Módulo 2: Captura de datos

El dato y sus fuentes

Data BootCamp
Cámara Comercio Bilbao

Germán Alonso Lascurain
german@campus2b.com

TABLA DE CONTENIDOS

Metodologías de gestión basadas en el dato

- Fundamentalmente encontramos 3 metodologías de gestión. Las describen los procesos a acometer durante el ciclo de vida del dato.

    - CRISP-DM
    - SEMMA
    - KDD

Metodologías de gestión basadas en el dato

- CRISP-DM

Metodologías de gestión basadas en el dato

- CRISP-DM

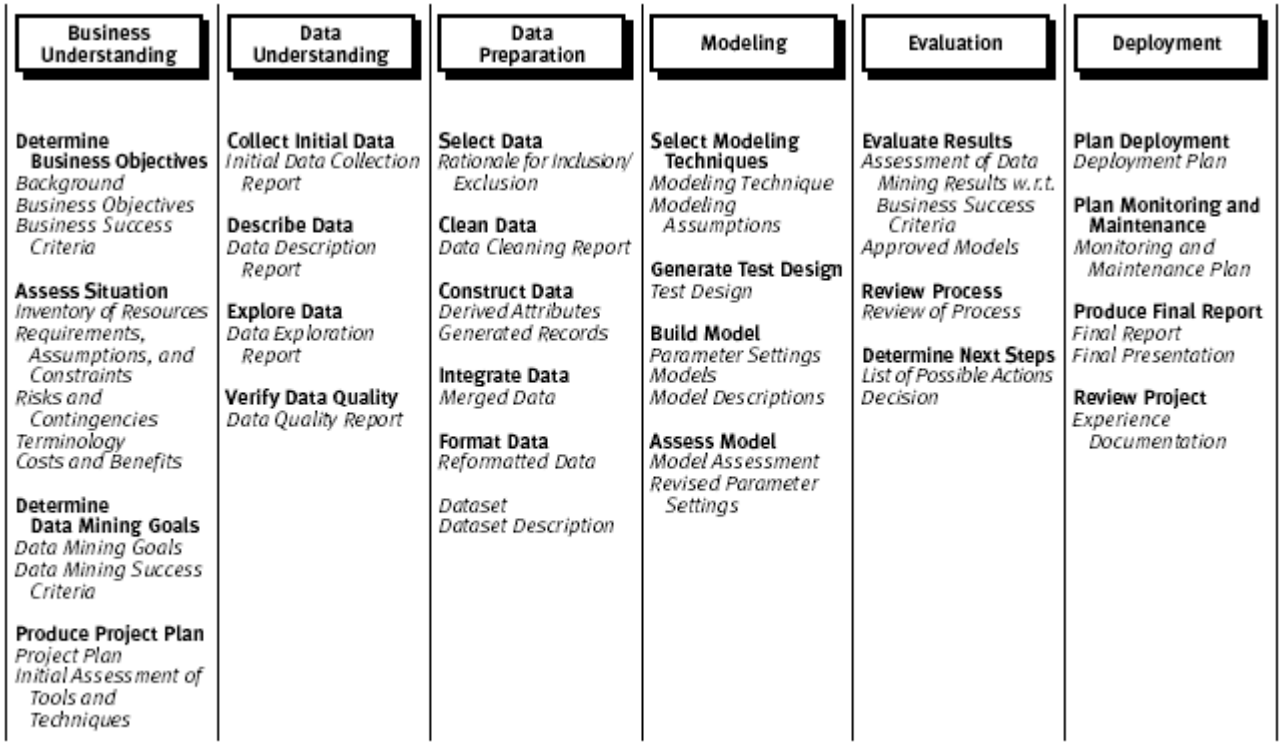
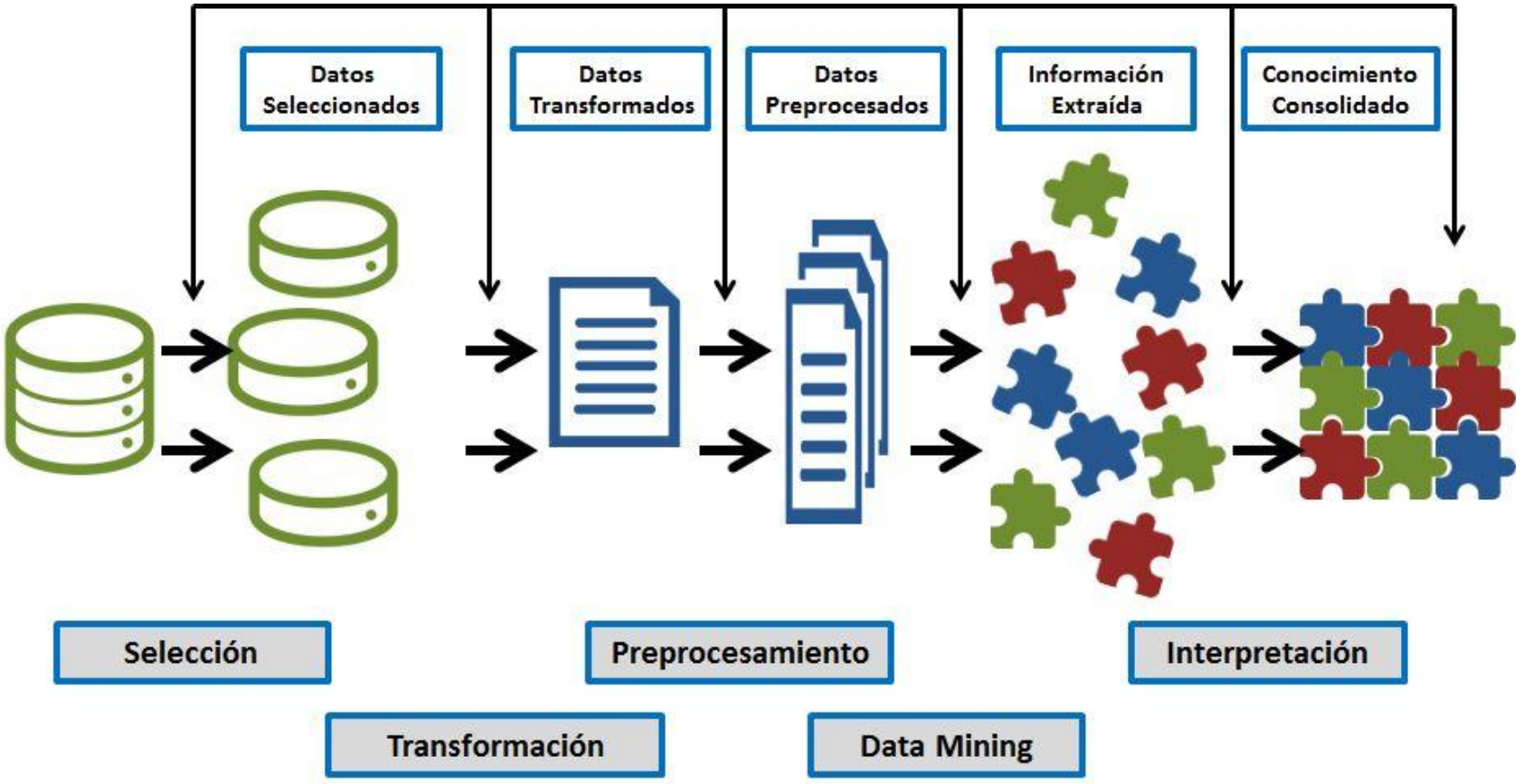


Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Metodologías de gestión basadas en el dato

- KDD



KDD - Etapas

Selección, limpieza, reducción y transformación

A menudo, los resultados van a depender más de la calidad de los datos en relación al problema que de la MD.

- Hablaremos de ruido en los datos, de relevancia de variables... siempre respecto a un objetivo.
- Una variable que consideremos ruido puede ser información útil para un problema distinto

KDD - Etapas

Selección, limpieza, reducción y transformación

- Selección de datos (variables).

  Ej. Utilizaremos el índice de masa corporal para caracterizar el riesgo de infarto

- Extracción de características.

  Ej. al procesar datos multimedia se extraen características que permitan construir vectores del tamaño necesario

- Selección de características descartables para reducir el número de variables

KDD - Etapas

Selección, limpieza, reducción y transformación

- Limpieza de datos.

  Recuperación de valores perdidos (imputación de datos)

  Tratamiento de valores anómalos (outliers)

  Suavizar ruido

  Eliminar inconsistencias

KDD - Etapas

Transformación de los datos

- Construcción de atributos.
  - Agrupamiento, separación, fecha, enteros a categóricos, ...
- Discretización
  - Convertir datos continuos en discretos

KDD - Etapas

Reducción de la dimensionalidad

- Reducción de casos / filas.
  - Puede hacer más eficiente el proceso de DM.
- Selección de variables (feature selection).
  - Selección del conjunto de atributos adecuado para la tarea.
  - Es uno de los pre-procesamientos MAS IMPORTANTES.
  - Técnicas usadas:
    - Estadísticas,
    - Basadas en búsquedas combinadas con métodos empíricos, ...

KDD - Etapas

Minería de datos

- Objetivo, producir nuevo conocimiento para ser usado.
  - Construcción de modelos (basados en los datos) que permitan describir los patrones y relaciones entre los datos para generar predicciones, explicar situaciones pasadas o entender mejor los datos.
- Decisiones a tomar:
  - ¿Qué tipo de conocimiento buscamos? → Predictivo, Descriptivo.
  - ¿Qué técnica es la más adecuada? → Clasificación, Regresión, Clustering, Asociaciones, ...
  - ¿Qué tipo de modelo? → Clasificación: Reglas asociación, árboles decisión, SVM ,...

KDD - Etapas

Evaluación, interpretación y presentación de resultados.

- La fase de MD puede producir varias hipótesis de modelos
  - Es necesario establecer qué modelos son los más válidos
  - Criterios: los patrones descubiertos deben ser precisos, comprensibles, e interesantes (útiles, novedosos)
- Medidas de evaluación de modelos. Dependen de la tarea.
- La interpretación de los mejores modelos (visualización, simplicidad, posibilidad de integración, ventajas colaterales,...) ayuda a la selección del modelo final

