

APRENDIZAJE AUTOMÁTICO (MACHINE
LEARNING)
Y
CIENCIA DE DATOS (DATA SCIENCE)

UPV (VALENCIA)

Benito J. Palacios Cerro

Diciembre, 24th 2023

Contents

Contents	1
1 Módulo 1: Introducción a la minería de datos y ciencia de datos	3
1 Motivación	3
2 Minería de Datos y Ciencia de Datos	5
2.1 Agente bancario	6
2.2 Gerente supermecado	7
2.3 Gerente de personal	8
2.4 Supervisor de fábrica	9
2.5 Papel de la inteligencia empresarial	10
2.6 Minería de datos	12
2.7 Almacenamiento de datos multidimensional	13
2.8 OLAP vs Minería de Datos	17
2.9 Ejemplos de Ciencia de Datos	19
3 Proceso de extracción de conocimiento	21
3.1 Estándar CRISP-DM	23
4 Tareas, técnicas y herramientas	25
4.1 Tareas	25
4.2 Técnicas	27
4.3 Herramientas	32

2	Módulo 2: Evaluación de modelos de aprendizaje automático	34
1	Métricas de Clasificación	34
2	Métricas para regresión	36

Chapter 1

Módulo 1: Introducción a la minería de datos y ciencia de datos

1

Motivación

- Motivación
 - ¿Qué es esto? Sopa de letras.
- Minería de datos y ciencia de datos
 - Ejemplos de minería de datos
 - El papel en la inteligencia empresarial
 - Almacenaje de datos y OLAP
 - Ejemplos de ciencia de datos
- El proceso de extracción de conocimiento
 - El proceso D2K
 - CRISP-DM
- Tareas, técnicas y herramientas

En general de lo que se trata es de la **transformación de datos en conocimiento**, partimos de datos que pueden estar en distintos formatos, el objetivo principal es que por medio de ciertos procedimientos convertir esos datos en conocimiento. Este conocimiento es el que nos puede servir en las organizaciones, lo que referimos como *Inteligencia Empresarial*, que puede ser útil para determinar que un intruso que entra en una red o en

un entorno físico, se considere como peligroso o no, y actuar en función del conocimiento o de los modelos que se extraen de los datos.

Ahora veamos definiciones:

Minería de Datos , está asociada a herramientas e inteligencia empresarial.

Ciencia de Datos , incluye cualquier acto, procedimiento o utilidad que hagamos sobre los datos.

Análisis de Datos (Inteligente) , proveniente de la estadística, se restringe a la parte analítica de los datos.

Análisis predictivo , es un tipo de *análisis de los datos*, existen más tipos de análisis como los *descriptivos* o la simple visualización de éstos.

Big Data , incluye una serie de aspectos que no son de análisis sino más bien de gestión de la información de grandes volúmenes de datos y a grandes velocidades.

Extracción de conocimiento, desde bases de datos , es la base de la transformación de los datos a conocimiento.

El **aprendizaje automático** aparece en todas las áreas que hemos visto como una herramienta muy flexible a la hora de extraer conocimiento a partir de los datos.

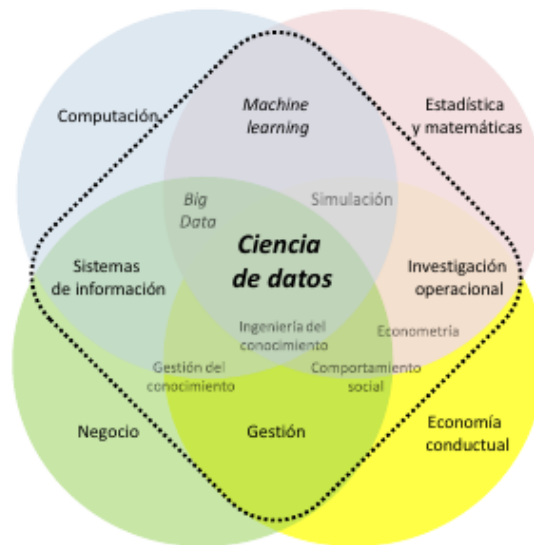
Para aclarar la diferencia entre *Big Data* y *Ciencia de Datos*, es que **no** todo lo que se hace en ciencia de datos tiene que ser con grandes volúmenes de datos, a gran velocidad o con gran variabilidad.



- Podemos tener un conjunto de datos pequeños, por ejemplo, pruebas diagnósticas de un número reducido de pacientes, y analizar esa información. Formaría parte del *Análisis de Small Data*.
- Podemos tener millones de registros de pacientes con medicamentos y tratamientos, de los que queremos analizar esos datos, hablaríamos de *Análisis de Big Data*.

- Por último, podemos tener los datos de todas las compras que se han realizado en un país durante unos cuantos meses o años y visualizar esos datos o resumirlos o hacer un informe y tener una gran complejidad a la hora de integrar esos datos o transformarlos, pero sin un análisis, hablaríamos de *Big Data*.

Es importante ver una serie de términos que están relacionados cuando hablamos de ciencia de datos y el papel del machine learning junto a la ciencia de datos.



2

Minería de Datos y Ciencia de Datos

Para entender lo que es la *ciencia de datos*, centrándonos particularmente en la *minería de datos*, la transformación de datos a conocimientos, vamos a ver cuatro ejemplos:

- Agente bancario. ¿Debo ofrecerle una hipoteca al cliente?. Veremos que con determinadas herramientas, que extraen conocimiento, modelos a partir de datos podremos resolver la pregunta.
- Gerente de supermercado. ¿Cuando los clientes compran huevos, también comprarán aceites?. Aquí veremos que podemos utilizar herramientas, de minería de datos, no tanto de análisis descriptivo o de aprendizaje automático.
- Gerente de personal. ¿Qué tipo de empleados tengo?.
- Supervisor de fábrica. ¿Cuántos fallos, para el módulo *X* esperamos cada mes?, o ¿Cuántas ventas vamos a hacer?.

2.1 Agente bancario

Cuando un cliente solicita una hipoteca o un préstamo de autoconsumo, la respuesta debe ser si se le concede o no. Para responder a esta situación, lo que le interesa al banco es saber si *los clientes anteriores han sido buenos clientes o malos para el banco*.

• AGENTE BANCARIO:

¿Debo ofrecerle una hipoteca a este cliente?

Histórico:

cld	Crédito-p (años)	Crédito-a (euros)	Salario (euros)	Tiene casa	Cuentas adeudadas	...	Devolución crédito
101	15	60.000	2.200	sí	2	...	no
102	2	30.000	3.500	sí	0	...	sí
103	9	9.000	1.700	sí	1	...	no
104	15	18.000	1.900	no	0	...	sí
105	10	24.000	2.100	no	0	...	no
...

Minería de datos

Patrón / Modelo:

```
If Cuentas adeudadas > 0 then Devolución crédito = no  
If Cuentas adeudadas = 0 and [(Salario > 2.500) or (Crédito-p > 10)] then Devolución crédito = sí
```

Un cliente que no devuelve un crédito es un *mal cliente para el banco*, pero puede haber otros conocimientos, como por ejemplo, un mes no se paga la mensualidad, no es un problema muy grande si después el cliente se recupera y puede ser incluso beneficioso para el banco. Lo que debe tener registrado el banco no es sólo si se devolvió o no el crédito, sino si realmente ha sido beneficioso o no para el banco.

Sit tuviéramos miles y miles de clientes anteriores a los que hemos concedido un préstamo, podríamos saber si al final, y al cabo de un cierto tiempo, esa persona ha tendido problemas o no a la hora de devolver el crédito.

Cuando un cliente solicita un crédito suelen hacérsele una serie de preguntas: cuanto tiempo durará el crédito, cuanta cantidad se solicita, que ingresos mensuales tiene, así como otro tipo de preguntas sobre su entorno familiar o si tiene otro tipo de propiedades que pudieran utilizarse como aval. La table mostrada está bastante simplificada, en relación al tipo de preguntas que se se pueden realizar.

Toda esta información se registra en una base de datos relacional. Ahora recibimos un nuevo cliente, que nos pide unos años, nos indica el dinero que quiere pedir y le hacemos otra serie de preguntas, lo que no sabemos es qué le responderemos. ¿Cómo podemos responder a esta petición?. Para ello lo que haremos será transformar los datos de la tabla en un modelo que a partir de las entradas del nuevo cliente nos de la predicción de si debemos conceder o no el crédito.

En este caso estamos realizando una **tarea predictiva**, y en este caso de **clasificación**,

donde los datos de entrada se corresponden con las conlumnas de la tabla, salvo la última columna, «*devolución crédito*», que se corresponde con nuestra variable de salida o predicción.

Como ejemplo, el modelo nos indica las siguientes reglas:

- que si *las cuentas adeudar son mayores que cero, entonces, no devolverá el crédito*
- *si las cuentas adeudadas son igual a cero y el salario es mayor de 2.500€ o los años son mayores que diez, entonces, normalmente se devuelve el crédito.*

Se trataría de un modelo muy simplista, pero nos da una solución al problema.

2.2 Gerente supermercado

Es un caso muy general en el ámbito del análisis de datos, puede ocurrir en cualquier tienda online o cualquier medio por el que se compran productos conjuntamente.

• GERENTE DE SUPERMERCADO:

Cuando mis clientes compran huevos, ¿cogen también aceite?

