUN y Efrain TURBAN (2018). Business Intelligence, Analytics and Data Science: A Managerial Perspective. New Jersey, USA: Pearson Education. Versión Global Edition (2019).

SHARDA, Ramesh; Dursun DELEN y Efrain TURBAN (2020). Analytics, Data Science, & Artificial
Intelligence. Systems for Decision Support. Pearson.

# APENDICE I. METODOLOGÍA DE MINERÍA DE DATOS CRISP-DM

La Minería de Datos, como señalaban en los primeros años de despliegue de Ciencia de Datos, Provost y Fawcett (2013) es un arte que implica la aplicación de una cantidad sustancial de ciencia y tecnología, pero la aplicación propia en sí también es un arte. El proceso de Minería de Datos más aceptado universalmente sigue el proceso experimentado y probado en la industria como CRISP-DM.

El proceso o metodología CRISP-DM —*Cross Industry Standard Process for Data Mining* (www.crisp-dm.org)— fue propuesto en la segunda mitad de la década de los noventa por un consorcio europeo de empresas —entre sus fundadores se cuentan Daimler, Chrysler, SPSS y NCR— y se ha convertido en una metodología de Minería de Datos abierta y no propietaria. Esta metodología presenta la ventaja de haber sido diseñada y construida sobre la base de experiencias reales, no de modo teórico, por empresas de gran prestigio en la industria y empresas de tecnologías de la información. La figura 5.2 ilustra el proceso propuesto en CRISP-DM, que consta de seis etapas que comienzan con un buen conocimiento del negocio y el dominio de la aplicación y termina con el despliegue de la solución que cumple la necesidad específica del negocio. En origen, el proceso es secuencial (avances), por naturaleza, aunque en algunas etapas se consideran flujos de datos bidireccionales, lo que significa que algunas fases permitirán revisar parcialmente o en su totalidad las fases anteriores. La Minería de Datos se apoya en el conocimiento; en consecuencia, es muy importante tener en cuenta la experiencia y la experimentación.

CRISP-DM actúa como una metodología y como un proceso. La metodología incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. El modelo de proceso ofrece un resumen del ciclo vital de la Minería de Datos. Las seis fases o etapas son:

1. Comprensión del negocio.
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Despliegue

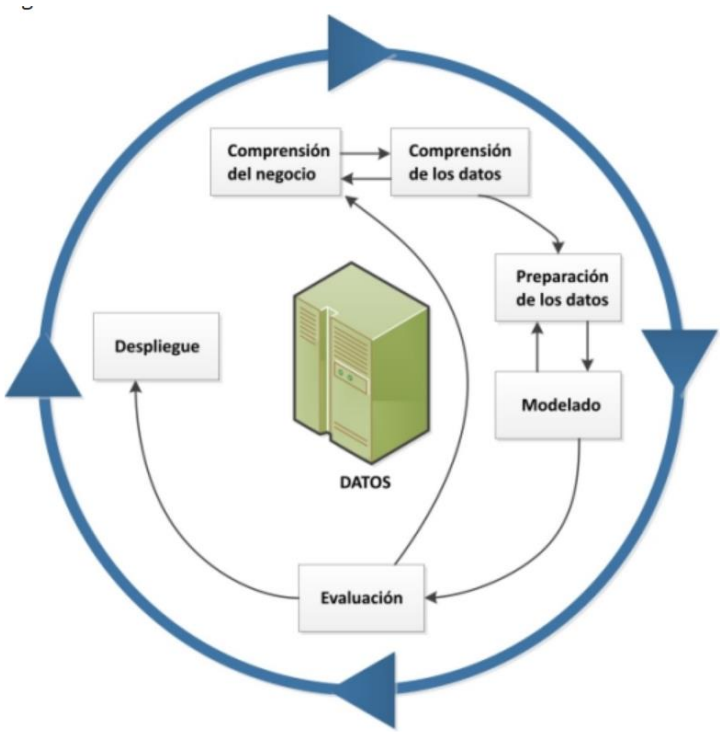



**Figura 5.3. Etapas del proceso de Minería de Datos CRISP-DM. (Modelo de referencia)**

El ciclo de vida del modelo tiene las seis etapas (fases) anteriores, que indican las dependencias más importantes y frecuentes. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario. El resultado de cada fase determina que dicha fase o la tarea específica de una fase tenga que ser realizada después. Las flechas indican las más importantes y frecuentes dependencias entre fases. De hecho, el diagrama del proceso explicita el hecho de que la iteración es la regla en lugar de la excepción. Normalmente, el proceso completo es una exploración de los datos y después de la primera iteración el equipo de Ciencia de Datos conoce mucho más. La siguiente iteración puede estar mucho más informada. El círculo externo de la figura 5.1 representa la naturaleza cíclica de la Minería de Datos, que no se termina una vez que la solución es desplegada. Examinemos las etapas en detalle. La figura 5.2 muestra la fuente original de la metodología

## Comprensión del negocio

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y en la definición de las necesidades del cliente. Este conocimiento de los datos, después,

se convierte en la definición de un problema de Minería de Datos y en un plan preliminar diseñado para alcanzar los objetivos. En esta fase, se trata de entender los objetivos del negocio y requerimientos del proyecto, desde una perspectiva del negocio y no técnica.

## Comprensión de los datos

La fase de comprensión y estudio de datos comienza con recopilar los datos, descubrir conocimiento preliminar sobre ellos y continúa con las actividades que permiten familiarizarse con ellos, identificar los problemas de calidad y analizar las primeras potencialidades, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

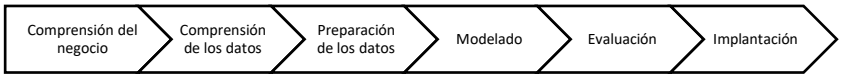



**Figura 5.4 Modelo de referencia CRISP_DM**
**Fuente: CRISP_DM 1.0 (p. 12)**

## Preparación de los datos

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado), a partir de los datos iniciales brutos. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan. En resumen, en esta fase se realiza el análisis de datos y la selección de características.

El objetivo de esta fase es obtener la vista "minable". Se realiza la selección de datos, a los que posteriormente se aplicarán las técnicas de modelado (variables y muestras), la limpieza de datos, la generación de variables adicionales, la integración de diferentes conjuntos de datos y los cambios de formato.

## Modelado de los datos

En esta fase, se seleccionan y aplican las técnicas de modelado pertinentes al problema (cuantas más, mejor) y se calibran sus parámetros a valores óptimos. Típicamente, hay varias técnicas para el mismo tipo de problema. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Se seleccionan diversas técnicas de modelado adecuadas a un conjunto de datos ya preparado (la vista "minable"), a fin de centrarse en las necesidades específicas del negocio.

## Evaluación

En esta etapa del proyecto, se han construido uno o varios modelos que parecen alcanzar una calidad suficiente desde la perspectiva del análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo, revisar los pasos ejecutados para crearlo y comparar el modelo obtenido con los objetivos del negocio. Un objetivo clave es determinar si hay alguna cuestión importante del negocio que no haya sido considerada lo suficiente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos. El resultado final de esta fase es la obtención de resultados.

Aquí, se evalúa el modelo de la fase anterior, es decir, si el modelo nos sirve para responder a algunos de los requerimientos del negocio.

## Despliegue

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es el de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica, quizás automatizada, de un proceso de análisis de datos en la organización. El objetivo final de esta fase es la distribución o desarrollo (despliegue) y la puesta en producción.

La fase de despliegue (implementación, implantación o distribución) trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisiones de la organización, difundir informes sobre el conocimiento extraído.

# APENDICE II. PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO: KDD

Otra metodología que tiene también gran aplicación en el proceso de Minería de Datos es KDD. La extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento, conocido como descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases*, KDD), que se refiere al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información[1]. No es un proceso automático, es iterativo que, exhaustivamente, explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad y que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos.

Durante la década de los noventa, el citado término fue un sinónimo de Minería de Datos. KDD es un proceso que utiliza métodos de Minería de Datos para

encontrar información útil y patrones de datos, al contrario que el concepto de Minería de Datos, que implica utilizar algoritmos para identificar patrones de datos a través del proceso KDD.

El proceso KDD fue el primer modelo aceptado por la comunidad científica que estableció las etapas principales de un proyecto de explotación de datos. En su origen, el modelo KDD establece que la Minería de Datos es una etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos. Aunque en la bibliografía científica y profesional el término KDD y Minería de Datos son sinónimos y se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento del conocimiento, desde un punto de vista de proceso de extracción y análisis de datos, KDD es un proceso completo que comprende a la Minería de Datos como una de sus etapas.

KDD es un proceso no trivial que identifica patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos.

**Etapas del proceso KDD**

El proceso de extracción/descubrimiento del conocimiento en bases de datos —Fayad et al.[7] (1996) y Dunham (2003)— consta de las siguientes etapas:( figura 5.5):

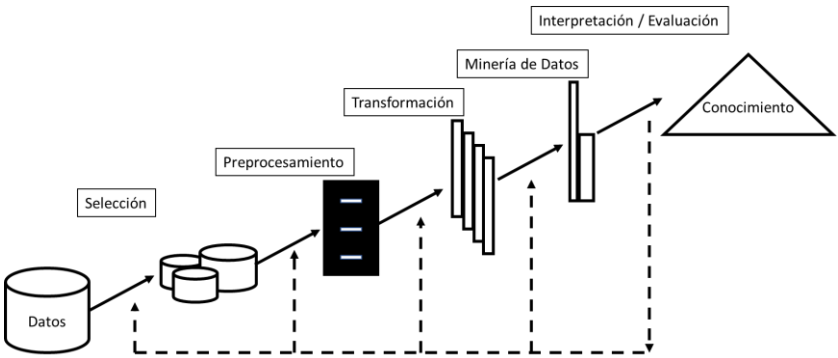



**Figura 5.5 Proceso de descubrimiento del conocimiento (KDD)**
**Fuente: (Fayyad, 1996)**

1. **Selección de datos.** En esta etapa, se determinan las fuentes de datos y el tipo de información que se ha de utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde las fuentes de datos.

[7] Fayyad et al. (1996). *The KDD Process for Extracting Useful Knowledge from Volumes of Data.* Communications of the ACM, 39(11), 27-34.

2. **Preprocesamiento.** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa, se utilizan diversas estrategias para manejar datos faltantes o ausentes, en blanco, datos inconsistentes o que están fuera de rango, y se obtienen al final una estructura de datos adecuada para su posterior transformación.
3. **Transformación.** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes, con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.
4. **Minería de Datos.** Es la fase de modelado propiamente dicho, en donde se aplican métodos inteligentes con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles, que están contenidos u "ocultos" en los datos.
5. **Interpretación y Evaluación.** Se identifican los patrones obtenidos, que son realmente interesantes, que se basan en algunas medidas; se realiza una evaluación de los resultados obtenidos.

Además de las fases descritas, frecuentemente, se incluye una fase previa de análisis de las necesidades de la organización y definición del problema, en la que se establecen los objetivos de la Minería de Datos. También es usual incluir una etapa final, donde los resultados obtenidos se integran al negocio para la realización de acciones comerciales.

La Minería de Datos es una etapa del modelo KDD aunque con mucha frecuencia en la literatura se suelen superponer o identificar los dos términos.

**Proceso de Minería de Datos**

A grandes rasgos, el proceso de Minería de Datos puede dividirse en seis pasos:

1. *Selección del conjunto de datos*: aquí se decide cuáles van a ser las variables objetivas (aquellas que se quieren predecir o inferir), las variables independientes y la selección de registros (datos) que se han de utilizar.
2. *Análisis de las propiedades de los datos:* mediante, por ejemplo, histogramas y/o diagramas de dispersión. Búsqueda de valores atípicos (*outliers*) y ausencia de datos.
3. Transformación o preprocesamiento del conjunto de datos de entrada: en este paso, se normalizan los datos a una misma escala. También se decide cómo se van a tratar datos faltantes, atípicos o dudosos. Una posibilidad es tratarlos como un tipo de dato especial o bien decidir descartarlos.
4. *Selección y aplicación de técnicas de Minería de Datos:* se construye un modelo que se utilizará sobre los datos para predecir las clases mediante clasificación o para descubrir grupos similares mediante segmentación

5. *Extracción de conocimiento:* una vez aplicado el paso anterior, se buscan patrones de comportamiento en los valores de las variables del problema o las relaciones de asociación entre dichas variables.
6. *Interpretación y evaluación de datos:* el modelo debe ser validado, comprobando que las conclusiones arrojadas son válidas y satisfactorias. Si el modelo final no supera esta evaluación, el proceso puede repetirse desde el principio o a partir de cualquiera de los pasos anteriores.