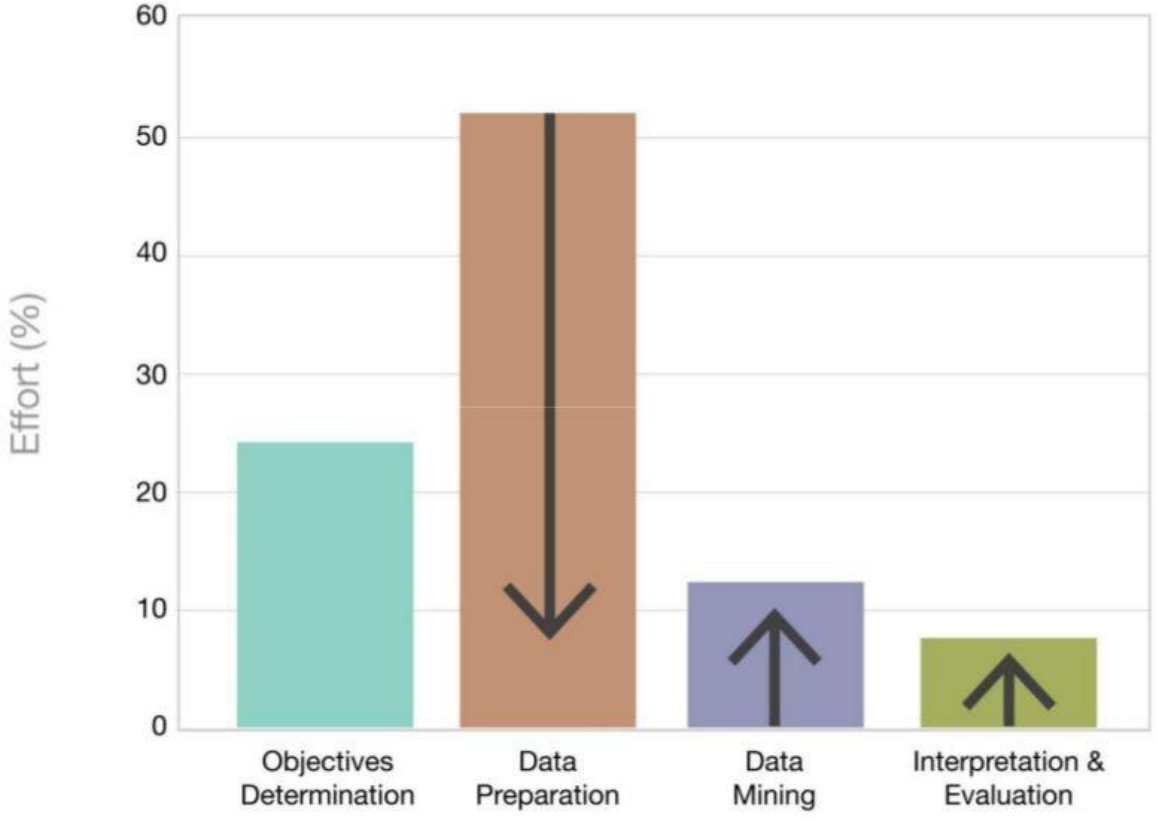
KDD - Etapas

Evaluación, interpretación y presentación de resultados.

- La fase de MD puede producir varias hipótesis de modelos
  - Es necesario establecer qué modelos son los más válidos
  - Criterios: los patrones descubiertos deben ser precisos, comprensibles, e interesantes (útiles, novedosos)
- Medidas de evaluación de modelos. Dependen de la tarea.
- La interpretación de los mejores modelos (visualización, simplicidad, posibilidad de integración, ventajas colaterales,...) ayuda a la selección del modelo final

Metodologías de gestión basadas en el dato

- Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos.



Metodologías de gestión basadas en el dato

- SEMMA

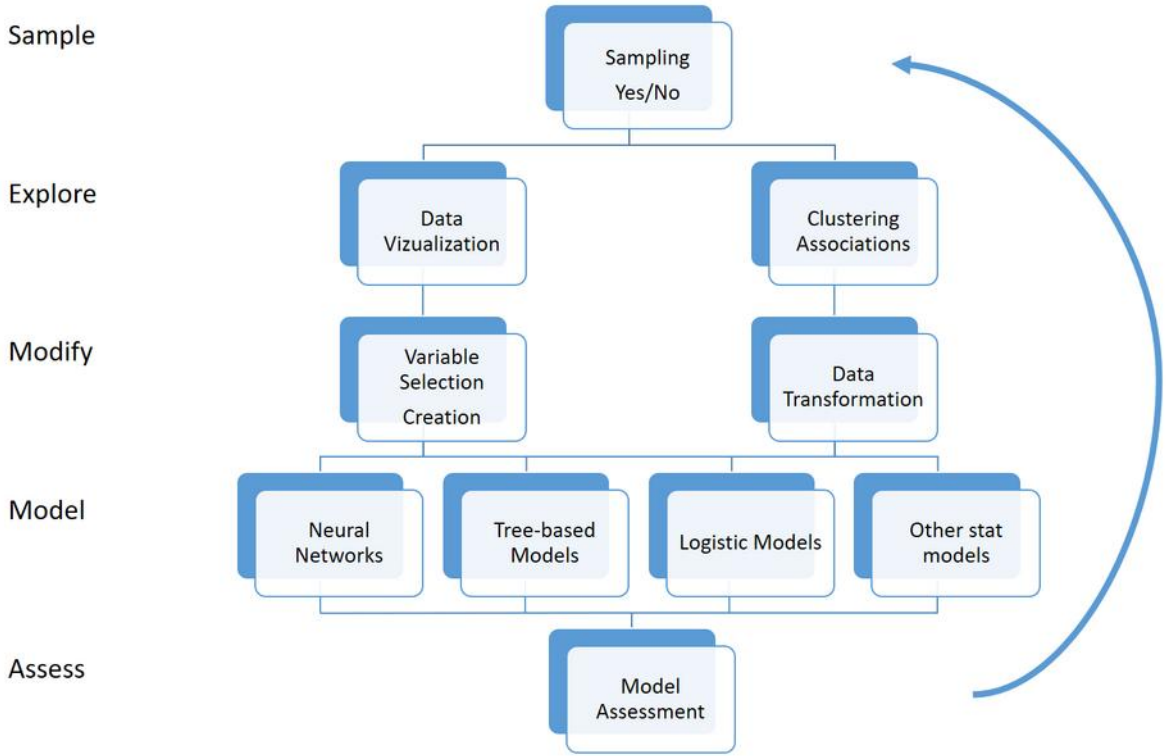


TABLA DE CONTENIDOS

Fuentes de datos, su tipología e importancia

- Existen varios tipos de datos:

- **No estructurados**
  - Texto, datos de vídeo, datos de audio, reclamaciones clientes, tweets, …
  - Para poder hacer uso de los datos no estructurados, es necesario transformarlos primero (con la herramienta que fuera).