Histórico:

Nº cesta	Huevos	Aceite	Pañales	Vino	Leche	Mantequilla	Salmón	Endivia	...
1	sí	sí	no	sí	no	sí	sí	sí	...
2	no	sí	no	no	sí	no	no	sí	...
3	no	no	sí	no	sí	no	no	no	...
4	no	sí	sí	no	sí	no	no	no	...
5	sí	sí	no	no	no	sí	no	sí	...
6	Sí	no	no	sí	sí	sí	sí	no	...
7	no	no	no	no	no	no	no	no	...
8	sí	sí	sí	sí	sí	sí	sí	no	...
...

Patrón / Modelo:

Minería de datos

Huevos → Aceite: Confianza = 75%, Apoyo = 37%

Uno puede preguntarse, ¿Qué productos se compran conjuntamente? o cuando uno sale con la cesta, qué suele acompañar a éstos productos. En principio, si son pocos productos, se puede observar directamente, pero si son muchísimos productos, como un supermercado que puede tener miles de productos en su stock, si queremos ver que productos se compran conjuntamente, se pueden buscar todas las compras hechas durante y una semana, que previamente se ha registrado en una tabla como la de la imagen, y comprobar si los «*síes*» de una columna están asociados con los «*síes*» de la otra columna.

En la table vemos que aceites y pañales no estarían demasiado asociados.

Lo que queremos es responder a la pregunta general ¿qué productos se compran conjuntamente? Esto nos obliga a ver todas las combinaciones posibles de pares entre los miles de productos; para ello podemos recurrir a *herramientas de minería de datos* que lo que hacen es extraer todas las asociaciones, relaciones entre atributos que son muy frecuentes.

Cuando ejecutamos el análisis de datos, en relación con la pregunta "*cuando los clientes compran huevos, ¿compran también aceite?*", obtendremos un patrón que dice que cuando se compran huevos, entonces se compra aceite y la confianza es un 75%, quiere decir que de 100 compras de huevos, en 75 ocasiones también se compra aceite. Aún así, puede que ambos productos estén muy relacionados pero sean poco frecuentes en conjunto.

En el caso visto, los atributos de la tabla son todos de entrada, no hay variables de salida, en este caso hemos obtenido un modelo que establece una regla o patrón que nos indica la relación entre dos productos.

2.3 Gerente de personal

En este caso, un gerente de personal se pregunta *¿qué empleados tengo?*. De los empleados a su cargo conoce algunos, pero dependiendo del número de empleados de la empresa es probable que desconozca a la gran mayoría. Necesito información, datos, para tomar decisiones respecto a la reestructuración de la empresa, si tengo que formarlos de alguna manera o cualquier otra actividad que les afecte.

• GERENTE DE PERSONAL:

¿Qué tipo de empleados tengo?

Histórico:

Id	Salario	Casado	Coche	Niños	Alquilado/ Propietario	Sindicado	Bajas/año	Años de trabajo	Género
1	10000	si	no	0	alquilado	no	7	15	M
2	20000	no	si	1	alquilado	si	3	3	F
3	15000	si	si	2	propietario	si	5	10	M
4	30000	si	si	1	alquilado	no	15	7	F
5	10000	si	si	0	propietario	si	1	6	M
6	40000	no	si	0	alquilado	si	3	16	F
7	25000	no	no	0	alquilado	si	0	8	M
8	20000	no	si	0	propietario	si	2	6	F
15	8000	no	si	0	alquilado	no	3	2	M
...

Patrón / Modelo:

Minería de datos

- **Grupo 1:** Sin niños y alquilados. Baja participación sindical. Muchos días de baja.
- **Grupo 2:** Sin niños y con coche. Alta participación sindical. Pocos días de baja. Más mujeres en casas alquiladas.
- **Group 3:** Con niños, casados, con coche. Mayormente hombres y propietarios de casa. Baja participación sindical.

Toda empresa suele tener una base de datos con la información importante relacionada con sus empleados, como puede ser: el salario, si tiene coche, si está casado, ... A partir de esa información, quiero poder conocer **qué tipologías o grupos de empleados tiene la empresa**.

A simple vista podríamos tener una idea general, pero la gradación que puede haber de unos empleados a otros, teniendo atributos similares, es tanta que nos perderíamos y no concretaríamos dichas tipologías. Para ayudarnos hay herramientas de **aprendizaje automático, de agrupamiento, de clustering** que nos permitirán, a partir de los datos de la tabla, todos los atributos son de entrada, desarrollan un algoritmo que nos generará tres grupos, los cuales agrupan registros que se parecen y difieren de los registros de otros grupos.

En el ejemplo vemos que los tres grupos se componen:

- **Grupo 1:** sin hijos, con vivienda en alquiler, baja participación sindical y con muchos días de baja.
- **Grupo 2:** sin hijos, que tienen coche, alta participación sindical, pocos días de baja y más mujeres que hombres en casas alquiladas.
- **Grupo 3:** con hijos, casados, con coche, mayoritariamente hombre y con vivienda en propiedad, baja participación sindical.

La herramienta crea los grupos, en función de nuestras necesidades.

2.4 Supervisor de fábrica

Aquí puede interesarnos contestar a la pregunta *¿Cuántos fallos para el módulo «X» esperamos cada mes?*. Disponemos de una serie de datos relativos a ese módulo, entre ellos como ha ido fallando.

• SUPERVISOR DE FÁBRICA:

¿Cuántos fallos para X módulo esperamos cada mes?

Histórico:

Módulo (F*)	Mes-12	...	Mes-4	Mes-3	Mes-2	Mes-1	Mes
X	20	...	52	14	139	74	?
Y	11	...	43	32	26	59	?
Z	50	...	61	14	5	28	?
W	3	...	21	27	1	49	?
R	14	...	27	2	25	12	?
...

Minería de datos

Patrón / Modelo:

Modelo lineal: Fallos de X para el siguiente mes:

$$FX(\text{Mes}) = 0.62 \cdot FX(\text{Mes-1}) + 0.33 \cdot FX(\text{Mes-2})_X + 0.12 \cdot FZ(\text{Mes-1}) - 0.05$$

Se trata de un problema más complejo que los anteriores ya que involucra el *tiempo*, los

otros ejemplos vistos no son atemporales, pero en éste tendremos información relativa a los fallos que cada módulo ha podido tener cada mes.

Muchas veces las predicciones, sobre todo relativas al tiempo, se basan en los valores, no sólo del tiempo inmediatamente anterior, sino de valores agregados en tiempos o en ciclos anuales, mensuales, ... A partir de aquí, con el procesamiento de los datos, indicándole qué queremos predecir, los fallos para el mes que viene, partimos de la información sobre fallos de los meses anteriores, información que puede depender del propio módulo, pero también de fallos de otros módulos, con lo que la información que podemos incluir para predecir el valor de X , puede depender de sus valores, pero también de la información de otros registros del resto de módulos.

En este caso, utilizando la tabla referida y mediante una herramienta de minería de datos, con una preparación concreta, procesaremos los atributos necesarios como entradas, para obtener la última columna que se corresponde con la salida tras el procesamiento.

En este caso, podríamos utilizar una función lineal como la que se indica a continuación, en concreto *regresión lineal*, que predice un valor numérico, los fallos de el módulo X :

$$FX(Mes) = 0.62 * FX(Mes_1) + 0.33 * FX(Mes_2) + 0.12 * FZ(Mes_1) - 0.05$$

- Mes_1 , mes anterior al que queremos predecir.
- Mes_2 , dos meses antes.
- FX referido a los fallos de otros módulos.
- Los coeficientes: 0.62, 0.33, 0.12 y 0.05, son proporcionados por la herramienta de minería utilizada.

El modelo es **predictivo**, pero en este caso es un **modelo de regresión**.

2.5 Papel de la inteligencia empresarial

Inteligencia Empresarial es un concepto que engloba una serie de herramientas que intenta extraer conocimiento y tomar decisiones a partir de la información. Son herramientas que se estructuran en varios niveles, se conoce como estructura piramidal, desde los datos más básicos hasta el conocimiento.



Un ejemplo sencillo sería que *quiero reservar un vuelo desde Valencia a Montevideo, con varias escalas*, lo que haré será mirar a través de una aplicación, que me solicitará el origen, el destino, los días, etc; esta aplicación lo que hará es localizar los vuelos que tiene en su base de datos que coincidan con los datos aportados y si hay espacio en ellos proporniendolos como rutas posibles. Una vez que quiero realizar la transacción, la reserva, la aplicación deberá comprobar en todo momento que sigue habiendo plazas disponibles, y finalizada la transacción, si había plaza, confirmar la transacción.

La decisión que toma el sistema **no es estratégica** es una decisión a nivel **operacional**, que los sistemas pueden tomar en todo momento basándose en información, que normalmente se recoge en bases de datos, que internamente se consultan y modifican a través de lenguajes de gestión de bases de datos, como SQL.

Todo el proceso visto constituiria la base de la piramide, la interacciones más sencillas entre un dispositivo o aplicación y una persona.