***Knowledge Discovery in Databases* (KDD, descubrimiento de conocimiento en bases de datos).** Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos (Fayyad et al., 1996).

**Minería de Datos.** Proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten y Frank, 2016). La Minería de Datos es la etapa más importante del KDD que integra los procesos de aprendizaje y métodos estadísticos para la obtención de hipótesis de patrones y modelos.

## APENDICE III. METODOLOGÍA DE PROCESOS DE MINERÍA DE DATOS SEMMA

La metodología SEMMA, abreviatura de *sample* (muestreo), *explore* (exploración), *modifiy* (modificación), *model* (modelado) y *asses* (valoración) es también muy conocida y utilizada. Se puede definir como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos". Fue desarrollada por el SAS Institute (2005).

El proceso de Minería de Datos SEMMA se compone de las siguientes etapas:

**Muestreo**

Genera una muestra representativa de datos. Se identifican los datos (entrada de datos, ejemplos, partición de datos).

**Exploración**

Visualización y descripción básica de los datos. Se exploran los conjuntos de datos para observar relaciones o patrones, y se generan análisis diversos, identificación de variables importantes y análisis de asociación (exploración distribuida, múltiples particiones, intuición, asociación, selección de variables).

**Modificación**

Selección de variables y transformación de la representación de variables. Se preparan los datos para el análisis (transformación de variables, filtros a los datos fuera de rango, agrupamiento, ruido).

**Modelado**

Utiliza técnicas diversas de estadística y modelos de aprendizaje automático (regresión, arboles, redes neuronales).

**Evaluación (Valoración)**

Evalúa la precisión y utilidad de los modelos (evaluación, medidas, reportes).


**Figura 5.6. Etapas del proceso de Minería de Datos SEMMA**

SEMMA comienza con una muestra representativa y estadística de los datos que facilita aplicar técnicas de exploración, estadística y visualización; selecciona y transforma las variables predictivas significativas, modela las variables para predecir resultados y confirma una precisión del modelo (Sharda, 2020).

Al evaluar el resultado de cada etapa del proceso SEMMA, el desarrollador del modelo puede determinar cómo modela nuevas cuestiones surgidas en los resultados anteriores y retroceder a la fase de exploración para refinamientos adicionales de los datos. *Al igual que la metodología CRISP-DM, SEMMA se controla mediante un ciclo iterativo de experimentación*.

La principal diferencia entre ambas metodologías de Minería de Datos es que CRISP-DM se centra en un enfoque más exhaustivo (completo) de los procesos (incluyendo comprensión del negocios y datos relevantes), mientras que SEMMA supone implícitamente que las metas y objetivos del proyecto y negocio, junto con las fuentes de datos apropiadas, han sido identificadas y comprendidas (Sharda, 2020).

# CAPÍTULO 6
# VISUALIZACIÓN DE DATOS: INFORMES Y CONSULTAS, CUADROS DEMANDO (DASHBOARDS) Y CUADRO DE MANDO INTEGRAL (CMI)

## 6.4 TIPOS DE GRÁFICOS

Existen numerosos tipos de gráficos (*diagrama* es un término sinónimo muy empleado) y uno de los aspectos más importantes y, a veces más complejos, es la selección del tipo de gráfico que se ha de utilizar en un informe o cuadro de mando. Describiremos los más populares que son utilizados en los diagramas de visualización, así como otros diferentes gráficos que han alcanzado gran notoriedad en los cuadros de mandos o paneles de control y en los medios de comunicación tradicionales y digitales. Los gráficos tradicionales siempre se han agrupado en torno a las tablas y cuadros, aunque en la actualidad la visualización de datos contempla un sinfín de técnicas y herramientas que facilitan el acceso a datos más significativos y representativos, así como las relaciones entre ellos.

**Gráfico de barras**

Es una representación gráfica en un eje cartesiano de las frecuencias de una variable discreta o cualitativa. Las barras se colocan vertical u horizontalmente, unas al lado de las otras. En uno de los ejes se colocan las distintas modalidades de la variable cualitativa o discreta y en el otro eje, el valor o frecuencia de cada categoría en una determinada escala. Esta visualización es muy útil en la realización de comparaciones sencillas entre valores adyacentes.

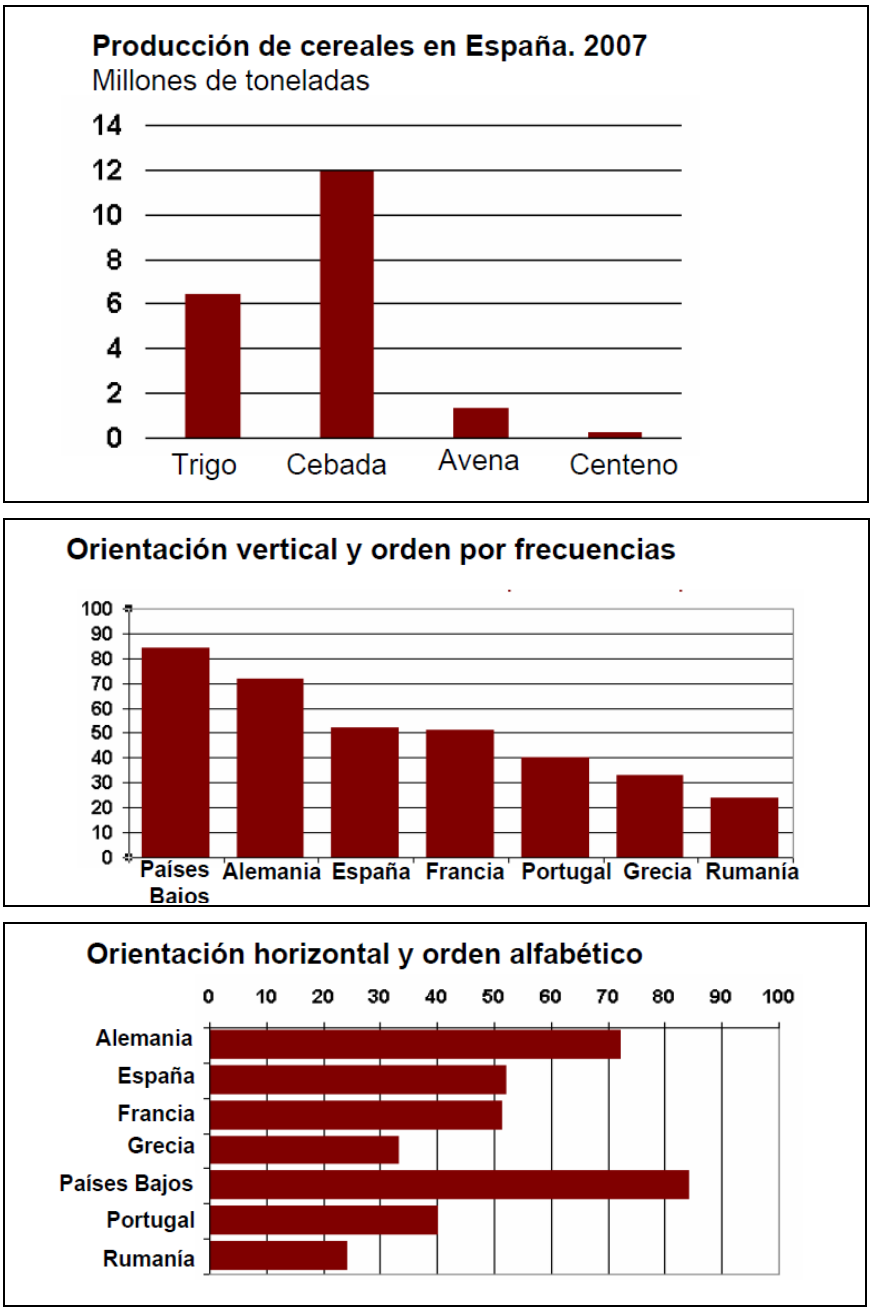



**Figura 6.1. Diagrama de barras**

La orientación del gráfico puede ser vertical (las categorías se sitúan en el eje horizontal y las barras de frecuencia crecen verticalmente) u horizontal (las categorías se sitúan en el eje vertical y las barras crecen horizontalmente).

Los gráficos de barras pueden ser:

- **Sencillo** (contiene una única serie de datos).
- **Agrupado** (contiene varias series de datos y cada una de ellas se representa por un tipo de barra, diferentes colores y texturas).
- **Apilado** (contienen varias series de datos y la barra se divide en segmentos de diferentes colores o texturas y cada una de ellas representa una serie de datos).
- **Bidireccional** (tiene orientación horizontal y contiene dos series de datos, cuyas barras de frecuencia crecen en sentidos opuestos).

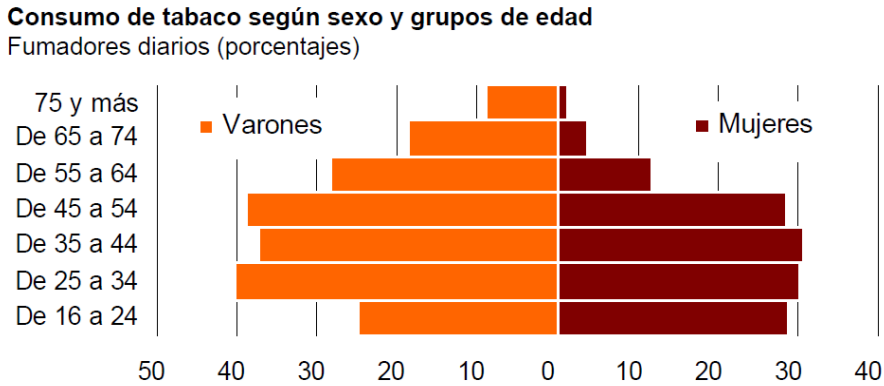



**Figura 6.2. Diagrama de barras. Fuente: Encuesta Nacional de Salud 2006, INE**

Es una representación gráfica en un eje cartesiano de la relación que existe entre dos variables que refleja con claridad los cambios producidos. Se emplea para representar tendencias temporales: las variaciones en el precio de las acciones de una empresa con el paso del tiempo, los ingresos durante un periodo de tiempo. En cada eje se representa cada una de las variables cuya relación se desea observar (por ejemplo, meses o trimestres); en el eje vertical, la otra variable (por ejemplo, ingresos).

El gráfico de líneas puede mostrar una, dos o múltiples series de datos.

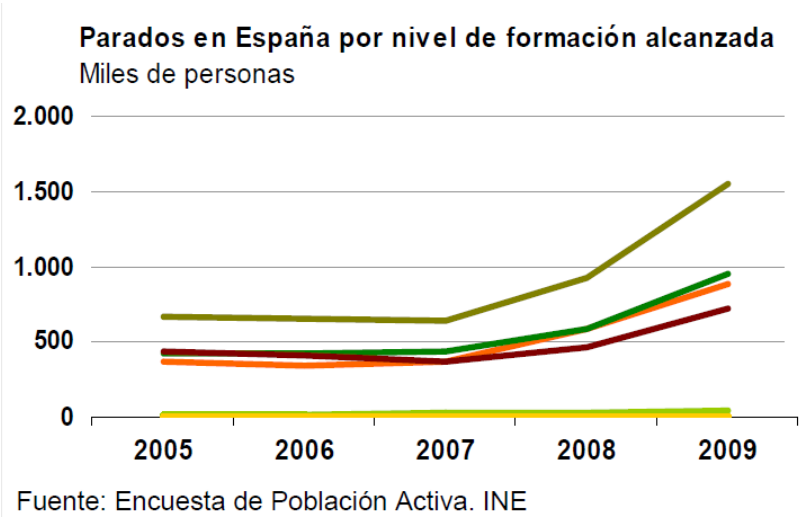



**Figura 6.3. Diagrama de líneas**

**Gráfico circular (de sectores)**

Un gráfico circular o de sectores es una representación circular de las frecuencias relativas de una variable cualitativa o discreta, que permite su comparación de una manera rápida o sencilla. El círculo representa la totalidad que se quiere representar (por ejemplo, total de turistas o viajeros alojados en hoteles) y cada porción (sector) es la proporción de cada categoría de la variable. Suele expresarse en porcentajes del total. El gráfico se puede representar como un pastel si se desea expresar grosor o simplemente resaltar los sectores. Muestra diversas porciones de una tarta donde cada trozo representa los datos como porcentaje del total.