- **Semiestructurados**
  - A veces se conocen como “multiestructurados”.
  - Tienen un formato y flujo lógico de modo que pueden ser entendidos pero el formato no es amistoso al usuario (HTML. XML…, datos de web logs

Fuentes de datos, su tipología e importancia

- Existen varios tipos de datos:

→ **No estructurados**

  - Texto, datos de vídeo, datos de audio, reclamaciones clientes, tweets, …
  - Para poder hacer uso de los datos no estructurados, es necesario transformarlos primero (con la herramienta que fuera).

→ **Semiestructurados**

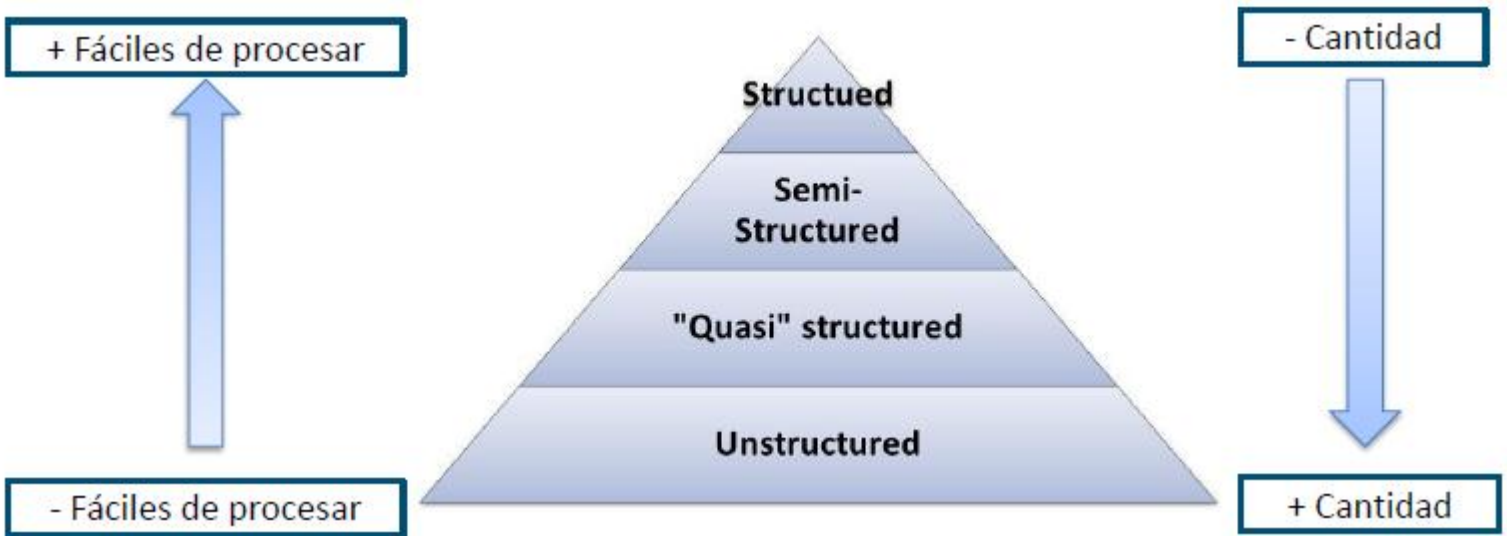
  - A veces se conocen como “multiestructurados”.
  - Tienen un formato y flujo lógico de modo que pueden ser entendidos pero el formato no es amistoso al usuario (HTML. XML…, datos de web logs

Fuentes de datos, su tipología e importancia

- **Estructurados**
  - Hojas de cálculo, BBDD tradicionales, …
  - Son los más fáciles de manejar.