Por encima del **nivel operacional** hablaríamos del **nivel táctico**, en él podemos hacer consultas mucho más complejas, como el informe que llega a la mesa de algunas personas, que toman decisiones en la organización, como pueden ser las ventas de la última semana, mes o año, desglosados por zonas geográficas y por tipos de productos. Son informes que van más allá de lo que es el dato preciso o concreto, no es información hipotética, sino de hechos ocurridos y, a partir de los cuales, tomar decisiones. Esta información se puede obtener a partir de consultas complejas, mediante instrucciones de lenguaje de programación como *group by*, agrupamiento, información que suele acompañarse de gráficos.

La información obtenida a nivel táctico podemos transformarla en **conocimiento** y observar si las ventas han ido como queríamos o si hay algo sorporendente.

El **nivel de planificación**, está por encima del nivel táctico, e incluye decisiones más complejas, por ejemplo decidir si *tengo que hacer una oferta a un clinte o conceder un crédito a un usuario*, no son preguntas que se puedan obtener de forma directa de los registros de una base de datos, ya que lo que necesitamos es anticipar o entender el con-

junto de datos, de que disponemos, entender el histórico y, a partir de ahí, extrapolar (puede ser desde el punto de vista predictivo o de grupos o tipologías). Este tipo de decisiones a este nivel, si queremos automatizarlas, requieren de herramientas de *aprendizaje automático*, la minería de datos nos permite tener modelos descriptivos o predictivos que nos permitan tomar decisiones.

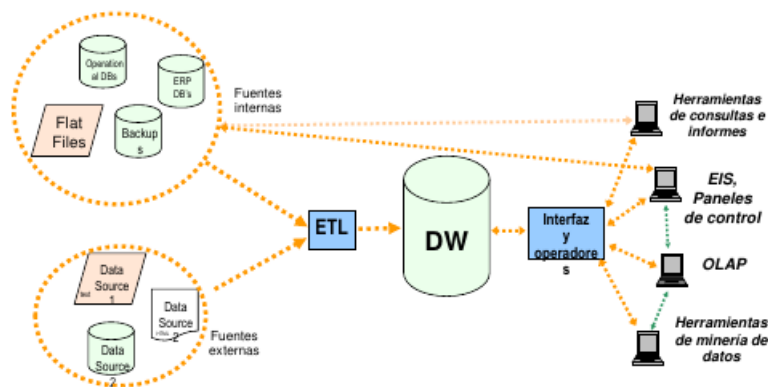
Vemos que a medida que ascendemos en la pirámide se requieren herramientas cada vez más sofisticadas.

La *inteligencia empresarial* es un término que engloba las herramientas de tecnología de la información y normalmente suelen estar compuesta de tres áreas:

- herramientas de almacenamiento de datos y OLAP, permiten manejar volúmenes de información y combinarlos, agregarlos y rotarlos.
- paneles de control y herramientas de supervisión, permiten a través de gráficas ver como va la organización en todo momento o unirlos a los informes.
- herramientas de minería de datos, permiten extraer de estos datos modelos con los que tomar decisiones; son modelos que pueden ser predictivos o descriptivos.

2.6 Minería de datos

Una estructura básica, sin entrar en detalles, de como se organiza la *inteligencia empresarial* sería la representada en el gráfico.



En primer lugar, al analizar datos e integrarlos para tomar decisiones estratégicas utilizaremos tanto las fuentes internas de la organización como las externas. Fuentes internas suelen ser el sistema de información de la organización, información que se vuelca en un *almacén de datos*, que generalmente está desconectado del trabajo diario de la organización. Normalmente, la fuentes internas, no permiten una visión completa del entorno de

la organización, en la organización funciona normalmente un entorno económico y social y se requieren datos e información del exterior, como pueden ser: tendencias del sector, qué eventos están ocurriendo. Temas como meteorología y calendarios, son fundamentales a la hora de hacer modelos predictivos; es información que podríamos necesitar integrarla, no es información necesaria desde el punto de vista operacional o trasaccional, pero es fundamental a la hora de extraer modelos que sean realmente predictivos y descriptivos sobre lo que ocurre alrededor de la organización.

Es una información, que como externa, puede estar en muchas fuentes diferentes, en bases de datos que uno compra o que están disponibles gratuitamente, de informes textuales, muchas veces en páginas web, ...

Como hemos dicho todo se integra en una base de datos o almacén de datos, que podemos ver que podemos organizarlos de diferentes maneras y, sobre esta información integrada, podemos aplicar una serie de herramientas con las que extraer informes.

Por último, tenemos *paneles de control* y *sistemas de información ejecutivos* que pueden hacer lo mismo. Las herramientas *OLAP* son herramientas que permiten, una vez realizada una consulta compleja que incluye la selección de varias dimensiones de los datos y agregadas a distintos niveles, entre otras:

- cambiar rápidamente entre niveles
- avanzar la consulta sin tener que esperar a que ésta acabe

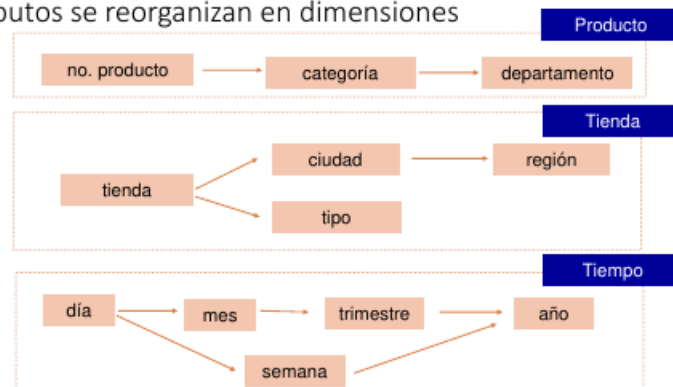
Las herramientas de minería de datos permiten a través de los datos históricos generar vistas *minables* y extraer modelos descriptivos y predictivos a partir de estas vistas.

2.7 Almacenamiento de datos multidimensional

Los almacenes de datos normalmente se organizan como una base de datos, integrando información de diferentes fuentes que posean información valiosa desde el punto de vista de responder a cuestiones de tipo estratégico, más que de tipo transaccional.

Almacenaje de datos multidimensional

- Los atributos se reorganizan en dimensiones



Para el trabajo transaccional utilizamos la base de datos de toda la vida, generalmente relacional.

Cuando el almacén de datos es sólo para realizar consultas, no habrá actualizaciones ya que estas normalmente van a la base de datos transaccional, ¿qué tipo de organización es más habitual?. Desde el punto de vista de la organización del almacén de datos tenemos, lo que se conoce como **modelo multidimensional**, es un modelo muy intuitivo ya que la información se suele dividir en dimensiones, que nos permiten agregar o desagregar la información a través de esas dimensiones.

Tomemos el ejemplo de un supermercado. En un supermercado puedo realizar las típicas preguntas o adverbios interrogativos:

- ¿Qué es lo que se vende en el supermercado?, hablamos del producto.
- ¿Dónde se vende?, nos aparece la dimensión tienda.
- ¿Cuándo se vendió?, nos aparece la dimensión temporal, cada producto se ha vendido en una tienda en un momento determinado.

Podemos hacer más preguntas como el tipo de pago, la hora, etc.

Nos vamos a centrar en qué, dónde y cuándo.

- El **qué**, hablamos del producto, cuando intentamos analizarlos vemos que se organizan jerárquicamente o en diferentes niveles. En general, del producto tenemos la siguiente información:
 - número del producto o código.
 - categorías.
 - departamentos.

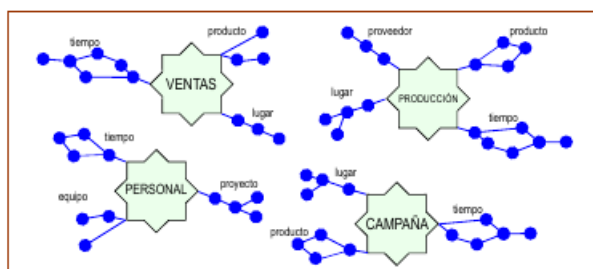
Puede haber más subdivisiones, aunque por simplicidad sólo mostramos tres.

- El **dónde**, referencia al punto de venta, una tienda concreta o venta online. En una tienda podríamos subdividir, incluso, por cajas. Podrían agregarse el tipo de tienda: minis, grandes o grandes superficies, ... Podemos agregar una dimensión *geográfica*: ciudad, región, país, ...
- El **cuándo**, el tiempo o momento en que se realizó la venta, bien el día, semana, mes, año, etc, incluso añadir más nivel de granularidad como la hora y minuto.

Este tipo de dimensiones y jerarquías de los niveles es parte del *diseño multidimensional* de este tipo de almacén de datos.

¿Cómo funcionan?. Hablabamos de ventas, para este caso sería la información que tendríamos sobre productos, lugar y tiempo.