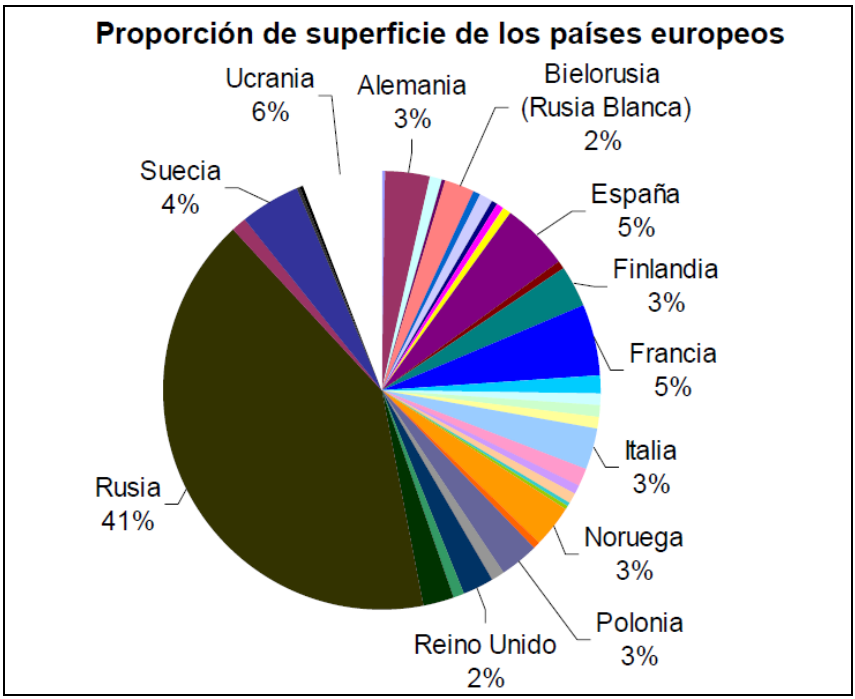



**Figura 6.4 Diagramas de sectores**

Es una variante del gráfico de barras. Se diseñó para reemplazar los indicadores, medidores y termómetros de los paneles de control. En estos gráficos, se compara una variable principal (por ejemplo, ingresos del año hasta una fecha determinada) con una o más medidas distintas (como ingresos anuales), y se presenta en el contexto de métricas de rendimiento definido (por ejemplo, cuota de ventas).

La observación de un gráfico de balas indica instantáneamente el rendimiento de la medida primaria con los objetivos generales (por ejemplo, el cumplimiento de un departamento de conseguir el logro de su cuota anual de ventas).

#### Gráfico de dispersión

Un gráfico o diagrama de dispersión muestra en unos ejes cartesianos la relación que existe entre dos variables e informa acerca del grado de correlación entre ellas. Es muy útil para mostrar la correlación entre dos conjuntos de datos (variables) y mostrar la fuerza y la dirección de dicha relación.

El tipo de correlación se muestra mediante un formato de nube de puntos y puede ser de dos formas: **correlación lineal**, donde existe una relación lineal negativa o positiva, y **correlación no lineal**, donde existe una relación entre las variables, pero no es lineal. También podría existir una **correlación nula**.

En la correlación lineal, a veces, se dibuja una recta de regresión que se obtiene gracias a un ajuste lineal, mediante un método matemático.

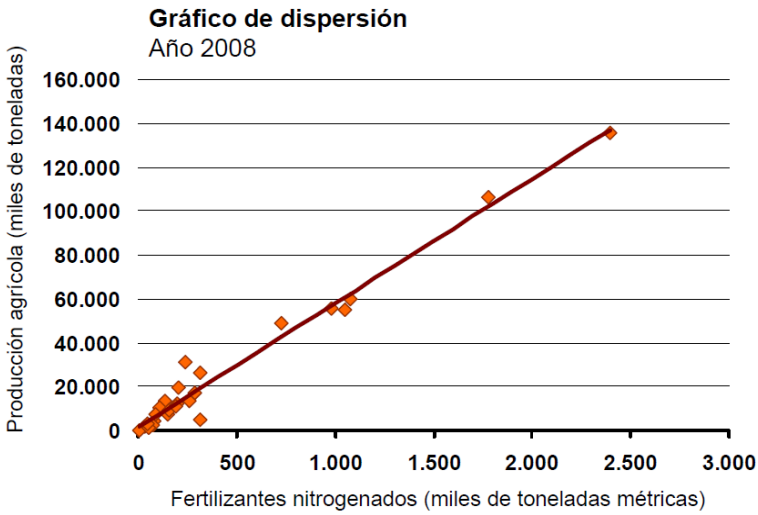



**Figura 6.5. Gráfico de dispersión**

### Gráfico de burbujas

Es una variante del gráfico de dispersión o de mapas al que se añade una tercera dimensión, vinculada al tamaño de los puntos (o burbujas); o una cuarta dimensión, vinculada con el color de cada burbuja. Las burbujas representan con sus diámetros una variable que se añade a otras dos variables que indican su posición.

Se emplea para acentuar datos en diagramas de dispersión y también para realizar superposiciones sobre mapas.

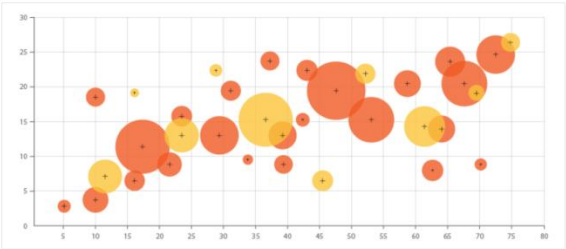


**Figura 6.6. Gráfico de burbujas**

### Gráfico en cascada

Es un tipo de gráfico normalmente utilizado para comprender cómo un valor inicial se ve afectado por una serie de cambios intermedios, positivos y negativos.

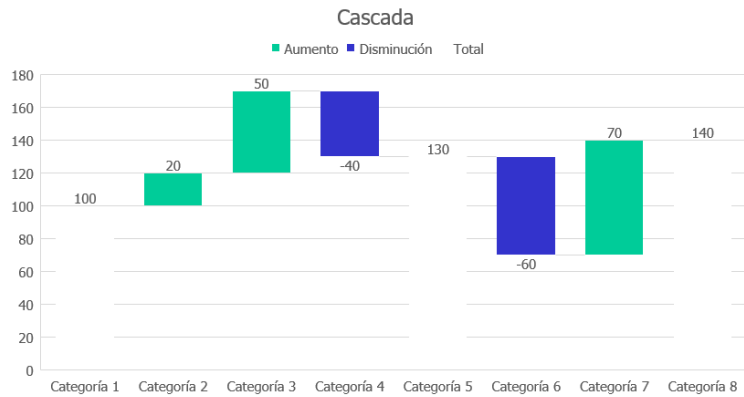



**Figura 6.7. Gráfico en cascada**

### Diagrama de Pareto

Es un tipo de gráfico de barras vertical ordenado por frecuencias de forma descendente, que identifica y da un orden de prioridad a los datos. Suele tener un eje horizontal, dos ejes verticales opuestos en los extremos del eje horizontal y una línea que atraviesa el diagrama de barras de izquierda a derecha, que representa un porcentaje acumulado de dichas frecuencias respecto al total.

Este gráfico representa el principio de Pareto, formulado por el estadístico italiano de igual nombre, que se sintetiza en el principio básico "pocos vitales, muchos triviales" (es decir, hay muchos problemas sin importancia frente a unos pocos graves. Por lo general, el 20 % de las causas totales originan el 80 % de los efectos; por esta razón se le conoce también como el principio 20-80).

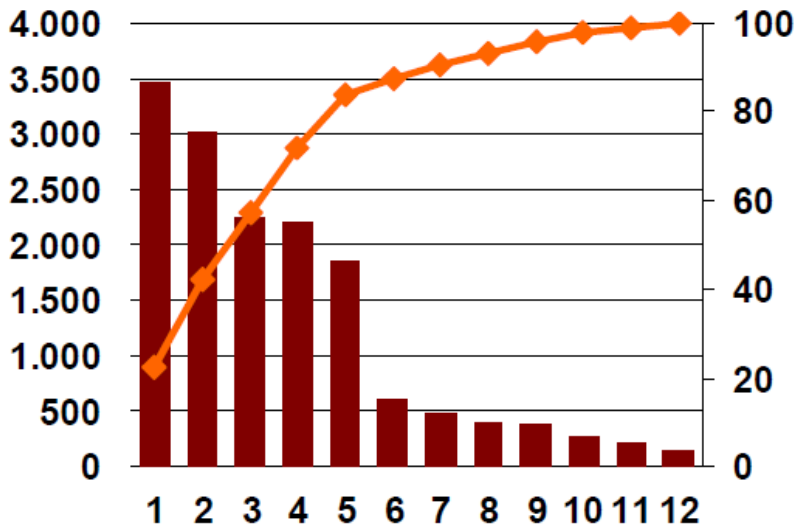



**Fig. 6.8. Diagrama de Pareto**

**Histograma**

Es un tipo especial de diagrama de barras que se usa para representar las frecuencias de una variable cuantitativa continua. En uno de los ejes se posicionan las clases de la variable continua (intervalos o marcas de clase, que son los puntos medios de cada intervalo); en el otro, las frecuencias. No existe separación entre las barras.

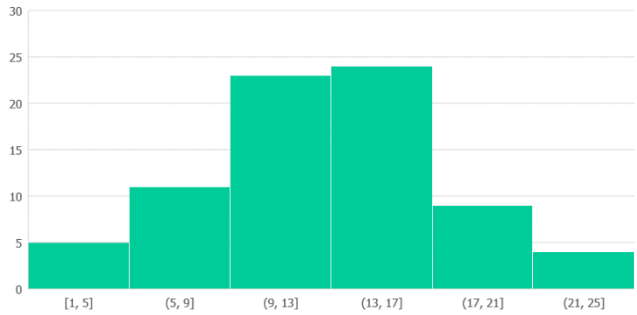



**Figura 6.9. Histograma**

Existen también histogramas bidireccionales que contienen dos series de datos cuyas barras de frecuencia crecen en sentidos opuestos.