- A su vez, en función de la ubicación desde la que se obtengan los datos, éstos podrán ser:
  - **Internos** (clientes, productos, empleados, proveedores, …)
  - **Externos** (OpenData, APIs, Redes Sociales, Otros)

Fuentes de datos, su tipología e importancia



Fuentes de datos, su tipología e importancia

- **Fuentes de datos abiertas**

https://opendata.euskadi.eus/inicio/

https://datos.gob.es/

TABLA DE CONTENIDOS

Gestión de los datos y su enriquecimiento

- Los datos como tales no nos aportan valor sí:
  - No están actualizados.
  - No están normalizados.
  - No tienen calidad suficiente.

  Todo ello implica que nuestros datos no son confiables.

Gestión de los datos y su enriquecimiento

- Para poder gestionar los datos, necesitaremos definir:

1)Lenguaje común, ¿qué significan cada uno de los indicadores?, ¿representan las mismas unidades, escalas, ...?

2)Cómo y con qué frecuencia se tienen que medir

3)Almacenamiento de los datos

4)Responsable del dato en cada etapa

5)Procesos de recogida y agregación a seguir

6)Criterios de validación del dato

Gestión de los datos y su enriquecimiento

- Objetivo:
  - Pasar de un proceso de **toma de decisiones basado en la experiencia** (subjetivo) a otro basado en los **datos (objetivo)**.
  - La toma de decisiones basa en los datos se conoce como Data-Driven

Gestión de los datos y su enriquecimiento

- No es gratis.

  Los procesos de gestión de los datos y su enriquecimiento, suponen el 80% de tiempo de un proyecto.

  Sólo el 20% del tiempo se dedica a obtención de conocimiento.

Gestión de los datos y su enriquecimiento

- Razones
  - **Datos disgregados** en diferentes fuentes de datos que es necesario consolidar
  - **Datos duplicados** y con valores dispares que es necesario consolidar
  - Fuentes de origen **poco accesibles**
  - **Datos desactualizados** o con frecuencia de actualización insuficiente
  - **Datos sensibles** que es necesario **anonimizar** para cumplir la normativa vigente
  - **Datos sin un responsable de negocio** que entienda lo que significa y los valide
  - Datos con **formato incorrecto**
  - Datos **incompletos**
  - Datos **no estandarizados**
  - Datos **no documentados**

TABLA DE CONTENIDOS

Del análisis de dato descriptivo al predictivo

- Análisis descriptivo.

  Se basa en obtener tendencias a partir de **datos existentes**.

  Suele ser el tipo de análisis más habitual en las empresas No data-driven.

  Técnicas como la **segmentación** se basan en análisis descriptivos.

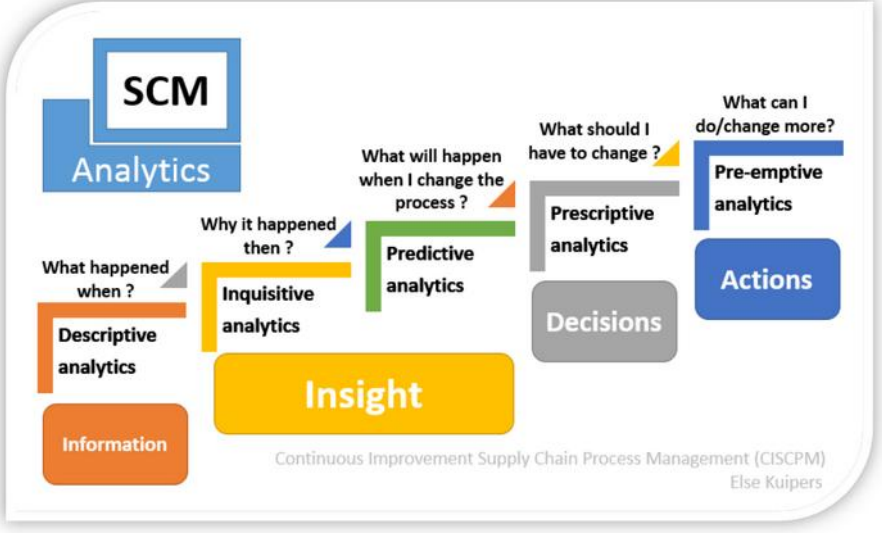
Del análisis de dato descriptivo al predictivo

- Análisis descriptivo.

  Puede tomarse como base para la realización posteriormente análisis predictivo.