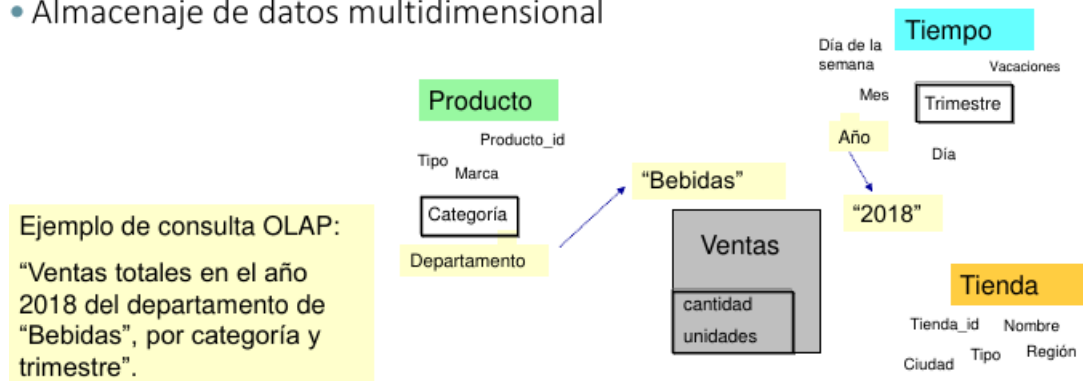
Cada diagrama se llama *datamart*:



Podríamos tener otros subalmacenes que normalmente se concocen con el nombre de **datamarts** en el que la información es de otro tipo. Para producción, podríamos tener información sobre campañas, que podrían ser independientes de la información de ventas. Cada una de ellas sería un *datamart* y cada uno de éstos tiene su estructura multidimensional.

Analicemos el *datamart de ventas*. Tenemos un ejemplo de una consulta, de momento ignoramos la parte de OLAP.

- Almacenaje de datos multidimensional



Quiero saber las ventas totales en el año 2018, del departamento de bebidas, por categorías y por trimestre.

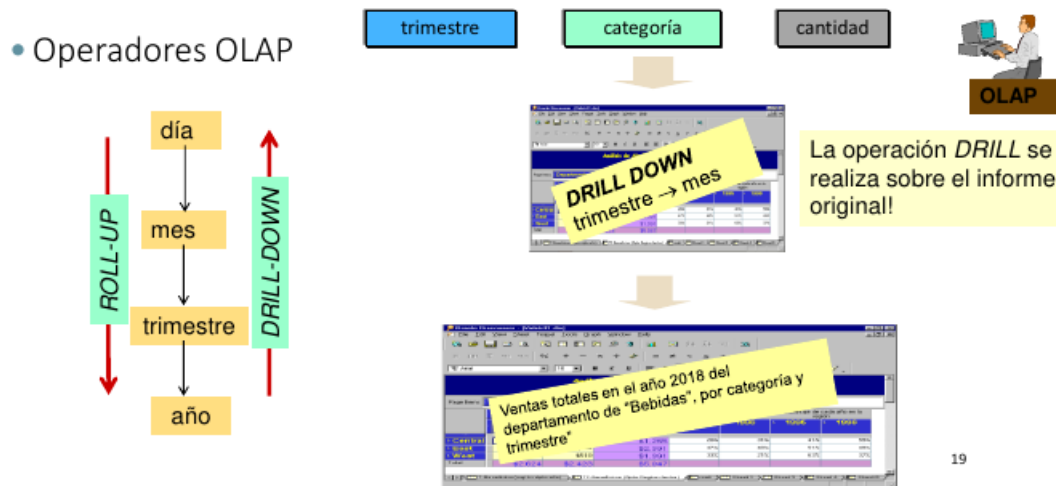
1. Nos centramos en el *datamart de ventas*, encontramos que hay una tabla central que tiene la información de los indicadores, en este caso cantidad(precio) y unidades.
2. Como cualquier pregunta suelen estar asociados con el qué, cuándo y cómo. En nuestro caso, el indicador son las ventas en el año 2018, tendremos que irnos al tiempo, en esta dimensión agregará *año*, en la dimensión de productos agregará la dimensión *departamento, bebidas*.
3. Por último, categoría y trimestre. Estos atributos no son selecciones sino agrupamientos.

Obtendríamos una tabla con la información seleccionada. Esto nos da un resultado que puede representarse de esta forma:



Podríamos ponerlo como una tabla en el que las categorías son las filas y los trimestres las columnas, y cada celda, las ventas totales.

Ahora podríamos querer cambiar algo en la consulta, para ello utilizaremos las *herramientas OLAP* (*OnLine Analytic Processing* o *Proceso analítico online*) para realizar un refinamiento de las consultas sin tener que lanzar una nueva consulta. En nuestro caso podríamos hacer, lo que se conoce con *drill down*, ir de trimestre a mes sin tener que realizar un informe nuevo, sino desde el informe original.



2.8 OLAP vs Minería de Datos

¿Es **OLAP** lo mismo que aprendizaje automático, que la minería de datos?. En principio, son cosas diferentes, aunque existe una delgada línea entre las preguntas que pueden responderse con herramientas de minería de datos o técnicas de aprendizaje automático y las que se pueden responder con herramientas *OLAP*, un almacén de datos y sin extrapolar.

La línea que separaría ambos ámbitos, es que *OLAP*, aunque sea de manera agregada, de manera compleja, todo el tipo de consultas estratégicas que hacemos sobre el almacén de datos no dejan de ser consultas sobre datos que existen en dicho almacén, la consulta no se inventa nada, no hace ninguna extrapolación ni hipótesis sobre los datos. Sin embargo, cuando hablamos de *minería de datos*, aquí existe un razonamiento hipotético, normalmente inductivo.

En la tabla tenemos preguntas que podemos clasificar en un de los dos ámbitos vistos.

OLAP

Minería

¿Cuál es el promedio de accidentes entre los fumadores y los no fumadores?	¿Cuáles son los mejores vaticinadores para los accidentes?
¿Cuál es el promedio de la factura de teléfono de mis actuales clientes vs. mis exclientes?	¿Dejará X la compañía? ¿Qué factores afectan a las dimisiones?
¿Cuál es el promedio de compras diarias entre los usuarios de tarjetas de crédito robadas y usuarios legítimos?	¿Qué patrones están asociados al uso de tarjetas de crédito fraudulentas?

Las preguntas que aparecen en la columna de *OLAP* vemos que son datos que se encuentran en el almacén de datos, y que lo que hacemos es lanzar consultas sobre datos reales recogidos en ellas, sin tener que elaborarlos. Por el contrario, en el lado de la *minería*, la información tienen que se interpretada, requiere la elaboración de un *modelo hipotético* que relacione las entradas, en el caso de la primera pregunta, las características de los clientes como: edad, sexo, años de antigüedad del carnet, tipo de comportamiento, dónde vive y demás, con respecto al número de accidentes que pueda tener; esto se hace a través de información histórica y no de información futura, que nos permite entender cuál es esta relación y describe lo ocurrido de forma abstracta.

En relación con la pregunta ¿qué patrones están asociados al uso de tarjetas de crédito fraudulentas?, la pregunta es bastante más amplia, aquí lo que podemos responder son cuestiones que son anómalas, por ejemplo, realizar dos o más compras de un tipo concreto, puede estar asociado a un uso fraudulento de las tarjetas, ya que normalmente son compras, por ejemplo, en joyería, alguien que nunca ha comprado este tipo de artículos y por cantidades elevadas, son pagos bastante inusuales para el perfil del individuo; aquí se recurriría a herramientas de aprendizaje automático principalmente, para extraer ese tipo de patrones.

De las siguientes afirmación, cuales pueden resolverse mediante herramientas **OLAP**:

- A) ¿Cuál es el producto más vendido entre el rango de edad 25-45?
- B) ¿Cuántos productos de la categoría «hogar» se espera vender el año que viene?
- C) ¿Qué diferencia de tamaño hay entre las cestas de la compra de los sábados y entre semana?
- D) ¿Qué días del mes suele haber más rotura de stock?
- E) ¿Qué tipo de productos tengo atendiendo a sus características: peso, precio, etc.?

Las respuestas correctas son A), C) y D).

2.9 Ejemplos de Ciencia de Datos

La *ciencia de datos* en un contexto de *inteligencia de negocio* es una actitud más que una disciplina desde el punto de vista de una organización. Cuando una organización contrata un científico de datos, lo que quiere es una persona con un rol muy particular.

Podemos pensar que un científico de datos es lo mismo que otros roles existentes en las organizaciones: director de sistemas de información o *data manager* o el director de datos o *data officer*. En empresas relacionadas con la tecnología, son roles muy importantes en la organización, son roles de primer nivel, como el caso de Google o Facebook, donde lo más importante son los datos. El científico de datos va en esa línea, debería ser una persona con un conjunto integrado de habilidades que abarcan:

- matemáticas
- aprendizaje automático
- inteligencia artificial
- estadística
- optimización de bases de datos

pero con un profundo conocimiento en la elaboración de problemas para el diseño de soluciones efectivas. Su función debe ser extraer valor de los datos y sacar valor a partir de ellos, haciendo que éstos sean el motor de la organización, no simplemente una herramienta que existe en la organización que está ahí para el funcionamiento de las aplicaciones, sino conseguir que esos datos sean los que proporcione valor y hagan que la organización funcione.

Nos podemos preguntar ¿dónde se originan esos datos? y ¿para quién queremos obtener ese valor?. Dependiendo de estas preguntas podemos clasificar distintos tipos de aplicaciones o problemas que podemos resolver mediante la ciencia de datos.

- Los datos que tengo son valiosos para mí ($in \rightarrow in$), prácticamente el 100% de las organizaciones que tengan datos pueden sacar rendimiento a partir de ellos, a parte de lo que pueda ser el sistema transaccional, pueden sacar valor, sacar patrones que ayuden a la hora de la toma de decisiones. Aunque hablamos de organizaciones es extensible a los individuos. Esto es lo que se ha conocido como *inteligencia empresarial clásica*.