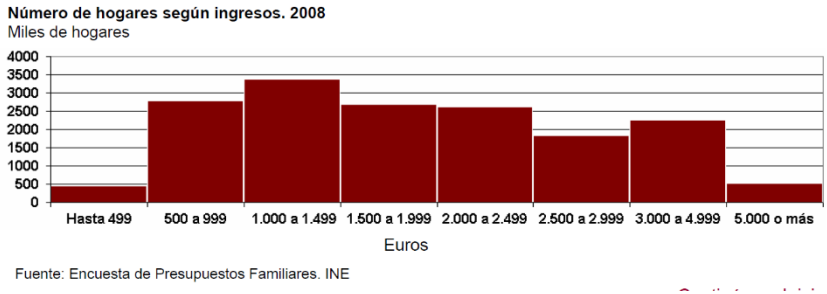



**Figura 6.10a. Histograma unidireccional**

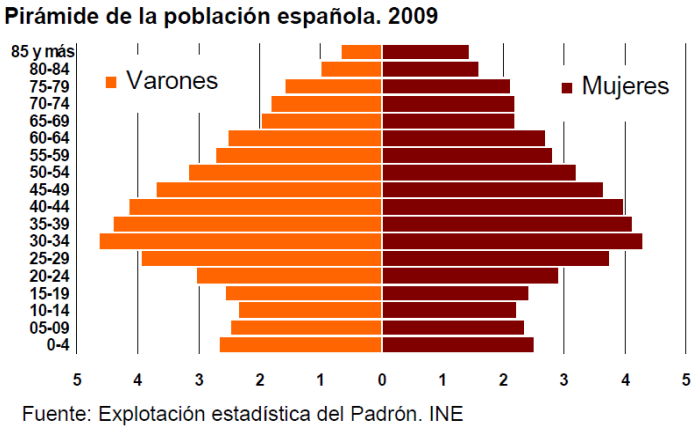
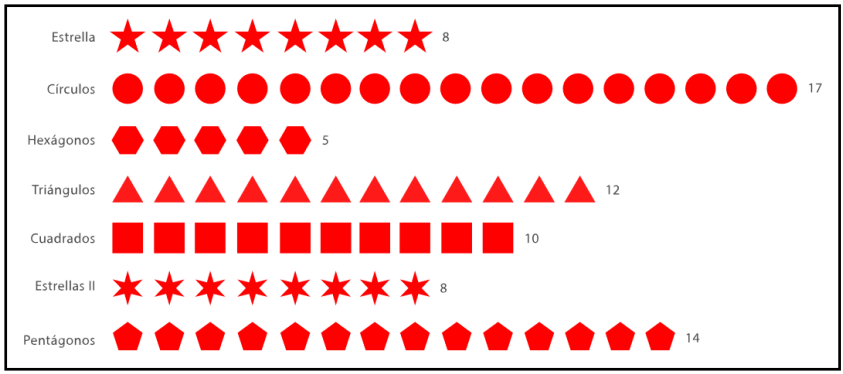


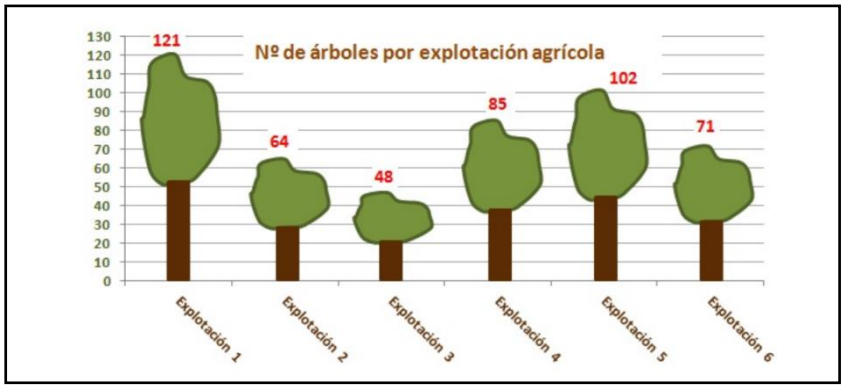

**Figura 6.10b. Histograma bidireccional**

### Pictograma

Es un gráfico que representa mediante figuras o símbolos las frecuencias de una variable cualitativa o discreta. Al igual que los gráficos de barras, se emplea para comparar magnitudes o ver la evolución en el tiempo de una categoría concreta.




**(a)**




**(b)**

**Figura 6.11. Pictograma**

**Fuente:**

**(a) <https://datavizcatalogue.com/ES/metodos/grafico_de_pictogramas.html>**

**(b) <http://www.universoformulas.com/estadistica/descriptiva/pictograma/>**

## 6.10 CUADRO DE MANDO INTEGRAL (CMI) / *BALANCED SCORECARD*

El cuadro de mando integral (CMI) o *Balanced Scorecard* es un sistema de gestión que permite traducir los objetivos estratégicos de una organización en resultados. El CMI es una herramienta de gestión empresarial muy útil para medir la evolución de la actividad de una compañía, sus objetivos estratégicos y resultados desde el punto de vista estratégico y de perspectiva general.

CMI es una metodología de planificación estratégica basada en indicadores (métricas y KPI), creada por los profesores de la Universidad de Harvard, Robert S. Kaplan y Daniel P. Norton, que relaciona factores medibles de procesos con la consecución de objetivos estratégicos. Surgió a principio de los años noventa como una respuesta ante la necesidad de analizar las organizaciones desde un punto vista estratégico, en lugar del financiero, más utilizado. El método propone un nuevo modelo de medidas que permite conocer mejor a las organizaciones

El CMI se aplica en empresas de distinto tamaño y sector, ya que representa un cambio fundamental en la cultura empresarial al permitir una trazabilidad entre la estrategia definida por la dirección, los objetivos propuestos para alcanzarla y la medición de los objetivos aplicando indicadores. En esencia, el Cuadro de Mando Integral mide la eficacia, la eficiencia y la calidad.

De acuerdo con lo anteriormente expuesto puede definirse el CMI como un método de planificación estratégica basado en métricas y procesos que relaciona factores medibles de procesos con la consecución de objetivos estratégicos.

El objetivo final de Kaplan y Norton era establecer un nuevo modelo de medidas que permitiera conocer mejor las organizaciones. El modelo define un cuadro de mando corporativo organizando indicadores clave de negocio en cuatro grupos o perspectivas, como veremos más adelante. En esencia, el modelo es una herramienta de gestión que permite a la dirección de la organización centrar su atención en lo que considera más importante para conseguir los objetivos estratégicos previstos.

El Cuadro de Mando Integral permite (Muñiz, 2016):

- "Evaluar de una forma eficaz si se están cumpliendo los resultados esperados.
- Confirmar que se avanza hacia los objetivos marcados en la estrategia.
- Comunicar a la organización cómo conseguir los objetivos estratégicos".

El Cuadro de Mando Integral es un cuadro de mando o tablero de control (*dashboard* o *scorecard*) que actúa como una herramienta de medición (indicadores, fundamentalmente KPI), una herramienta de gestión estratégica (actúa como coordinación esencial entre los objetivos estratégicos y las iniciativas a corto plazo para la consecución de los objetivos) y una herramienta de comunicación (permite comunicar los objetivos estratégicos a toda la organización).

## 6.10.1 PERSPECTIVAS DEL CUADRO DE MANDO INTEGRAL

El Cuadro de Mando Integral trata de cumplir la misión y la estrategia explícita de una organización con un amplio conjunto de medidas de actuación para la

creación de valor, organizadas de forma coherente en cuatro perspectivas diferentes o ámbitos en los que, normalmente, se divide la empresa para realizar las tareas anteriores.

Una característica importante del método es que utiliza tanto indicadores financieros como no financieros que, a su vez, tratan de ser un equilibrio entre los objetivos a corto y a largo plazo, entre las medidas financieras y las no financieras, entre los indicadores de retraso y entre las perspectivas internas y externas.

Las perspectivas (puntos de vista) son las dimensiones clave de la empresa y van a permitir relacionar los objetivos de todas las áreas o departamentos. Una perspectiva es el punto de vista respecto del cual se monitoriza el negocio. El CMI considera que toda empresa se organiza en cuatro perspectivas: *finanzas, clientes, procesos internos, aprendizaje y crecimiento*.

Las cuatro perspectivas básicas permiten un equilibrio entre los objetivos a corto y largo plazo, entre los resultados deseados y las iniciativas para conseguirlos. Los objetivos estratégicos estarán vinculados por las relaciones de causa y efecto de las diferentes perspectivas.

Sin embargo, el método tiene en cuenta que las cuatro perspectivas básicas se pueden ampliar o reducir en número, en función de las necesidades de la organización (no son obligatorias las cuatro). Se deben elegir las perspectivas que sean necesarias para describir los objetivos de la estrategia y su consecución. El número de perspectivas por elegir debe captar la atención de todos los interesados (*stakeholders*) y comprender los objetivos estratégicos.

Por ejemplo, en el caso de una empresa de fabricación de calzado y ropa deportiva, se debe contemplar una quinta perspectiva, además de la de clientes: la perspectiva de consumidores, ya que para una empresa de este sector son tan importantes los distribuidores como sus clientes finales, los usuarios de la ropa y del calzado.

Las cuatro perspectivas del Cuadro de Mando Integral permiten el equilibrio entre (Muñiz, 2016: 154):

- Los objetivos fijados para el corto y largo plazo.
- Los diferentes tipos de indicadores financieros y no financieros, de futuro y de pasado.
- Los intereses de los accionistas y clientes (externos), empleados y procesos internos.

**Tipos de perspectivas**

El método CMI engloba en las cuatro perspectivas la totalidad de los indicadores o variables (endógenas y exógenas), que repercuten en la cuenta de resultados de la compañía, elemento fundamental de valor para el accionista.

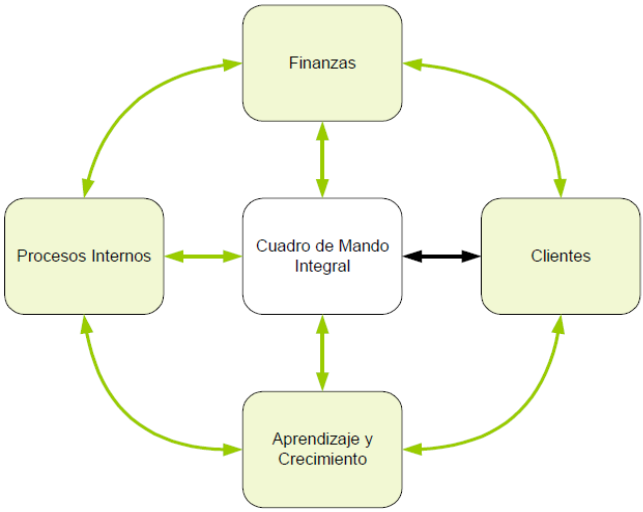



**Figura 6.30. Perspectivas del Cuadro de Mando Integral**

**Perspectiva financiera**. Permite medir las consecuencias económicas de las acciones tomadas en las organizaciones; incorpora la visión de los accionistas y mide la creación de valor de la empresa.

Responde a preguntas tales como: ¿Qué debemos hacer para satisfacer las expectativas de nuestros accionistas? ¿Cómo nos ven nuestros inversores y accionistas? ¿Qué resultados económicos esperan de la empresa sus grupos de interés?

**Perspectiva de cliente.** Refleja el posicionamiento de la empresa en el mercado o sectores de mercado donde se desea competir.

Responde a preguntas tales como: ¿Cómo nos ven nuestros clientes? ¿Qué debemos hacer para satisfacer las necesidades de nuestros clientes? ¿Qué aspectos de la relación con el cliente afectan a los resultados financieros?

**Perspectiva interna (procesos internos).** Explica las variables internas consideradas como críticas, así como define la cadena de valor generado por los procesos internos de la empresa.

Responde a preguntas tales como: ¿Cuáles son los procesos internos en los que debemos sobresalir para satisfacer a nuestros clientes? ¿En qué procesos internos debemos ser excelentes para satisfacer esas necesidades? ¿En qué debemos ser óptimos?

**Perspectiva de aprendizaje y crecimiento**. Identifica la infraestructura que la organización debe construir para crear crecimiento y valor a largo plazo.

Responde a preguntas tales como: ¿Cómo conseguir que la plantilla esté alineada con la estrategia definida por la dirección y mejorar así mi organización? ¿Cómo podemos mejorar y crear valor? ¿Qué competencias son clave para innovar y mejorar? ¿Qué debemos hacer para desarrollar los recursos internos necesarios para lograr la excelencia en los procesos clave?

