**TAREAS Y PROCESOS**

1. **Clasificación**: Toma como entrada de un modelo una entidad y como salida dice a qué clase pertenece. Las clases pueden ser excluyentes.

Ejemplo: Para un cierto producto queremos saber si un cliente lo comprará o no. Clases: 1--> Lo comprará, 2 --> No lo comprará.

En este caso solo habría dos clases y serían excluyentes.

Similares: *scoring*, estimación de probabilidad de clase (Probabilidad de que siga el cliente, aunque nos dé un número lo que subyace es una categoría).

1. **Regresión** (“estimación de valor”): Dado una entidad intentará estimar o predecir el valor numérico de alguna de sus variables.

Ejemplo: Queremos saber para un cliente el uso que le va a dar a un servicio. Para ello podemos comparar con gente que sea similar y mirar sus histórico de uso de ese servicio.

**Nota**: Aunque la clasificación está relacionada con la regresión, la clasificación nos dirá **si algo ocurrirá o no** mientras que la regresión nos dirá **cuánto** ocurrirá.

1. **Clustering** (“agrupación): Intenta agrupar una población por su similitud, pero no por un propósito en concreto.

Ejemplo: Se puede utilizar para una exploración preliminar de clientes para ver por ejemplo que productos ofrecemos o desarrollamos.

1. **Coincidencia de similitud**: Teniendo el conocimiento de una entidad se intenta encontrar otra similar.

Ejemplo: IBM está interesado en encontrar compañías similares a las que le compran sus productos. Usan sus datos “*firmográficos*” (tamaño, el valor de los activos, el número de empleados, área donde actúa esa empresa).

Usos: La similitud se usa en tareas de minería de datos como clasificación, regresión y *clustering*(agrupación).

1. **Agrupación de co-ocurrencia**: Intenta encontrar asociaciones entre entidades basándose en las transacciones que las involucran.

Ejemplo: ¿Qué productos se compran juntos de manera frecuente? Quizá el libro A se suele comprar a la vez que el libro B o cierto tipo de carne se comprar junto a cierto tipo de salsa.

Usos: Para análisis de bolsa de mercado o para mejorar los sistemas de recomendación.

Sinónimos: Minería de objetos frecuentes, descubrimiento de reglas de asociación, análisis de cesta de mercado.

1. **Profiling** (perfilar o descripción de comportamiento): Intenta caracterizar el comportamiento típico de un individuo, grupo o población.

Ejemplo: ¿Cuál es el tipo de móvil que usa un determinado segmento de clientes? Es complejo porque podría ver cuánto tiempo emplea escribiendo, cuánto viaja, uso internacional, etc.

Usos: Se puede usar para detectar fraude, monitorizar intrusos en sistemas informáticos, mirando por ejemplo el horario de accesos o si hace alguna compra anormal.

1. **Predicción de conexiones**: Sugiere conexiones entre objetos de datos y estima la fuerza de ese enlace.

Ejemplo: Viendo que Karen y tú tenéis 10 amigos en común, ¿te gustaría ser su amiga? También se utiliza para los sistemas de recomendación mirando el grafo de las películas que un cliente ha visto o puntuado bien.

1. **Reducción de datos:** Se planea reemplazar un conjunto muy grande de datos por otro más pequeño que contenga información más importante, así es más sencillo de procesar.

Ejemplo: Se quiere saber las preferencias de un consumidor con respecto a las películas que prefiere ver. Se puede aplicar una reducción de datos filtrando por las preferencias que tiene en base al género de las películas.

1. **Modelado causal**: Intenta arrojar luz a que eventos o acciones influyen realmente en otros. Siempre debe acabar con una conclusión causal con las suposiciones exactas hechas para una conclusión causal vigilando el “efecto placebo” de pasar por alto ciertas suposiciones.

Ejemplo: Tras elegir un público para presentarles un anuncio las ventas suben. ¿Ha sido por el anuncio o por elegir bien el público? Para saberlo hay que invertir en métodos sofisticados.

Técnicas: Experimentos aleatorios controlados como los test A/B. Se basan en un análisis contrafactual el cual intenta entender que diferencias puede haber entre una situación y otra.

**Métodos de Aprendizaje supervisado vs Aprendizaje no supervisado**

El origen de estos términos viene del campo del aprendizaje automático (*machine learning*) y referente a sus métodos viene a decir que si el agrupar por ejemplo clientes en segmentos no tiene un propósito o un objetivo estaríamos ante métodos de **aprendizaje no supervisado**, mientras que si realizamos una segmentación por ejemplo para responder a una pregunta concreta estaríamos ante un problema de **aprendizaje supervisado.**

Los métodos utilizados son distintos si hablamos de un problema supervisado o no. Por ejemplo, el *clustering* es una tarea no supervisada que evidentemente al agrupar por similitudes es probable que no nos proporciones mucha información útil. Sin embargo, para poder utilizar minería de datos supervisada es que debe haber datos sobre el objetivo. Por ejemplo, es difícil saber si un cliente estará 6 meses con nosotros si apenas tenemos información de su permanencia (quizá solo lleve 2 meses con nosotros). De hecho, **la clave en una inversión de ciencia de datos suele ser la obtención de este tipo de datos**. Estas variables suelen ser llamadas etiquetas del individuo.

Aprendizaje supervisado: Clasificación, regresión y modelado causal. También podrían incluirse coincidencia por similitud, predicción de conexiones y reducción de datos.

Aprendizaje no supervisado: Clustering, agrupación por co-ocurrencia, perfilado, detección de anomalías.

Clasificación y regresión se distinguen por el objetivo (una es numérica y la otra es categórica, normalmente binaria).

**PROCESO DATA MINING**

El proceso no tiene un ciclo perfecto en el que hacer cada fase en un orden determinado una vez y al pasar a la siguiente fase ya no habría que volver a la anterior, sino que se basa en la iteración y en la vuelta y revisión de fases anteriores para mejorarlas y refinarlas.



Ilustración 1- The CRISP data mining process

**TÉRMINOS**

* **Leaks**: cosa a tener muy en cuenta cuando se preparan los datos. Se trata de información que aparece en los datos históricos pero que no está realmente disponible cuando se tiene que tomar la decisión. Ejemplo el número de páginas que visita un usuario en una sesión, ya que hasta que no acabe la sesión no se puede conocer este dato.