Cuando hablamos de **datos internos de la organización** para dar servicio a la organización, un ejemplo habitual es el de una compañía de seguros de automóviles donde queremos predecir qué póliza va a comprar un cliente. Teniendo en cuenta su historial de transacciones, miro el almacén de datos y extraigo los datos que me

permitan establecer un *modelo predictivo* sobre qué póliza es la más adecuada. No necesita buscar datos fuera de la organización.

- Los datos que he visto exteriores a la organización, los podría utilizar para sacar patrones que me ayuden en mis decisiones ($out \rightarrow in$), dichas fuentes pueden ser redes sociales.

Como ejemplo, tenemos la *Comisión del Crimen de Detroit*, en su momento intentaron utilizar información sobre redes sociales para ver realmente dónde y en qué momentos se estaba produciendo tráfico de drogas u otro tipo de delincuencia. Es sorprendente, pero algunos delincuentes son capaces de decir lo que están haciendo en las redes sociales o de lo que van a hacer o dónde van a quedar, a partir de ahí la policía puede saber qué puntos calientes hay, e incluso poder abortar algunos de los posibles delitos. Son datos completamente externos que le permiten tomar decisiones estratégicas.

- Nosotros disponemos de datos, que son valiosos para nosotros, pero también podrían serlo para otras organizaciones externas ($in \rightarrow out$), como sacar esos datos es un servicio que los pone disponibles. Podemos vender los datos directamente o el «conocimiento», como se extrae dicho conocimiento.

Cuando hablamos de externos, por ejemplo una compañía de telefonía, sus clientes no son externos son internos, al hablar de externos nos referimos que somos conscientes de que tengo información sobre miles de usuarios, que son internos, y que esa información puede ser útil para terceros. Por ejemplo, tendría información de dónde se encuentran los usuarios en una ciudad, si tengo el 20% de cuota de teléfonos móviles en una localidad, puedo saber dónde están, no a nivel individual, esta información no se puede proporcionar a terceros por razones legales, pero sí a nivel *agregado*. Puedo saber qué proporción de gente se mueve en una zona de la ciudad en un rango de horas. Si mi porcentaje es lo suficientemente representativo, podré extrapolar y saber los flujos de población en la ciudad. Esta información puede ser importante para ayuntamientos, comercios, transportes, etc. Evidentemente, el poseedor de la información sacará un beneficio de proporcionar este conocimiento.

- En ocasiones, yo no tengo los datos, ni siquiera soy el fin, soy un intermediario ($out \rightarrow out$), un experto en ciencia de datos que dispone de los datos y establece que si se extrajeran determinados patrones, los datos podrían ser útiles para una organización o usuarios. Muchos servicios hoy en día intentan hacer este tipo de cosas.

Cómo en casos anteriores, recurrimos a las redes sociales. La gente, con frecuencia, cuando se encuentran mal, enfermos, pueden indicar sus síntomas e incluso que acudieron al médico e indicar qué pronóstico les dió en redes sociales. Se intenta hacer asociaciones entre síntomas, como se toman los medicamentos, etc. Es una información que es externa, y que estaba dirigida para otro tipo de uso.

- Finalmente, la visión más ambiciosa, no hay datos ($\emptyset \rightarrow out$), pero si los hubiera

habría valor, es decir puedo intentar crear aplicaciones cuyo objetivo final sea crear datos que den valor. Gran parte de la economía digital reciente se basa en esta idea. Si volvemos a las redes sociales, éstas a partir de la nada han creado datos, a partir de éstos conocen el perfil de la gente, sus aficiones, gustos, horas de contacto, amigos, familiares, es decir, no es que esa información existiera, pero mucha se crea a través de esas aplicaciones que se crean inicialmente de la nada.

Una de las aplicaciones más típicas, relacionadas con este punto de vista, es que instalamos una aplicación en nuestro dispositivo móvil que nos informa, a través de datos que recoge de otros usuarios de la misma aplicación. El ejemplo típico son aplicaciones de tráfico, nos permiten saber las rutas más cortas y menos congestionadas.

Otra forma de ver la ciencia de datos es en las áreas donde se aplica, aquí podemos ver algunas de ellas:

- | | |
|--|---|
| <ul style="list-style-type: none"> • Datos de telecomunicaciones <ul style="list-style-type: none"> ▪ Valioso para comerciantes, tráfico, ayuntamiento, policía... • Otros datos de geolocalización (Flickr, Instagram, Wikiloc, ...) ▪ Valioso para agencias de viaje... • Datos en consumo de energía <ul style="list-style-type: none"> ▪ Valioso para anuncios de televisión... • Datos del transporte público (bus, metro, tren, taxi, tráfico, ...) ▪ Valioso para turismo, consumo, contaminación, comercio... • Datos de redes sociales. <ul style="list-style-type: none"> ▪ Valioso para casi todo... | <ul style="list-style-type: none"> • Datos de uso de tarjetas de crédito <ul style="list-style-type: none"> ▪ Valioso para comercios, ayuntamientos, ... • Datos de policía <ul style="list-style-type: none"> ▪ Valioso para aseguradoras, agentes inmobiliarios, ... • Datos comerciales (Amazon, Ebay, segundamano.es, ...) ▪ Valioso para salud, demografía, sociología... • Datos climatológicos <ul style="list-style-type: none"> ▪ Valioso para comercios. • Datos de búsquedas web. <ul style="list-style-type: none"> ▪ Valioso para casi todo. |
|--|---|

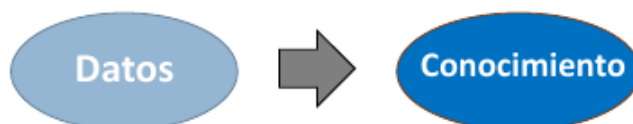
3

Proceso de extracción de conocimiento

Cuando nos centramos en la parte de la *minería de datos*, más predictiva o descriptiva, nos centramos en convertir datos en conocimiento, proceso que se conoce por sus siglas en inglés de **D2K (Data to Knowledge)**, de datos a conocimiento.

“Proceso no trivial de identificar datos válidos, novedosos, potencialmente útiles y comprensibles”.

(Fayyad et al. 1996)



Este proceso podemos traducirlo como extracción de conocimiento a partir de datos o bases de datos de cualquier tipo. Es un proceso que se conoce hace tiempo, no trivial de identificar patrones válidos, potencialmente útiles y comprensibles.

- Es un **proceso no trivial**, si fuera sencillo no serían necesarias herramientas, de conocimiento o reglas o técnicas para resolver el problema.
- **Identificar patrones**, son patrones que van más allá de los datos, son tendencias, reglas, grupos que no son sólo datos agregados; normalmente inferen otro tipo de información a partir de los datos ya existentes. Como son datos hipotéticos, que extrapolan a partir de otros datos, históricos, nos interesa que sean válidos, que los patrones tengan un porcentaje de error mínimo. La parte de evaluación es fundamental.
- **Novedoso**, por ejemplo que en una unidad de maternidad de un hospital, el porcentaje de ingresos de mujeres siempre será del 100%, este procedimiento no es novedoso. También se da el caso de ser novedoso pero inútil, puedo suponer que dos productos, champú y una marca de tomate se compran conjuntamente, ¿qué puedo hacer con esa información?, poner el champú cerca del tomate, probablemente será una curiosidad, pero difícilmente será útil.
- La información debe de ser **comprensible**, hablamos de técnicas de aprendizaje automático, redes neuronales o redes profundas, máquinas de vectores, que proporcionan información poco comprensible para un humano, frente a técnicas que producen reglas que sí son comprensibles por un humano, como pueden ser los árboles de decisión.

Este proceso normalmente se desglosa en fases más concretas, no sólo en cómo convertir datos en conocimiento



en la parte izquierda tenemos el conocimiento, el final aparece como decisiones, a partir de ese conocimiento, es lo que se denomina extracción del conocimiento a partir de datos.

El número de tareas o de fases puede variar, pero generalmente se empieza por una parte de integración de datos, con fuentes internas y externas con distintos formatos, que integraremos generalmente en un repositorio, que no tiene por qué ser multidimensional, pero sí un repositorio donde tenemos todos los datos juntos, integrados y consistentes. Aunque tengamos todo los datos almacenados, será necesario prepararlos para el modelo que queremos extraer. Este proceso de preparar para extraer conocimiento es lo que se conoce como *crear la vista minable*.

En la vista minable pongo una filas y columnas con la información que quiero, una vez que tengo todo esto puedo aplicar minería de datos, o aprendizaje automático, para obtener patrones, que normalmente tendré que evaluar, tendrán un error, fallarán con alguno de los ejemplos del histórico. Habrá que seleccionar aquel patrón que mejor se ajuste a los datos y que tenga un menor error.

Una vez realizado el proceso, dispondremos de un **modelo evaluado y validado**, y hablamos realmente de conocimiento, ya podremos utilizarlos y tomar decisiones.

Estos modelos en ocasiones quedarán obsoletos, las situaciones cambian o aparecen problemas inesperados; podemos ir hacia atrás, al inicio o a un reentrenamiento de los modelos dependiendo del problema que detectemos.