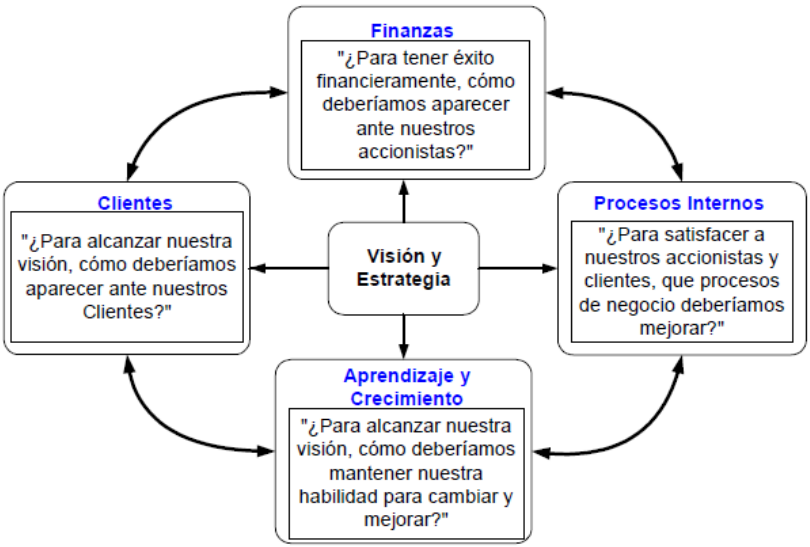
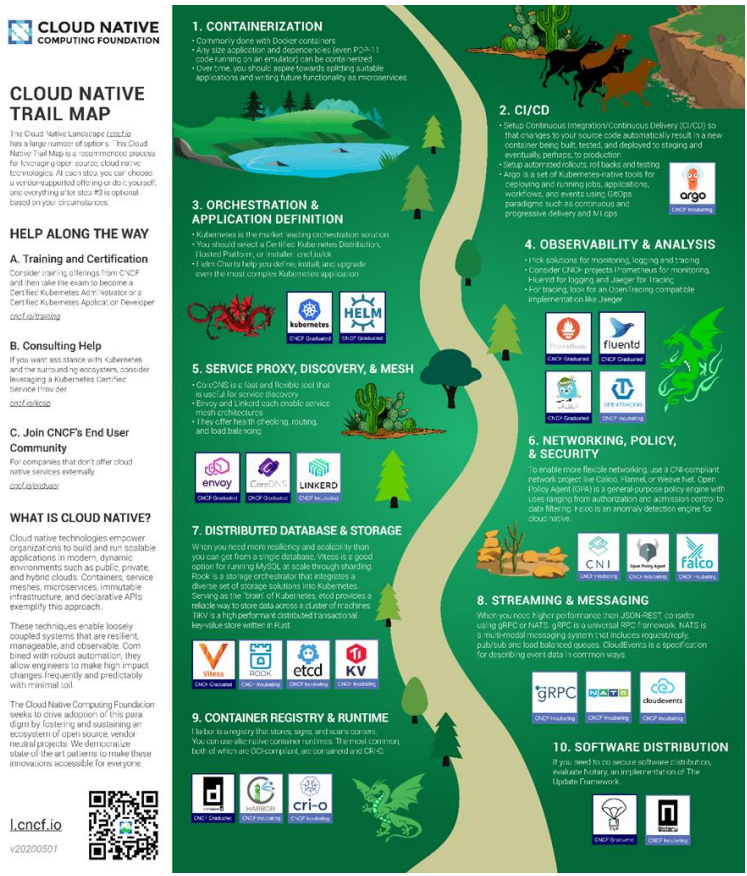



**Figura 6.31. Etapas en el diseño del CMI**

# CAPÍTULO 8
# COMPUTACIÓN EN EL BORDE, LA NUBE NATIVA Y LA COMPUTACIÓN SIN SERVIDORES

## PANEL DE CLOUD NATIVE COMPUTING FOUNDATION (CNCF)




**CNCF Serverless White Paper (<github.com/cncf/wg-s>)**

# CAPÍTULO 9
# ALMACENAMIENTO DE DATOS: DATAWAREHOUSES Y DATALAKES

## 9.3 MARCO DE TRABAJO (*FRAMEWORK*) DE UN SISTEMA DE ALMACENAMIENTO DE DATOS

Un *data warehouse* de una empresa (EDW*, Enterprise Data Warehouse*) es un almacén de datos que se utiliza en la empresa y es el depósito más importante de datos disponibles para el desarrollo de arquitecturas de Inteligencia de Negocios y sistemas de apoyo a la decisión. El término almacenamiento de datos indica el conjunto completo de actividades interrelacionadas implicadas en su diseño, implementación y uso.

Las organizaciones, públicas y privadas, recogen datos, información y conocimiento de modo continuo, que se almacena en sistemas computarizados. El mantenimiento y uso de esos datos se vuelve complejo, sobre todo, cuanto escalan a volúmenes mayores. El trabajo y la gestión con múltiples bases de datos integradas o no en un *data warehouse* se vuelve tarea compleja y extremadamente difícil. Una buena manipulación de estas herramientas proporciona grandes beneficios que compensará los excesivos costes que supone su implementación y actualización continua.

El marco de trabajo o entorno de un sistema de *data warehouse* y de *data mart* comienza extrayendo datos desde fuentes de datos (*data sources*), mediante un software denominado ETL (*Extract*, *Transform* y *Load*: extraer, transformar y cargar), y se carga en un área provisional, donde se transforman y limpian.

Una vez leídos los datos, se cargan en el depósito de datos. A continuación, son analizados, minados, presentados y visualizados, utilizando herramientas de negocio final como OLAP, Minería de Datos y de Texto, paneles o tableros de control (*dashboards*).

### 9.3.1 COMPONENTES DE UN *DATA WAREHOUSE*

El marco o entorno de trabajo de un *data warehouse* incluye, al menos, los siguientes componentes fundamentales (Sharda, Dursun y Turban, 2017: 137-138):

- Fuentes del sistema. Internas, externas y personales, que proporcionan datos al *data warehouse* o *data mart*, según su caso.
- Integración de datos. Tecnologías y procesos que se necesitan para preparar los datos para su uso (sistemas ETL: procesos de "extracción", "transformación" y "carga" de los datos); es decir, se extraen los datos utilizando un software comercial denominado ETL o software escrito por el usuario (organización o empresa).
- Arquitectura de almacenamiento de datos para su almacenamiento en el *data warehouse* o *data mart* de una organización.
- Herramientas y aplicaciones para los diferentes tipos de usuarios, que deberán aprender a utilizar o, en su caso, desarrollar.
- Acceso a los datos (*middleware*). Herramientas que facilitan el acceso al *data warehouse*. Actúan de intermediación entre los dispositivos de almacenamiento y los usuarios que utilizan los datos con las aplicaciones y herramientas adecuadas.
- Metadatos, calidad de datos y procesos de gobierno, que deberán cumplir con sus especificaciones y asegurar que los almacenes de datos cumplen sus objetivos. Debido a sus características especiales entre tecnologías y políticas de gobierno de TI, les dedicaremos unos apartados especiales.

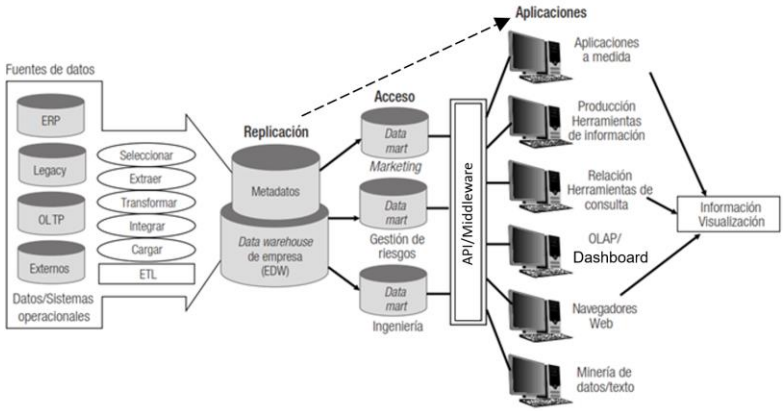



**Figura 9.3. Marco de trabajo de un sistema de *data warehouse***
**Fuente: (Sharda, Dursun y Turban, 2017: 138) traducida y adaptada**

### 9.3.2 FUENTES DE DATOS

Normalmente, los datos procederán de múltiples fuentes: sistemas operacionales/transaccionales OLTP (*On Line Transaction Processing*: procesamiento de transacciones en línea), sistemas de planificación de recursos ERP, datos de sitios web, datos propios o ajenos heredados, *legacy* (datos de estadísticas del gobierno, de la Unión Europea, de Eurostat, de la oficina del Censo de Argentina), datos de terceras partes. La tendencia actual de Big Data es incluir datos procedentes de sensores, etiquetas (*tags*), chips, RFID y NFC, geolocalización. Estos sistemas de fuentes utilizarán diferentes sistemas de software como Oracle, SAP, IBM, y almacenarán los datos en diferentes formatos (bases de datos relacionales, multidimensionales, jerárquicas).

Las fuentes de datos más típicas de una empresa son las bases de datos operacionales, bases de datos relacionales (bidimensionales), y *data warehouse*; además, utilizarán sistemas multidimensionales (bases de datos multidimensionales y OLAP).

Un tema que deben afrontar las empresas es la redundancia de los datos: tienen muchos sistemas que contienen los mismos datos; en estos casos, las empresas deben seleccionar cuál es la mejor fuente o el mejor sistema.

Otro concepto importante para tener en cuenta es la granularidad, es decir, el nivel de detalle con el que se necesitan los datos; no se requiere el mismo detalle en las transacciones diarias de un cliente que en un banco, o al momento de solicitar un préstamo personal o una hipoteca para comprar una casa.

### 9.3.3 CATEGORÍAS DE DATOS

En lo relativo a la alimentación al *data warehouse*, los datos se pueden agrupar en tres grandes categorías: datos internos, datos externos y datos personales (Vercellis, 2009).

**Datos internos**

Los datos internos almacenados en la mayoría de las bases de datos se conocen como sistemas transaccionales o sistemas operacionales y constituyen la espina dorsal (*backbone*) de un sistema de información de empresa. Los datos internos se recolectan de aplicaciones transaccionales en las operaciones rutinarias de una empresa (administración, contabilidad, logística). Este conjunto de aplicaciones de software se denomina ERP (*Enterprise Resources Planning*).

Los datos almacenados en los sistemas operacionales, normalmente, tratan con las entidades principales implicadas en los procesos de negocio de las empresas: clientes, productos, ventas, empleados, proveedores y socios. Estos datos provienen de diferentes componentes de un sistema de información:

- Sistemas de *back office*. Reúnen registros transaccionales básicos (pedidos, inventarios, producción, factura, y procesos logísticos).
- Sistemas de *front office*. Contienen datos que proceden de atención al cliente, actividades de centros de llamadas (*call centers*) y compañías de marketing.
- Sistemas basados en la web. Reúnen datos de ventas de sitios web en operaciones de comercio electrónico, visitas a sitios web, historial de visitas.

**Datos externos**

Los datos externos pueden proceder también de diferentes fuentes y se utilizan para ampliar la información extraída de las bases de datos internas y considerar la situación actual y futura de la empresa. Por ejemplo, datos de gran interés son los datos estadísticos publicados por organizaciones nacionales e internacionales, consultoras y bancos que muestran tendencias de mercado, hábitos de compras, entre otros.

Hoy en día existen numerosas y significativas fuentes externas de información: Sistemas de Información Geográfica (SIG o GIS, *Geographic Information Systems*), cuyas aplicaciones facilitan la adquisición, organización, almacenamiento y presentación de datos del territorio y, esencialmente, datos de posición geográfica específicos (coordenadas de latitud y longitud) de la organización, departamentos, sucursales, y, fundamentalmente, datos de geolocalización y sistemas GPS.

Otras fuentes de información vitales son los datos procedentes de medios sociales (redes sociales generalistas, como Facebook, Twitter o Tuenti, o redes sociales profesionales como Linkedin, Vimeo, Xing).

**Datos personales**

Cada día es más frecuente el análisis de la información almacenada en hojas de cálculo, bases de datos locales, mensajes de textos, boletines de noticias, archivos y documentos adjuntos profesionales en correos electrónicos. También, es habitual la integración de información operacional de la compañía con datos estructurados y no estructurados de empleados, clientes, proveedores (naturalmente, con el cumplimiento estricto de las leyes de protección de datos y privacidad vigente) que permitan su conversión en conocimiento para la toma de decisiones. En otras palabras, los datos personales formarán parte sustancial del sistema de gestión del conocimiento de la empresa y el uso adecuado de herramientas colaborativas de flujo de trabajo condicionará la mejora en los procesos de tomas de decisiones.

### 9.3.4 INTEGRACIÓN DE DATOS

Dada la gran cantidad de fuentes de información distribuidas en diferentes lugares de una organización, en muchos casos, heterogénea, hace falta integrar toda esta

información así como promover y facilitar su acceso. Se necesita extraer los datos de las diferentes fuentes, transformarlos (limpiar) para hacerlos eficaces y cargarlos en un *data warehouse* o, en su defecto, un *data mart*. Este proceso, de modo tradicional, se ha conocido como ETL (*Extraction, Transform, Load*), pero cada día más, se comienza a utilizar el término *integración de datos* debido, precisamente, al número creciente de modos en que los datos fuente pueden ser manipulados, tanto por la secuencia de operaciones como por la disparidad de fuentes sociales, sobre todo, porque en muchos casos no pueden ser tratados por herramientas ETL tradicionales.