  Nos permite responder preguntas del tipo: ¿Qué ha pasado?, ¿Cuándo ha pasado?, ¿Cómo se agrupan mis clientes?, ¿Qué productos o servicios se pueden asociar a otros productos o servicios?

  Análisis basado en hechos pasados.



Del análisis de dato descriptivo al predictivo

- Análisis inquisitivo.

  Análisis realizado mediante las técnicas de Business Intelligence existentes hasta la aparición del Big Data. **Aborda la historia de la empresa**

  Nos permite responder preguntas del tipo: ¿Porqué sucedió ésto en el pasado?

Del análisis de dato descriptivo al predictivo

- Análisis predictivo.

  Se basa en el uso de **técnicas estadísticas avanzadas y de predicción** para entender qué puede suceder en el futuro. Entraría en el mundo del Business Analytics.

  Nos permite responder preguntas del tipo: ¿Qué podría pasar? ¿Cómo afecta a mis ventas un cambio en los precios? ¿Cuál es la propensión al abandono de mis servicios por parte de mis clientes? ¿Cuánto venderé el año que viene?

TABLA DE CONTENIDOS

Terminología

- **Columna** → Incluye los datos de un determinado tipo. Todos los datos en ella deben tener la misma escala y tener un significado relativo.
  - Proceso calidad de datos
- **Fila** → Cada fila representa una entidad u observación.
  - Las columnas describen las propiedades de la fila.
- **Celda** → Valor contenido en una determinada fila y columna.
  - Booleano, categoría, entero, ...

Terminología

- **Perspectiva Estadística**
  - Variables dependientes
  - Variables independientes
- **Perspectiva Ciencias de la Computación**
  - Fila → Entidad, instancia, ejemplo...
  - Columna → Atributo, característica...

	A	B	C	D
1		Attribute 1	Attribute 2	Output Attribute
2	Instance 1	2.2	2.3	1
3	Instance 2	2.3	2.6	0
4	Instance 3	2.1	2	1
5				

Preparación de los datos

Las técnicas de minería de datos (generalmente)

- Trabajan con una única tabla de datos
- Trabajan únicamente con los datos que hay en esa tabla
  - No consideran información (por evidente que pueda ser) más allá de los valores almacenados en la tabla
- Si tenemos en un campo fecha “01/01/2016”

Preparación de los datos

- Si tenemos en un campo fecha “01/01/2016”
  - No saben que ese día fue viernes
  - No saben que el viernes es parte del fin de semana (o no)
  - No saben que ese día es Año Nuevo
  - No saben que Año Nuevo es festivo
  - No saben que el segundo “‘01” significa “enero”
- Por eso es importante la preparación de datos para su procesamiento por estas técnicas.

Preparación de los datos

- Por eso hablaremos de **datasets**, en lugar de tablas, bases de datos...
- Por eso hablaremos de **atributos** en lugar de columnas, campos...
- Estos atributos pueden extraerse de la base de datos
  - En (muchas) ocasiones deberán de calcularse a partir de la información almacenada
- Por eso hablaremos de **instancias/ejemplos**, en lugar de tuplas/filas
  - Cada ejemplo es una muestra relevante para el problema a abordar

Preparación de los datos

- **Son dependientes del problema a abordar. Deben ser relevantes para la pregunta a responder**
- Si tengo datos de **ventas**, y **quiero estudiar clientes**, para cada cliente, puedo extraer (entre otros)
  - Número de compras
  - Antigüedad
  - Frecuencia de compra (tiempo promedio entre compra y compra)
  - Frecuencia de compra (compras por mes)
  - Compra promedio (gasto promedio)
  - Categoría de productos adquiridos más frecuentemente
  - Preferencias de envío (ordinario o urgente)
  - Tipo de cliente (Normal o premium)
  - ...