3.1 Estándar CRISP-DM

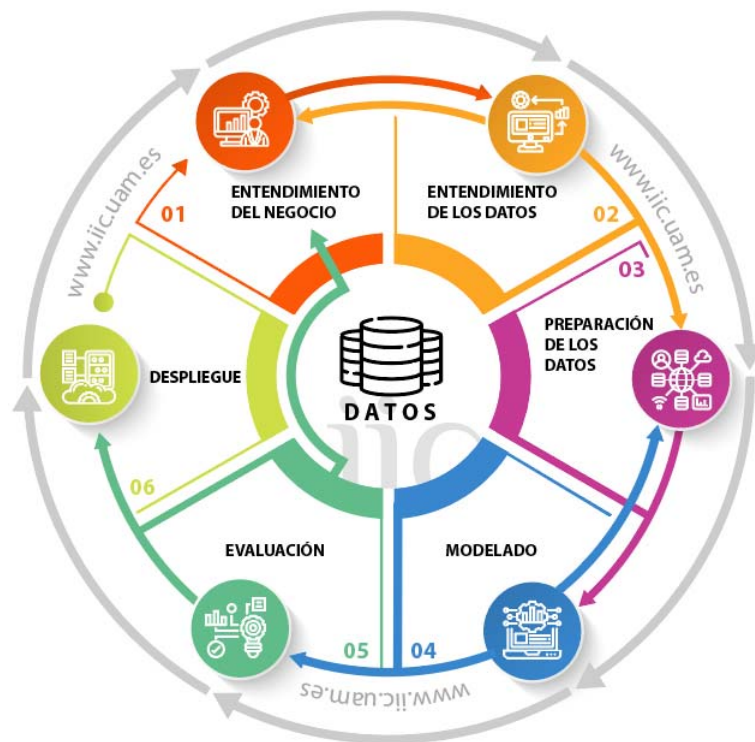
Hemos visto que el proceso de extracción de conocimiento se estructura en una serie de fases, que pueden variar dependiendo del contexto o de la organización, incluso podemos ver el proceso a partir de conocimiento muy simple.

Cuando vemos un proceso que tiene que seguir una serie de pasos aparecen estándares o metodologías que guían el proceso. Seguir estos estándares permite que diferentes usuarios, e incluso organizaciones, se entiendan a la hora de desarrollar un proyecto. Un proyecto ya finalizado, es más entendible si sigue una metodología.

Algo importante de las metodologías es que van acompañadas de documentación, que permite que los usuarios externos al proyecto puedan entender e integrarse en el proyecto más fácilmente. No es más que **gestión de proyectos** aplicado a ciencia de datos.

La metodología más común se conoce como **CRISP-DM**, apareció hace más de veinte años, como un proyecto de la comisión europea y de empresas, siendo una de las metodologías más utilizadas.

La metodología CRISP-DM se conceptualiza en 6 fases:



1. **Entendimiento del negocio**, el equipo de trabajo debe comprender los objetivos y requisitos del proyecto definidos por el cliente, para convertir el conocimiento en una definición técnica del problema. Requiere una comunicación intensa entre cliente y equipo técnico.
2. **Entendimiento o compresión de los datos**, el equipo técnico realiza un *análisis exploratorio* para obtener una visión general de lo que se puede conseguir con los datos. Tras el análisis debería tenerse una idea clara de la viabilidad del proyecto y de los resultados esperados, de ser así se avanza a las siguientes dos fase.
3. **Preparación de los datos**, cubre las actividades para construir el conjunto de datos definitivo que se empleará en la siguiente fase.
4. **Modelado de datos**, el equipo técnico realiza los análisis y modelos pertinentes de los que se deriven los resultados y conclusiones del proyecto, y su validación frente a errores.
5. **Evaluación**, el cliente determinará la calidad de los resultados obtenidos y decidirá cómo pueden explotarse antes del despliegue.
6. **Despliegue**, fase en la que el modelo se pone en producción, con el fin de tomar decisiones.

Se trata de un proceso de **carácter iterativo**. Aunque algunos consideran que esta metodología es un poco rígida, siempre puede ser utilizada como base y adaptarla a nuestras necesidades.

4

Tareas, técnicas y herramientas

4.1 Tareas

Algo muy importante es distinguir entre tareas, técnicas y herramientas. Nos vamos a centrar ahora en ¿qué son las tareas?.

Vamos a ver cinco grandes familias de tareas:

- Regresión
- Clustering
- Reglas de Asociación
- Análisis Factorial
- Dispersión o Multivariante

La primera pregunta que nos haremos, si vemos una tabla como la siguiente:

x1	x2	x3	x4	x5	x6	...	xn
20	315	High	1.9	Married	0.2	...	sí
135	310	Low	2.1	Single	0.3	...	no
...

si hay alguna/s variable/s de salida y/o de entrada. Las **variables de salida** representan lo que quiero **predecir** a partir de las **variables de entrada**. En el caso que tenemos, X_n es una *variable de salida*, y el resto de variables, X_1, X_2, \dots , serían *variables de entrada*. En el momento en que determinamos la existencia de una/s variable/s de salida, hemos establecido una **tarea predictiva**.

Por ejemplo, quiero saber si a un cliente hay que darle un préstamo, o una operación con tarjeta de crédito es fraudulenta o las ventas del próximo trimestre, etc. Todas son tareas que tienen un valor de salida a partir de valores de entrada.

Podemos distinguir dos tipos de *tareas predictivas*:

- **Clasificación/Categorización**. Tenemos una *variable de salida*, el resto son *variables de entrada*, el valor de la variable predictiva es **nominal** o **cualitativa**. Por

ejemplo, predecir si se concederá un crédito da como resultado una variable cualitativa o nominal, **sí o no**, hablamos de una *tarea de clasificación*.

- **Regresión.** Si la *variable de salida* es numérica. Por ejemplo, cuando queremos saber cuantos productos puedo vender en el trimestre.

Pero y cuando no hay ninguna **variable de salida**, hablaríamos de una **tarea descriptiva**, muchas de las tareas del *aprendizaje automático* son descriptivas, no predicen nada. Hay tareas *descriptivas* que tienen un valor a la hora de entender y extraer conocimiento a partir de los datos. Por ejemplo, si lo que quiero hacer es entender las filas o las columnas, tengo diferentes tareas descriptivas.

Los tipos de *tareas descriptivas* son.

- **Clustering**, si quiero entender la relación entre las filas de la tabla, relación entre individuos en este caso, si éstos fueran clientes puedo hacer un *agrupamiento de clientes*, *clustering de clientes*. Es una tarea cuyo objetivo es describir grupos de datos, que grupos de clientes tengo.
- **Análisis Exploratorio**, es el análisis de las relaciones entre columnas:
 - **Reglas de asociación, dependencias funcionales**, cuando las variables son **nominales, cualitativas**, no son números. Por ejemplo, las columnas que hacen referencia a *alto*, *bajo*, *casado*, *soltero*,... son nominales. Si el valor de *High* en la variable X_3 está asociado con un valor de *Married* en la variable X_5 , sólo podré verlo si hay muchísimos ejemplos, ver si esta relación es más frecuente que otras. Otro ejemplo sería el *análisis de la compra*, ver que productos se compran conjuntamente, ejemplo típico de las **reglas de asociación**.
 - **Análisis Factorial o Multivariante**, si las variables que queremos relacionar son numéricas. El concepto de *relación binaria*, cuando hablamos de este tipo de variables, se conoce tradicionalmente como **correlación**. Entre las variables X_4 y X_6 podría ver si estos valores están correlacionados, por ejemplo, si la edad y el sueldo están correlacionados. Es un ámbito inmenso, que restringimos normalmente a preguntas más básicas como calcular correlaciones y una matriz de correlaciones. Esta última clasificación, también la conocemos como **análisis factorial de correlaciones, análisis de dispersión o análisis multivariable**.

La pregunta es ¿qué ocurre si tengo varias variables numéricas y varias nominales?.

- Si tenemos *muchas numéricas y pocas nominales*, podemos *numerizar todas las variables* y realizar un **análisis de correlaciones**.

- Si tenemos *muchas nominales y pocas numéricas*, las numéricas podemos **discretizarlas** y realizar un **análisis de regla de asociación**.

Son tipos de variables que no podemos analizar al mismo tiempo, pero podemos convertir unas en otras y aplicar el análisis que más convenga.

Es una manera simple, pero elegante de entender qué tipos de **tareas** podemos tener en *aprendizaje automático*, las diferencias fundamentales son:

- **predictivas**, qué tipo de variable tenemos que predecir, numérica o nominal.
- **descriptivas**, queremos analizar las relaciones entre filas, *agrupamientos*, o analizar relaciones entre columnas, *análisis exploratorio*, si los valores son numéricos, *correlaciones*, si son nominales, *asociaciones*.

¿Qué tipo de tarea es «*diagnostiscar la presencia de una enfermedad dados unos síntomas*»?:

- A) Clasificación
- B) Regresión
- C) Asociación
- D) Clustering

Solución: A), clasificación.