La integración de datos se puede conseguir mediante técnicas diferentes que se agrupan, fundamentalmente, en dos categorías: extracción de datos mediante consulta con SQL (el lenguaje de consulta) en bases de datos o mediante software comercial especifico de integración de datos (ETL). También, es posible un tercer método, que es el desarrollo propio a medida del usuario, aunque cada vez es menos utilizado, ya que las herramientas ETL comerciales, o herramientas de aplicaciones web (apps) para escritorio o teléfonos inteligentes y tabletas, son cada día más abundantes.

El objetivo fundamental de la integración de datos es transformarlos haciéndolos más fáciles de utilizar. Por ejemplo, datos procedentes de diferentes fuentes se pueden integrar alrededor de una clave común, como puede ser la identidad digital de un usuario o cliente (nombre de usuario y contraseña).

Cada día, es más frecuente que el usuario acceda a Internet y, en consecuencia, a las aplicaciones corporativas, a través de diferentes fuentes: correo electrónico, la web, contactos directos, teléfono celular inteligente, mensajería instantánea (bien mediante los mensajes SMS tradicionales o con aplicaciones de mensajería instantánea (MI) como WhatsApp, Viber, Telegram, Line, Spotbros, Hangout, WeChat) y tabletas. Imaginemos el caso de un banco: un cliente puede acceder directamente en la oficina de la empresa, mediante un cajero automático (ATM), por correo electrónico a través del sitio web del banco (con diferentes dispositivos), a través de las redes sociales del banco (Facebook, Twitter). Por eso, es necesario integrar los datos de las diferentes fuentes del banco en un *data warehouse*.

Por otra parte, la integración de datos puede implicar otros tipos de transformaciones, por ejemplo, en lo relativo a formatos, género, fecha (día, mes, año), profesión, o en cuanto a duplicación de datos del cliente, originados por la diversidad de métodos de acceso y diferentes aplicaciones de una misma organización. Estas integraciones de datos pueden exigir no solo diferentes formas de extracción, sino de limpieza y carga de los datos.

Otra cuestión clave es la duración del tiempo de actualización. En muchos negocios, comienza a requerirse que los datos se almacenen en los *data warehouse* cada muy poco tiempo, minutos o decenas de minutos. Imaginemos el caso de grandes almacenes (Walmart en Estados Unidos, El Palacio de Hierro en México, El Corte inglés en España, Carrefour en Europa), que pueden necesitar almacenar los datos de compra o venta en el *data warehouse* de la empresa cada

10 o 20 minutos, para que estén disponibles para su análisis. Es decir, en los últimos años, han nacido herramientas de *data warehouse* en tiempo real, que han facilitado la integración de datos procedentes de múltiples fuentes mediante la actualización de datos en períodos prácticamente de tiempo real.

### 9.3.5 ALMACENAMIENTO DE DATOS

Existen diferentes arquitecturas para el almacenamiento de datos que ayudan en la toma de decisiones. La arquitectura más usual es un *data warehouse* central de empresa (EDW*, Enterprise Data Warehouse*). El *data warehouse* se configura como una plataforma central para cada organización sin uso de *data mart*. Este es el enfoque de la mayoría de las organizaciones y en esta arquitectura los datos del *data warehouse* son accesibles por todos los usuarios.

La segunda arquitectura es la de *data marts* independientes. Esta arquitectura se suele adoptar cuando el costo de la implantación de un único *data warehouse* no es asumible por la empresa o porque así lo consideran sus directivos o responsables de los sistemas de información. Los datos se almacenan en una o diversas aplicaciones, como finanzas, ventas, operaciones o marketing. Esta arquitectura está centrada en las aplicaciones para el almacenamiento de datos, aunque los datos se pueden utilizar para otras aplicaciones de la organización mediante los diferentes conectores (aplicaciones de *middleware* o intermediarias). En este caso, normalmente se requieren procesos de programación de la propia empresa o del proveedor del software.

Los *data mart* independientes no suelen ser muy eficaces en las grandes organizaciones; sin embargo, cumplen muy bien las necesidades de las pequeñas organizaciones o las de un departamento en una gran empresa. Uno de los inconvenientes de los *data mart* procede de las posibles inconsistencias de los datos y de la utilización de sistemas de fuentes diferentes (que pueden contener diferentes datos para los mismos elementos, por ejemplo, la dirección del cliente o definición de las actividades con la propia organización que pueden no estar anotadas en su totalidad).

Existe una tercera arquitectura que algunos proveedores recomiendan y adoptan las organizaciones: *hub and spoke* (concentrar y hablar). Esta arquitectura almacena datos en un *data warehouse* central mientras mantiene *data marts* que obtienen sus datos del depósito central de datos. Como los *data mart* adquieren sus datos del almacén central pueden tener más fiabilidad y consistencia, precisamente, por proceder de una fuente central. En el caso de los *data mart* dependientes, los datos se suelen almacenar con formatos apropiados al uso que se les va a dar, de este modo, proporcionan respuestas más rápidas para consultas y ejecución de aplicaciones.

### 9.3.6 *MIDDLEWARE*

Las herramientas de *middleware* (intermediación o adaptación) facilitan el acceso al *data warehouse*. Dado que existen numerosas aplicaciones para que los usuarios puedan ver los resultados de los análisis o visualizaciones, se requiere un componente idóneo para conseguir que las diferentes herramientas de los usuarios puedan acceder sin problemas a los *data warehouse* y *data mart*.

### 9.3.7 USUARIOS

Una vez que los datos ya se han almacenado en el *data warehouse* o *data mart*, pueden acceder a ellos los diferentes usuarios del sistema de Inteligencia de Negocios (desarrolladores TI, trabajadores directamente relacionados con BI, analistas, directivos y gerentes, proveedores, clientes, etcétera).

Los beneficios de un sistema de almacenamiento de datos incluyen que:

- Los usuarios finales pueden acceder rápida y fácilmente mediante navegadores web, ya que estos datos están localizados en lugares concretos.
- Los usuarios finales pueden realizar un análisis exhaustivo de datos con algún método que anteriormente no fuera posible.
- Los usuarios finales obtienen una visión consolidada de datos organizados.

Estos beneficios pueden, además de mejorar el conocimiento de negocio, proporcionar ventajas competitivas y mejoras al servicio y satisfacción del cliente, facilitando la toma de decisiones. Los *data warehouse*, pese a sus muchas ventajas, también presentan inconvenientes que será preciso tener presentes. Según el proyecto específico, la construcción y el mantenimiento pueden resultar muy caros. Otro inconveniente es la incorporación de datos desde sistemas informáticos obsoletos, que puede resultar difícil y costosa. Por último, suele producirse un hecho incontestable: las personas de un departamento pueden ser reacias a compartir datos con otros departamentos por numerosas causas (esta situación se produce igual que sucede en algunos casos de gestión del conocimiento, donde las personas no desean compartir sus conocimientos con otros empleados).

## 9.4 METADATOS, CALIDAD Y GOBIERNO DE UN DATA WAREHOUSE

A fin de documentar el significado de los datos contenidos en un *data warehouse*, se recomienda establecer una infraestructura de información específica conocida

como metadato. Un metadato es un dato que describe otro dato, es decir, son datos acerca de datos.

Tanto el personal especialista en TI, que opera y gestiona el *data warehouse*, como los usuarios que acceden a los datos, necesitan metadatos. El personal de TI necesita información relativa a las fuentes de datos, bases de datos, tablas, uso de datos. Las necesidades de los usuarios incluyen definiciones de datos, herramientas disponibles de informes/consultas (*report/query*), distribución de informes e información de contactos para ayuda y seguridad.

Los metadatos indican para cada atributo de un *data warehouse* la fuente original de datos, su significado y las transformaciones a las que ellos se han sometido. La documentación proporcionada por los metadatos debe mantenerse constantemente actualizada, con el objetivo de reflejar cualquier modificación en la estructura del *data warehouse*. La documentación debe ser accesible directamente a los usuarios de almacenes de datos a través de un navegador web o de un tablero de control (*dashboard*), mediante los derechos de acceso que cada usuario establezca.

Vercellis (2009) considera que los metadatos deben realizar las siguientes tareas informativas:

- Una documentación de la estructura de los almacenes de datos: diseño, vistas lógicas, dimensiones, jerarquía, datos derivados y localización de todos los *data mart*.
- Una documentación de la genealogía de los datos, obtenida por el etiquetado de las fuentes de datos de las cuales se extrajeron, que describe cualquier transformación realizada en ellos.
- Un listado que mantenga las estadísticas del uso del *data warehouse*, que indique especialmente cuántos accesos a campos o vista lógica se han realizado.
- Una documentación del significado de los almacenes de datos con respecto al dominio de la aplicación, que proporcione la definición de los términos utilizados y que describa las propiedades de los datos, así como el propietario de los datos y las políticas de carga que se han utilizado.

### 9.4.1 CALIDAD DE LOS DATOS EN UN ALMACÉN DE DATOS

La calidad de los datos en el *data warehouse* debe ser la adecuada para el cumplimiento de las necesidades del usuario. Si no sucede así, los datos no serán fiables y al final no se utilizarán. Muchas empresas consideran que los datos en los sistemas de fuentes de datos son pobres y han de mejorarse antes de almacenarlos.

La necesidad de verificar, preservar y mejorar la calidad de los datos es una preocupación constante del responsable de diseño y actualización de un *data warehouse*. Los principales problemas que pueden comprometer la validez e integridad de los datos son: datos incorrectos, datos no actualizados y pérdida de datos.

Los factores importantes que pueden afectar a la calidad de los datos son:

- Precisión (*accuracy*). Los datos deben ser altamente correctos.
- Completitud o compleción. Los datos deben ser completos y se ha de procurar que no se produzcan pérdidas de valores.
- Consistencia. El formato y contenido de los datos debe ser consistente a través de las diferentes fuentes de datos, después de los correspondientes procedimientos de integración.
- Oportunos. Los datos deben actualizarse con frecuencia basados en los objetivos del análisis (actualizaciones regulares: diarias, semanales).
- No redundantes. Se debe evitar la repetición y redundancia de datos para no malgastar memoria y prevenir posibles inconsistencias.
- Significativos. Los datos deben ser significativos (relevantes) a las necesidades del sistema de análisis y toma de decisiones (Inteligencia de Negocio), con el objetivo de añadir valor real a la ejecución de todos los análisis posteriores.
- Accesibilidad. Los datos deben ser fácilmente accesibles por los analistas y las aplicaciones de apoyo a decisiones.

### 9.4.2 GOBIERNO DEL DATA WAREHOUSE

En las organizaciones, es necesario asegurarse que los diferentes sistemas de información que se nutrirán de los datos de los almacenes, desde los ERP hasta los sistemas de BI, cumplen sus necesidades. Por estas razones, se deben poner en marcha buenas prácticas de gobierno de TI y, en particular, las relacionadas con los negocios fundamentales de la organización.

El gobierno del sistema de *data warehouse* debe ser una parte del gobierno de TI y requiere que las personas, comités y procesos se realicen en el momento oportuno y cumpliendo las reglas de gobiernos de las aplicaciones correspondientes. Las organizaciones suelen crear comités de personal técnico y de negocios que priorizan los proyectos, asignan recursos y aseguran que los negocios y los sistemas de información, en particular los almacenes de datos, estén alineados. Es necesario que estos comités y los equipos de trabajo que nombren supervisen los diferentes proyectos, para asegurar que se cumplen en su totalidad y de modo eficaz. También es preciso que los equipos de trabajo operacional, con la supervisión de los comités citados, ejecuten las tareas de

creación de la definición de datos, identificación y resolución de los problemas de los datos. Todos los comités de empresa requieren la colaboración y las contribuciones especializadas del personal de TI y de los procesos de negocio.