Preparación de los datos

- Si tengo datos de **ventas**, y quiero **responder**:
  - Las ventas por mes
    - “01/01/2022” → Atributo MES = Enero
  - Quiero ver si recibo más compras a primeros de mes
    - “01/01/2022” → Atributo DIA_DEL_MES = 1
  - Quiero ver si hay más ventas días festivos
    - “01/01/2022” → Atributo FESTIVO = TRUE
  - Quiero estudiar la serie temporal de ventas
    - “01/01/2022” → Atributo TIEMPO = XX (según empiecen mis datos)
  - Quiero ver si las ventas se elevan en fines de semana
  - “01/01/2022” → Atributo FIN_DE_SEMANA = FALSE

Preparación de los datos

- Si tengo **documentos**, puedo extraer (entre otros):
  - Número de palabras
  - Idioma
  - ¿Contiene Imágenes?
  - Temática
  - Cuántas veces aparece el término “Big Data” (o cualquier otro). Ojo:
    - “Big Data” ≠ “big data” ≠ “big-data” ≠ “Big data”
    - Requerirá de una preparación especial (dejar caracteres, minúscula, sin acentos...)
  - “Candidato” ≠ “Candidata”
    - Se puede extraer la raíz “Candidat” → “Candidato” = “Candidata” = “Candidatura”
  - Procedencia del autor
  - Fecha de creación/modificación → antigüedad

Preparación de los datos

- Si tengo **enlaces**, puedo extraer (entre otros):
  - Dominio
  - Activo o caído
  - Sitio Web
  - Contiene enlaces

Preparación de los datos

- Si tengo **imágenes**, puedo extraer (entre otros):
  - Con software especializado:
    - Objetos que aparecen
  - Sin software especializado:
    - Color promedio, mediano, modal, ...
    - Color promedio, mediano, modal, ... en una determinada zona
    - Porcentaje de rojo, verde, azul, ...

Preparación de los datos

- Para cualquier atributo es importante tener siempre presente **la información que nos interesa y la que no**.

  - ¿Nos interesa la fecha de nacimiento del cliente?
    - ¿Sólo nos interesa si es mayor de 65 o no?
    - ¿Sólo nos interesa si su edad está en los rangos 20-30, 30-40, 40-50, 50-60...?
    - ¿Sólo nos interesa si su edad está entre 20-30?
    - ¿Sólo nos interesa su signo zodiacal?
    - ¿Sólo nos interesa la cercanía a su cumpleaños?

  - ¿Nos interesa su localidad?
    - ¿Sólo nos interesa si es nacional o internacional?
    - ¿Sólo nos interesa la provincia?
    - ¿Sólo nos interesa el país?
    - ¿Sólo nos interesa la distancia desde nuestro almacén?
    - ¿Sólo nos interesa el idioma oficial de su localidad?

Modelos y Métodos

- Para cualquier atributo es importante tener siempre presente **la información que nos interesa y la que no**.

  - ¿Nos interesa la fecha de nacimiento del cliente?
    - ¿Sólo nos interesa si es mayor de 65 o no?
    - ¿Sólo nos interesa si su edad está en los rangos 20-30, 30-40, 40-50, 50-60...?
    - ¿Sólo nos interesa si su edad está entre 20-30?
    - ¿Sólo nos interesa su signo zodiacal?
    - ¿Sólo nos interesa la cercanía a su cumpleaños?

  - ¿Nos interesa su localidad?
    - ¿Sólo nos interesa si es nacional o internacional?
    - ¿Sólo nos interesa la provincia?
    - ¿Sólo nos interesa el país?
    - ¿Sólo nos interesa la distancia desde nuestro almacén?
    - ¿Sólo nos interesa el idioma oficial de su localidad?

Modelos y métodos

- Un modelo es, en general, una función o una estructura de datos
  - Que puede representar el conocimiento subyacente en un conjunto de datos.
  - Dependiendo del objetivo del problema, tipo de variables de entrada, tipo de salida esperada, complejidad de los datos, etc. habrá modelos más o menos apropiados
- Es una representación específica hecha a partir de los datos
  - Árbol de decisión, Red Neuronal, Conjunto de coeficientes,...

Modelos y métodos

Un **algoritmo** es una **secuencia de pasos con un fin**.

- Un algoritmo de aprendizaje se encarga de que el modelo aprenda de los datos (que el modelo se ajuste a los datos)

Es el proceso seguido para obtener un modelo

- C4.5
- Regresión lineal por mínimos cuadrados
- Probar coeficientes “a lo loco” hasta que esto funcione

Copyright (c) 2022 Germán Alonso Lascurain
This work (but the quoted images, whose rights are reserved to their owners*) is licensed under the Creative Commons “Attribution-ShareAlike” License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/

Germán Alonso Lascurain
german@campus2b.com
germanalonso@opendeusto.es