4.2 Técnicas

En un ámbito general, las tareas de minería de datos o en ciencia de datos (aprendizaje automático), se dividían en cinco clases principales, dependiendo de si eran:

- **supervisadas**,
- **no supervisadas**,
- **predictivas** o
- **no predictivas**

Es habitual confundir *tareas* con *técnicas* ya que muchas o algunas de las técnicas sólo resuelven un subconjunto o incluso una sola de las tareas que hemos visto.

La tabla simplifica la relación entre *tareas*

TÉCNICA	PREDICTIVA / SUPERVISADA		DESCRIPTIVA / NO SUPERVISADA		
	Clasificación	Regresión	Clustering	Reglas de asociación	Otros (factorial, correlación...)
Redes neuronales	✓	✓	✓ *		
Árboles de decisión	✓ (4.5)	✓ (CART)	✓		
Kohonen			✓		
Regresión lineal (local, global), exp.		✓			
Regresión logística	✓				
K-means	✓ *		✓		
A Priori (asociaciones)				✓	
Análisis factorial, análisis multivariable					✓
CN2	✓				
K-NN (vecinos más próximos)	✓		✓		
FBR	✓				
Clasificadores básicos	✓	✓			

La diferencia entre *trema* y *técnica* es importante ya que si no confundiremos el problema que queremos resolver.

Veamos de forma descriptiva alguna de estas técnicas:

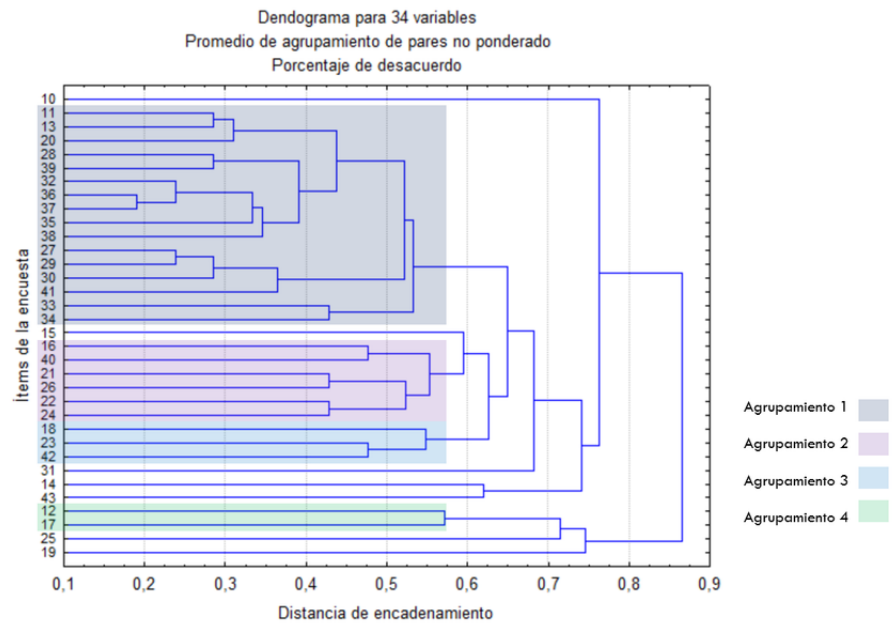
- **Técnicas descriptivas:**

- **Correlaciones y Asociaciones**, se suele denominar como *análisis exploratorio*, y aquí tenemos:

- * **Coefficiente de correlación**, cuando hablamos de variables numéricas y matrices de correlación, y queremos analizar más de dos variables. En principio el coeficiente de correlación es **bivariante**, ya que analizaría dos variables.
- * **Asociaciones**, hablamos de relaciones entre variables cualitativas, no numéricas, o categóricas. El ejemplo típico es la cesta de la compra, podemos preguntarnos si tabaco y alcohol están o no relacionados cuando se compran conjuntamente.
- * **Dependencias funcionales**, queremos conocer si una variable implica normalmente los valores de otra, se trata de una versión generalizada de las *reglas de asociación*.

- **Clustering**, normalmente se dividen en:

- * **Jerárquicos**, tenemos lo que se denomina un **dendograma**, en el que se agrupan ejemplos.



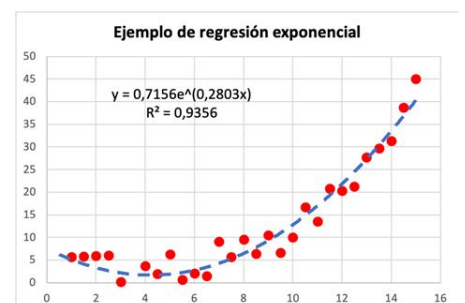
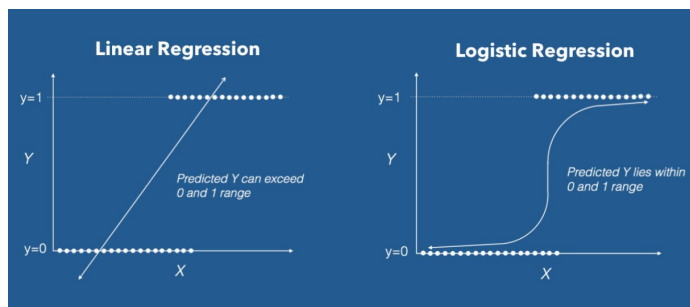
Son ejemplos que se agrupan por distancia de enlazado, al final podemos partir en los grupos que deseemos.

* **No Jerárquicos**, agrupan directamente en tres o cuatro grupos sin crear una jerarquía.

– kkkk

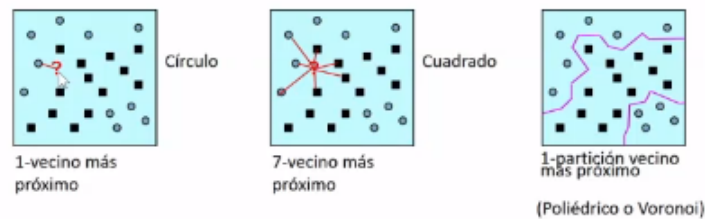
• **Técnicas predictivas.** Se refiere a los tipos de *regresión*:

- **Regresión lineal**, problemas de regresión, con variables de salida cuantitativa y numérica.
- **Regresión logística**, adaptación de la regresión lineal para problemas de clasificación, la variable de salida es *cualitativa*.
- **Regresión no lineal**, permite adaptarse a patrones, que de inicio, no son lineales.



Suelen verse juntas porque están muy relacionadas.

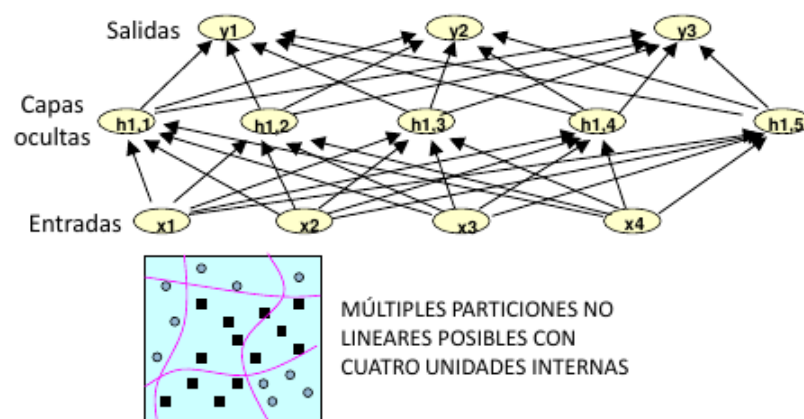
Veamos brevemente una serie de técnicas, una bastante antigua y usual es la denominada *vecinos más próximos*, muy intuitiva, se categoriza o predice dependiendo de sus vecinos; para determinar los vecinos hay una distancia, una medida de similitud y, a través de ella, se miden cuáles son los vecinos más próximos a un nuevo ejemplo. De ese nuevo ejemplo no tenemos la categoría, la clase. Lo que se hace es mirar cuáles son los vecinos más próximos y a partir de ahí determinar de qué clase es.



Si hablamos del vecino más próximo, sólo se mira uno, es menos robusto porque puede haber «ruido», puede ser que un grupo pueda estar rodeado de cuadraditos y el vecino más cercano pueda ser un círculo, cerca del punto que quiera predecir. La partición que se hace es muy parecida a la que haríamos con un lápiz si intentáramos determinar fronteras entre las clases, en la imagen entre cuadrados negros y círculos verdes.

Otro ejemplo bastante popular son las *redes neuronales*, y lo son por las aplicación de las *redes neuronales profundas* (las *redes neuronales no profundas* son sólo una capa oculta), se utilizan para muchas tareas, especialmente en las que las variables de entrada son bastante predictivas y no hay que crear nuevas características a partir de los datos.

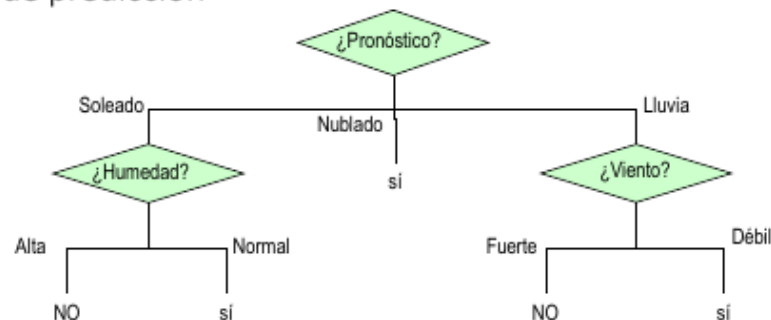
Redes neuronales



La ventaja que tienen las *redes neuronales*, de una capa o múltiples capas ocultas, es que permiten comportamientos **no lineales** como el mostrado en la imagen. Las fronteras se combinan a través de líneas que no son lineales, pudiendo ajustarnos a las formas que producen los datos como nosotros queramos.