## 9.5 HERRAMIENTAS ETL

La integración de los datos en un sistema de almacenamiento de datos, como ya se ha comentado, se realiza con tecnología de ETL (*Extraction, Transformation, Load*). ETL se refiere a las herramientas de software que se dedican a la ejecución automática de las tres tareas principales: extraer, transformar y cargar. El proceso de carga en un *data warehouse* requiere la realización de las tres fases citadas.

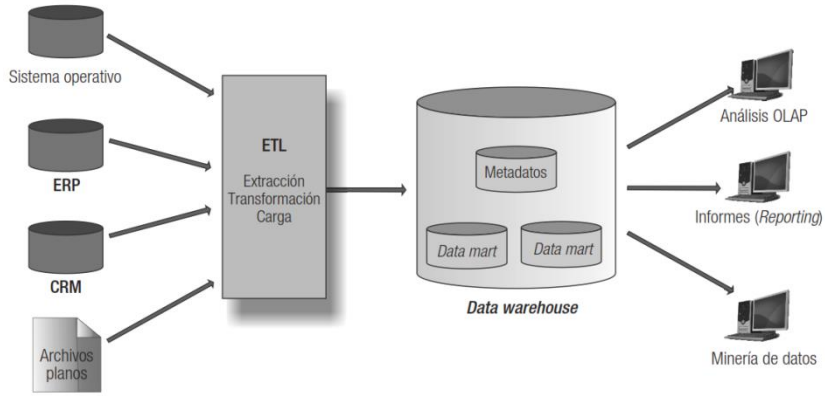



**Figura 9.4 Herramientas ETL**

### 9.5.1 EXTRACCIÓN

Los datos se extraen de las fuentes internas y externas disponibles (lectura de datos de una o más bases de datos). Las fuentes de datos pueden constar de archivos extraídos de bases de datos OLTP, hojas de cálculo, bases de datos personales (Oracle, Microsoft, Access) o archivos externos y archivos planos.

Es necesario considerar dos momentos en la extracción de datos: la extracción inicial, donde los datos disponibles relativos a períodos anteriores se introducen en el *data warehouse* vacío, y las extracciones incrementales posteriores que lo actualizan, utilizando nuevos datos disponibles a lo largo del tiempo.

### 9.5.2 TRANSFORMACIÓN

Es la conversión de los datos extraídos en su formato anterior al formato que se requiere para que se puedan situar en un *data warehouse* o, simplemente, en otra base de datos. Esta fase se conoce como transformación, pero en realidad se realizan dos tareas: transformación y limpieza de datos.

El objetivo de esta fase (transformación y limpieza) es mejorar la calidad de los datos extraídos de las diferentes fuentes mediante la corrección de inconsistencias, imprecisiones y pérdida de valores. Las principales que se eliminan durante la fase de transformación y limpieza de datos son:

- Inconsistencias entre valores registrados con atributos diferentes que tienen el mismo significado.
- Duplicación de datos.
- Pérdida de datos.
- Existencia de valores inadmisibles.

### 9.5.3 CARGA

Una vez que los datos se han extraído, transformado y limpiado, se deben cargar en el *data warehouse* para hacerlos disponibles a los analistas y las aplicaciones de apoyo a la decisión los utilicen.

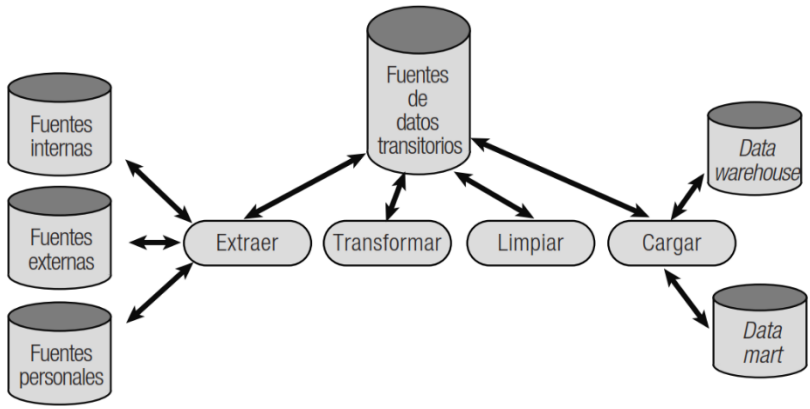



**Figura 9.5. Fases de un sistema ETL**

## 9.6 DESAROLLO DE UN SISTEMA DE *DATA WAREHOUSE*

Un proyecto de *data warehouse* es una actividad muy importante dentro de cualquier organización; es más complicado que muchos de los restantes proyectos

de computación, debido, fundamentalmente, a que afecta e influye en muchos departamentos e interfaces de entrada y salida. Además, son parte efectiva de los procesos de Inteligencia de Negocios de la compañía, desde los CRM de relación con los clientes hasta los sistemas de analítica como OLAP o de análisis de datos, pasando por los sistemas de información básicos de la empresa, como los sistemas ERP.

Los beneficios de los sistemas de almacenes de datos para las organizaciones son numerosos y los clasificaremos en beneficios directos e indirectos (Turban, 2011):

- Los usuarios finales pueden realizar análisis amplios desde numerosas vistas.
- Es posible una visión consolidada de los datos corporativos.
- La información es mucho mejor y oportuna.
- Se obtiene una mejora en el rendimiento o competencias del sistema.
- El acceso a los datos se simplifica.

Los beneficios indirectos se derivan del uso de los beneficios directos: mejora en el conocimiento del negocio, ventaja competitiva presente, mejora del servicio y de la atención al cliente, facilitación de la toma de decisiones y ayuda en las actualizaciones de los procesos de negocio. Por consiguiente, todos estos beneficios contribuyen a crear una ventaja competitiva con el uso del almacenamiento de datos.

Algunos autores añaden las siguientes características positivas de la utilización de un DW:

- Los usuarios finales pueden acceder a los datos que necesiten de un modo muy rápido y fácil, incluso vía navegadores web, ya que estarán situados en servidores web.
- Los usuarios finales pueden realizar análisis amplios de los datos en formas que no eran posibles anteriormente con los sistemas tradicionales.

### 9.6.1 TIPOS DE *DATA WAREHOUSE* POR SU FUNCIONALIDAD

Van den Hoven realizó una clasificación de los *data warehouse* que, pese a la antigüedad de su publicación, conserva en la actualidad toda su vigencia. Clasifica los almacenes de datos en tres categorías diferenciadas. La primera son los almacenes de datos que realizan funciones como localización, extracción, transformación, limpieza, transporte y carga de los datos. La segunda categoría es como herramienta de gestión de datos: un motor de base de datos que almacena y gestiona los *data warehouse* así como los metadatos. La tercera categoría es una herramienta de acceso a datos que proporciona a los usuarios finales acceso para el análisis de los datos. Esta última categoría incluye generadores de consultas, visualización, OLAP y Minería de Datos.

# CAPÍTULO 15
# ÉTICA, PRIVACIDAD, PROTECCIÓN DE DATOS Y COMPLIANCE EN LA CIENCIA DE DATOS: NORMAS LEGALES Y REGULACIONES, CONSORCIOS Y ORGANIZACIONES

## ANEXOS A

### GUÍA DE PRIVACIDAD DE INCIBE Y DE LA AEPD

**Guía sobre privacidad digital en Internet**

AEPD/INCIBE/OSI (2016). *Guía de Privacidad y Seguridad en Internet* (23 pp.) Disponible en: <https://www.aepd.es/media/guias/guia-privacidad-y-seguridad-en-internet.pdf>.

"El derecho a la privacidad se ha visto afectado por la irrupción de las nuevas tecnologías. El desarrollo de Internet y la era digital ha dado paso a nuevos retos en la protección de datos de usuarios. En esta guía sobre privacidad digital te contamos todo lo que debes saber sobre este delicado tema".

*Qué es la privacidad digital*

En términos generales, la definición de privacidad digital se entiende como el control que un usuario de Internet puede ejercer sobre sus datos, limitando el acceso de otras personas o instituciones a su información privada.

El significado de la privacidad digital ha ido evolucionando con el paso del tiempo. Los requisitos en materia de seguridad y protección de datos privados han ido aumentando con el desarrollo de Internet y la era digital. A ello se unen las nuevas exigencias que introduce el RGPD de aplicación en toda la Unión Europea.

*Privacidad en Internet*

La privacidad en la red siempre ha sido un tema polémico. Muchos usuarios desconocen cuáles son sus derechos sobre sus datos o quiénes están accediendo a su información personal. Por otro lado, algunas empresas necesitan para gestionar su privacidad digital abogados que les puedan aconsejar con el cumplimiento de la normativa.

El nuevo Reglamento General para la Protección de Datos hace hincapié en la regulación del entorno digital. Con esta nueva ley se pretende armonizar toda la normativa europea relativa a privacidad digital en Internet.

*Páginas web*

Una página web es un documento digital que se aloja en un servidor y muestra al usuario información en forma de texto, fotografías, videos. Existen muchos tipos de webs: páginas corporativas, tiendas online, blogs personales, medios de comunicación especializados.

Todas las páginas web es que deben cumplir con los aspectos legales de la privacidad digital. Cualquier empresa online debe informar al usuario sobre el tratamiento de sus datos. El individuo debe dar su consentimiento explícito y comprobable para el uso de su información personal.

Los requisitos legales en privacidad digital que deben cumplir las páginas web vienen reflejados en tres apartados obligatorios: aviso legal, política de privacidad y política de cookies.

*Aviso legal*

El aviso legal se refiere a los términos y condiciones de uso de una web. En él se explica el objetivo de la página, sus contenidos o su modo de funcionamiento. Tanto el propietario de la página como los visitantes deben respetar estas normas.

*Política de privacidad*

Si una página web tiene registro, estará tratando con la información personal de sus usuarios. Para establecer un límite al uso que una empresa puede hacer de estos datos, existe la política de privacidad.

Para cumplir con las políticas de privacidad, se debe indicar a la AEPD el tratamiento que se hará de los ficheros. También se debe redactar un texto legal en un apartado visible de la web. Las políticas de protección de datos y privacidad han de ser aceptadas por el usuario mediante una casilla de verificación.

Cualquier persona puede denunciar una web si no cumple con las políticas de privacidad digital. Las multas pueden llegar a 600.000 euros.

*Cookies*

Las *cookies* son archivos que se guardan en el ordenador al navegar por determinadas páginas web. Estos archivos permiten realizar un seguimiento del comportamiento de los usuarios en Internet, normalmente con objetivos comerciales.

Toda página web debe informar el usuario de su política de *cookies*. Los visitantes deben dar su consentimiento expreso para su colocación. No cumplir con la normativa puede conllevar sanciones de 150.000 euros.

*Redes sociales*

No cabe duda de que la privacidad de la información digital se ha visto comprometida por la irrupción de las redes sociales. Estas plataformas han cambiado la forma en que la gente se relaciona en Internet. Las redes sociales ofrecen muchas ventajas para la comunicación, pero también tienen algunos riesgos para la privacidad digital.

¿Cómo se trata la información de los usuarios en redes sociales como Facebook, Twitter o Instagram? ¿Dónde realizar los ajustes de privacidad?

# ANEXO B

## 5 DE NOVIEMBRE DE 2018. PRIMERA NORMA INTERNACIONAL ISO/IEC PARA INTERNET DE LAS COSAS

**Estándar, ISO/IEC 30141 Internet de las Cosas (IoT)-Arquitectura de Referencia**

El estándar ISO/IEC 30141 proporciona un vocabulario común en todo el mundo para diseñar y desarrollar aplicaciones de IoT, lo que permite desplegar sistemas fiables, seguros, protegidos, respetuosos con la privacidad y capaces de afrontar ciberataques.

Nuevo ejemplo de cómo la normalización contribuye a garantizar el éxito de la transformación digital. Más de 600 normas internacionales están en revisión para el correcto funcionamiento de la industria 4.0. En España se ha publicado la Especificación UNE 0060 Industria 4.0. *Sistema de gestión para la digitalización.* Requisitos para favorecer la digitalización de la industria española.

Desde los vehículos autónomos hasta la agricultura de precisión, pasando por la fabricación inteligente, la sanidad electrónica o las ciudades inteligentes; Internet de las Cosas *(IoT)* ya está presente en todas partes y va en aumento. Se trata de integrar los objetos cotidianos en los sistemas informáticos, permitiendo así que los dispositivos electrónicos interactúen con el mundo físico, directamente entre ellos, sin intervención de las personas.

Para garantizar la seguridad, confianza y una base tecnológica con medidas y sistemas robustos, la Organización Internacional de Normalización (ISO) y la Comisión Electrotécnica Internacional (IEC) han publicado el primer estándar internacional que proporciona una arquitectura de referencia de *IoT* estandarizada internacionalmente que utiliza un vocabulario común, diseños reutilizables y las mejores prácticas del sector.

Este estándar, denominado ISO/IEC 30141 Internet de las Cosas (IoT)- Arquitectura de Referencia, proporciona un marco común para los diseñadores y desarrolladores de aplicaciones de *IoT*, que permite desarrollar sistemas fiables, seguros, protegidos, respetuosos con la privacidad y capaces de afrontar interrupciones debidas a catástrofes naturales o ciberataques.

El estándar ISO/IEC 30141 ha sido desarrollado por el comité técnico conjunto de la Organización Internacional de Normalización y la Comisión Electrotécnica Internacional: ISO/IEC JTC 1, Tecnología de la información; subcomité SC 41: Internet de las Cosas y tecnologías relacionadas.

La industria española puede participar en estos desarrollos de normas internacionales a través de la Asociación Española de Normalización, UNE, el miembro español de ISO e IEC.

Este primer estándar internacional de Internet de las Cosas es un claro ejemplo de cómo la normalización contribuye eficazmente a garantizar el éxito de la transformación digital. En conjunto, más de 600 normas internacionales están en revisión para garantizar el correcto funcionamiento de la industria 4.0, en aspectos como interoperabilidad, conectividad, ciberseguridad, robótica avanzada o impresión 3D.

Además del estándar recién publicado, el grupo de trabajo de ISO/IEC está elaborando futuras normas como la que establecerá el marco de aplicación para la gestión de energía de la demanda de instalaciones industriales o la de red de sensores para medidores de gas inalámbricos, entre otras.

# CAPÍTULO 16
# TENDENCIAS TECNOLÓGICAS DISRUPTIVAS EN CIENCIA DE DATOS EN ELHORIZONTE 2025


**Figura 16.1. Principales tendencias tecnológicas estratégicas para 2022 de Gartner Fuente: https://www.gartner.es/es/tecnologia-de-la-informacion/insights/principales-tendencias-tecnologicas**

**Tendencia 1: Tejido de datos (Fábrica de datos)**

El tejido de datos proporciona integración flexible y resiliente de las fuentes de datos entre distintas plataformas y usuarios comerciales para que estén disponibles desde cualquier lugar, independientemente de dónde se alojen. La fábrica de datos puede utilizar analítica para aprender y recomendar de forma activa dónde se deben utilizar y cambiar los datos. Gartner considera que esta propiedad puede reducir hasta un 70 % los esfuerzos de gestión de datos.

**Tendencia 2: Malla de ciberseguridad**

La malla de ciberseguridad es una arquitectura componible y flexible que integra servicios de seguridad distribuidos y dispares. La malla de ciberseguridad permite que las mejores soluciones de seguridad independientes se utilicen conjuntamente para una mayor protección global y que los puntos de control se sitúen más cerca de los activos a los que deben proteger. Puede verificar de forma rápida y fiable la identidad, el contexto y el cumplimiento de políticas en todo tipo de entornos, sean de la nube o no.

**Tendencia 3: Computación de mejora de la privacidad**

La computación de mejora de la privacidad garantiza el tratamiento de datos personales en entornos que no son lo bastante seguros (algo cada vez más importante ante los cambios normativos en materia de privacidad y protección de datos, en virtud de la creciente preocupación de los consumidores). Para ello utiliza varias técnicas de protección que permiten obtener valor de los datos y cumplir al mismo tiempo los requisitos vigentes.

**Tendencia 4: Plataformas nativas de la nube**

Las plataformas nativas de la nube son tecnologías que permiten construir nuevas arquitecturas de aplicaciones resilientes, elásticas y ágiles para seguir el rápido ritmo del desarrollo digital; mejoran el enfoque tradicional de tipo *lift-and-shift* (o realojamiento), que no aprovecha todos los beneficios de la nube y añade complejidad al mantenimiento.

**Tendencia 5: Aplicaciones componibles**

Las aplicaciones componibles se crean a partir de componentes modulares centrados en la empresa, al facilitar la utilización y reutilización de código, lo que acelera el tiempo de comercialización de nuevas soluciones de software y libera valor empresarial.

**Tendencia 6: Inteligencia de decisiones**

La inteligencia de decisiones es un enfoque práctico que mejora la toma de decisiones en la organización. Cada decisión se modela como un conjunto de procesos, utilizando conocimientos y análisis para conformar las decisiones,

aprender de ellas y perfeccionarlas, incluso tiene el potencial de mejorarlas con análisis aumentados, simulaciones e IA.

**Tendencia 7: Hiperautomatización**

La *hiperautomatización* es un enfoque empresarial disciplinado que busca identificar rápidamente, investigar y automatizar procesos comerciales y de Tecnología de la Información como sea posible. Facilita el redimensionamiento, las operaciones remotas y la transformación radical del modelo de negocio.

**Tendencia 8: Ingeniería de IA**

La ingeniería de IA automatiza las actualizaciones de datos, modelos y aplicaciones, y combinada con una sólida gobernanza de Inteligencia Artificial, la ingeniería de IA hace operativa la entrega de la IA para asegurar su valor comercial continuado.

La Inteligencia Artificial madurará a través de múltiples disciplinas a medida que se necesite más automatización. DataOps, ModelOps y DevOps son los pilares de ingeniería de la IA. Una sólida estrategia en esta materia facilitará el rendimiento, la escalabilidad, la interpretabilidad y la fiabilidad de los modelos de Inteligencia Artificial, al tiempo que proporcionará un mayor valor a las inversiones realizadas en esta tecnología, según Gartner.

Se trata, apuntan desde la firma de investigación, de una tendencia importante teniendo en cuenta que solo el 53 % de los proyectos de Inteligencia Artificial pasan de la fase de prototipado a la de producción. Y, según los analistas de la consultora:

> *El camino hacia la producción de la Inteligencia Artificial significa recurrir a la ingeniería de esta tecnología, una disciplina centrada en la gobernanza y la gestión del ciclo de vida de una amplia gama de modelos operativos de estas herramientas, como el Aprendizaje Automático o los gráficos de conocimiento.*

**Tendencia 9: Empresas distribuidas**

Las empresas distribuidas son un reflejo del modelo de negocio basado en priorizar lo digital y lo remoto para mejorar las experiencias de los empleados, digitalizar los puntos de contacto de consumidores y socios, y desarrollar experiencias de producto. Este tipo de empresas pueden satisfacer mejor las necesidades de los empleados y los consumidores remotos que están disparando la demanda de servicios virtuales y lugares de trabajo híbridos.

**Tendencia 10: Experiencia total**

La estrategia total es una estrategia empresarial que integra la experiencia del empleado, la experiencia del cliente, la experiencia del usuario y la multiexperiencia en varios puntos de contacto para acelerar el crecimiento e incrementar la confianza, la satisfacción, la fidelidad y la defensa de los clientes y empleados a través de una gestión holística de las vivencias de las partes interesadas.

**Tendencia 11: Sistemas autonómicos**

Los sistemas autonómicos, que pueden estar gestionados físicamente o mediante software, son capaces de aprender de su entorno y modificar dinámicamente sus propios algoritmos en tiempo real para optimizar su comportamiento en ecosistemas complejos. Los sistemas autonómicos crean un conjunto ágil de capacidades tecnológicas que pueden ayudar en caso de nuevos requisitos y situaciones, optimizar el rendimiento y defender contra ataques sin necesidad de intervención humana.

**Tendencia 12: Inteligencia Artificial Generativa**

La Inteligencia Artificial Generativa utiliza datos para aprender cómo son los artefactos y generar nuevas creaciones innovadoras que son parecidas al original, pero no una simple repetición; posee el potencial de producir nuevas formas de contenido creativo, por ejemplo videos, acelerar los ciclos de I+D en ámbitos que van desde la medicina hasta el diseño de productos.

**NOTA.** La mayoría de estas tendencias son fuente para muchos de los informes de tendencias de Ciencias de Datos, Inteligencia Artificial y Aprendizaje Automático que se publican a finales del 2021 o principios de 2022, como veremos en el siguiente apartado.

## 16.1.1. PRINCIPALES TENDENCIAS TECNOLÓGICAS ESTRATÉGICAS PARA 2022 (GARTNER)

En octubre del 2021 se presentó un estudio donde se identifican las tendencias tecnológicas que resultan esenciales para el negocio digital en el año 2022 y siguientes. Las principales tendencias tecnológicas estratégicas para 2022 (Figura 16.1) de Gartner[8] están en su gran mayoría relacionadas con la Ciencia de Datos y la Inteligencia Artificial, disciplinas afines como nube nativa, ciberseguridad, Inteligencia de Negocios. A continuación, se describen, según Gartner, las tendencias tecnológicas citadas en el informe

[8] *Disponible en: <https://www.gartner.es/es/tecnologia-de-la-informacion/insights/principales-tendencias-tecnologicas>. La consultora Gartner presenta las tendencias y realiza un análisis breve pero de gran interés. En esta dirección, también de modo gratuito, un e-book síntesis del informe completo.*

# GLOSARIO

**Bloque**. Agrupación de transacciones individuales en una cadena de bloques. En el caso de los bitcoines, todas las transacciones se comprueban, ordenan y almacenan en un bloque que se une al bloque anterior, creándose así una cadena. Cada bloque debe referirse al bloque anterior para ser válido. Esta estructura registra fielmente el momento de las transacciones y las almacena evitando que nadie pueda alterar el registro.

**Cadena de bloques**. Es un conjunto de computadores o servidores denominados nodos, que conectados en red utilizan un mismo protocolo o sistema de comunicación con el objeto de validar y almacenar la información registrada en una red P2P.

**Contrato inteligente**. Programa de computadora que controla directamente la transferencia de monedas o activos digitales entre partes bajo ciertas condiciones, almacenados en la tecnología *Blockchain*.

**Criptomoneda**. Moneda virtual que utiliza técnicas de criptografía para controlar cuándo se generan las unidades de la divisa y garantizar la transferencia segura de fondos.

**Libro de contabilidad distribuido (distributed ledger).** Lista de transacciones con marcas de tiempo que se difunde, copia y confirma simultáneamente a través de múltiples computadores de una red P2P.

**Método aleatorio o dispersión (hash/hashing**). Método criptográfico que usa una función llamada *hash* que resume cualquier cantidad de datos en una cadena alfanumérica de longitud fija. La transformación de una cadena de caracteres en un valor normalmente más corto, de longitud fija, o una clave que representa la cadena original (similar a la creación de un acortador de direcciones URL de la web como bitly.com).

**Minería**. Proceso computacional necesario que opera para asegurar su red. Los nodos de una red de criptomoneda compiten entre sí para añadir de modo seguro nuevos bloques a la cadena.

**Nodo**. Un computador conectado a una red P2P (*peer-to-peer*) con diferentes capacidades de cómputo. Las criptomonedas como bitcoin y Ethereum se componen de miles de nodos repartidos por todo el mundo. Todos los nodos han de poseer el mismo software y protocolo para comunicarse entre sí. En una *Blockchain* pública los nodos no tienen por qué identificarse, mientras que en una privada, los nodos se conocen entre sí y pueden ser iguales entre ellas.

**Redes igual a igual (P2P, peer-to-peer)**. Red descentralizada de computadores conectados directamente entre ellos, que se pueden comunicar entre sí directamente, sin pasar por un servidor central ni por un administrador. Napster fue una de las primeras redes P2P; en la actualidad BitTorrent es una de las más populares.

**Sistema descentralizado**. Todos los computadores conectados a la red son iguales entre sí y controlan dicha red. No existe una jerarquía entre los nodos en el caso de una *Blockchain* pública, pero si puede existir jerarquía en una privada. En un sistema centralizado, toda la información está controlada por un único computador o servidor.