Los *árboles de decisión* son muy comunes y es otra forma sencilla de entender los modelos extraídos utilizan técnicas que pueden representarse como un árbol, y se pueden representar en forma de reglas. Es una técnica muy utilizada en los entornos donde prima la *comprensibilidad*.

Árboles de predicción



A partir de los datos generamos un modelo, que dependiendo de tres variables: pronóstico, humedad y viento, podría determinar cuál será la clase de salida mediante los valores.

- Si pronóstico soleado,
- Si pronóstico nublado,
- Si pronóstico lluvia,
- Si viento fuerte,
- Si viento débil,
- Si humedad alta,
- Si humedad baja

Para tomar una decisión particular no tengo que evaluar todas las variables.

De la siguiente lista señal las **tareas**:

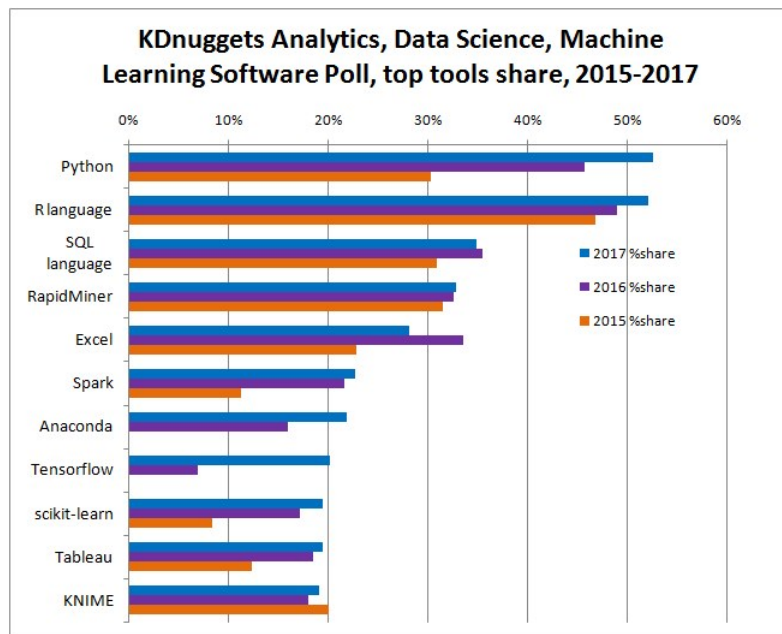
- A) Regresión lineal

- B) Clasificación
- C) Regresión
- D) Custering (agrupamiento)
- E) Redes neuronales
- F) Correlaciones
- G) Regresión Logística
- H) Reglas de asociación
- I) Vecinos más próximos

Solución: *B), C), D), F) y H)*

4.3 Herramientas

Para llevar a cabo las técnicas que hemos visto, y muchas más, se recurre a herramientas, en las que normalmente ya están implementadas dichas técnicas, ya sea en forma visual o mediante lenguajes de programación mediante el uso de librerías, o con programación orientada a objetos y librerías específicas como ocurre con *Python* o *R*.



Muchas de estas herramientas son gratuitas.

Lo interesante es ver que algunas de las herramientas se utilizan bastante en la comunidad:

- *R* y *Python* son muy utilizadas y además son gratuitas.
 - **R**, es un lenguaje interpretado, que dispone de muchas librerías, pero es necesario saber programar, crear código, lo que tiene ventajas y desventajas. Una de las grandes ventajas que tiene es que dispone de librerías de visualización muy potentes, por ejemplo **ggplot**.
 - **Python**, ha ido tomando espacio a *R* como lenguaje de programación predominante en el análisis de datos, especialmente a partir de aparecer la librería **Scikit-learn**, ésta contiene gran cantidad de técnicas de aprendizaje automático (*R* también dispone de algunas librerías de este tipo).
- Existen otro tipo de herramientas, estilo *RapidMiner*, que son más visuales, más sencillas de manejar, pero que suelen tener una licencia que ya no es completamente abierta.
- Están las herramientas comerciales, cuyo uso requiere de la compra de licencias, están relacionadas con el *BigData* y no tanto con el *aprendizaje automático*.
- Existe una serie de herramientas adicionales que podemos utilizar en muchos casos, que aunque no sean propiamente de análisis de datos o de aprendizaje automático son muy utilizadas; es el caso de *SQL*, muy utilizado cuando se trabaja con bases de datos, pero no es una herramienta de análisis, por el momento.

.

En la «nube» hay plataformas que permiten hacer el *aprendizaje automático* como son: AzureML y BigML, pero hay muchas otras.

Existen otras muchas herramientas que se *pagan por cómputo*, como **TensorFlow**, que podemos instalar en nuestra máquina pero si requiere hacer redes profundas y tienes unos clusters, en principio, a veces, es más cómodo lanzar todo esto mediante algún *servicio en la nube*, lo mismo ocurre con *Keras*. Todas estas herramientas, más sofisticadas, suelen utilizar en *aprendizaje profundo*, *reconocimiento de imágenes* o *de lenguaje natural*.

Chapter 2

Módulo 2: Evaluación de modelos de aprendizaje automático

La evaluación depende de la tarea que se realice. Distinguiremos entre tres tipos de tarea:

- **Aprendizaje supervisado**, donde tenemos variables de entrada y una variable de salida, que representa la solución deseada. La meta es aprender la *regla general* que convierte los datos de entrada en la solución correcta. Distinguimos entre:
 - **Clasificación**, donde la salida es una categoría.
 - **Regresión**, la variable resultado es numérica.
- **Aprendizaje no supervisado**, no se asigna ninguna etiqueta al algoritmo de aprendizaje, sólo existen variables de entrada. Distinguiremos:
 - Clustering
 - Reglas de asociación,
 - Correlaciones
- **Aprendizaje de Refuerzo**, donde un programa informático interactúa con un ambiente controlado en el que debe alcanzar una meta concreta: conducir un vehículo o los videojuegos son ejemplos de este tipo de tarea. Muy utilizado en Inteligencia Artificial y/o Robótica.

1

Métricas de Clasificación

La clasificación se predice a partir de una serie de entradas, para obtener una variable que es **categorica**:

- puede tener dos o más valores,
- no tiene un orden, puede ser si/no, a/b/c o d


el modelo lo que intentará predecir a cuál de las clases pertenecen los datos. Lo que tenemos que establecer es que si comparamos la clase predicha y la real coinciden.

La clase predicha se denota como $h(x)$ y la clase actual como $f(x)$, donde x es el ejemplo.

$$error_S(h) = \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$

Donde $\delta(a,b)=0$ si $a=b$ y de otra manera 1.

Clase predicha ($h(x)$)	Clase actual ($f(x)$)	Error
Comprar	Comprar	No
No Comprar	Comprar	Yes
Comprar	No Comprar	Yes
Comprar	Comprar	No
No Comprar	No Comprar	No
No Comprar	Comprar	Yes
No Comprar	No Comprar	No
Comprar	Comprar	No
Comprar	Comprar	No
No Comprar	No comprar	No

Errores / Total
 Error = 3/10 = 0.3

En el ejemplo, el modelo desplegado, a través de los datos etiquetados como *se ha comprado* o *no se ha comprado* un producto, se comparan con los que predice el modelo. Vemos que en algunos casos no concuerdan, esos casos son **errores**. En la clasificación lo que contamos es cuantas veces se falla, nos estamos refiriendo al *error*; en el ejemplo las veces que falla la predicción son 3 sobre 10, el error es de 30 % o 0.3. Podemos referirnos de forma inversa y decir que se acierta un 70 % de las predicciones, lo que normalmente se llama **porcentaje de acierto**, en inglés se designa con el término **accuracy**.

Este tipo de medida, porcentaje de acierto o porcentaje de error, es muy simple y en ocasiones se queda corta a la hora de entender como funciona realmente el modelo.

Una de las métricas habituales en clasificación son las que se conocen como:

- **Precision**, representa el porcentaje de documentos que son relevantes para la consulta, con **TP** no referimos a **True Positives**, *valores positivos que han sido verdaderos*, en el ejemplo TP /positivos predichos. ..
- **Recall**, representa el porcentaje de documentos que se devuelven por el estudio o modelo (TP), TP /positivos reales.

Si queremos integrar estados medidas, que nos dé más información, podemos recurrir otro tipo de medida, también habitual que se denomina **Medida F**, o también denominada **media armónica**:

$$\text{Medida } F = 2 \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}$$

Un modelo con una *Medida F* alta suele tener medidas altas en precision y recall.

TP o Positivos Predichos correctamente.

FP o Negativos Predichos como Positivos.

TN o Negativos Predichos correctamente.

FN o Positivos Predichos como Negativos.

TNRate o porcentaje de negativos verdaderos

estos parámetros nos permiten conocer el rendimiento de un modelo de clasificación binaria en circunstancias específicas.

2

Métricas para regresión
