Analítica de datos

Juan Fernández-Herrerín Álvarez

La analítica de datos

- Tradicionalmente se usa término analítica de datos → actualmente lo puedes encontrar como minería de datos.
- **Análisis de datos**: busca analizar la información para encontrar:
  - hechos
  - relaciones
  - Patrones
  - Tendencias
- Todo ello con el objetivo de generar conocimiento y apoyar en la mejor medida posible en la toma de decisiones.

La analítica de datos

- La **analítica de datos** es un concepto más amplio que el análisis, ya que incluye adicionalmente actividades, algunas de las cuales ya hemos visto:
  - Carga de datos desde las fuentes
  - Limpieza de los datos
  - Integración de los datos
  - Gobierno de los datos
    - Disponibilidad → datos disponibles cuando se necesitan
    - Usabilidad → datos válidos para el objetivo buscado
    - Integridad → datos correctos
    - Seguridad → acceso solo a autorizados

La analítica de datos

- Inteligencia de negocio (BI): uso de la analítica de datos para apoyar la toma de decisiones por parte del nivel táctico o estratégico.
- Otros casos de uso posibles de la analítica de datos:
  - Ciencia: comprensión de causas de algunos fenómenos y poder reaccionar de forma temprana.
  - Servicios: disminuir los costes y mejorar calidad.

Niveles de analítica de datos

- Podemos diferenciar 4 niveles:
  - Análisis descriptivo: ¿Qué ha ocurrido?
  - Análisis diagnóstico: ¿Por qué ha ocurrido?
  - Análisis predictivo: ¿Qué ocurrirá?
  - Análisis prescriptivo: ¿Qué hacer para que ocurra?

Análisis descriptivo

- Intenta describir qué es lo que ha ocurrido. Estudia el pasado.
- Produce como resultado informes o cuadros de mando estáticos mediante consultas a sistemas operacionales (OLTP). Ejemplo: CRM, ERP, Facturador.
- Responde a preguntas como:
  - ¿Cuál ha sido el beneficio mensual en los últimos 12 meses?
  - ¿Cuántas llamadas hemos recibido en el servicio de atención al cliente?
  - ¿Cuál es el producto con mayor margen?

Análisis diagnóstico

- Intenta entender la causa de lo que ha ocurrido o está ocurriendo.
- Detecta la información relacionada con esta causa.
- Normalmente debe obtenerse información de varias fuentes, que se almacenan en sistemas OLAP para facilitar el análisis.
- Mediante consultas interactivas, los usuarios emplean herramientas interactivas que permiten identificar patrones.
- Responde a preguntas como:
  - ¿Por qué hay más llamadas de reclamación en Galicia que en Madrid?
  - ¿Por qué están bajando las ventas de un producto concreto?

Análisis predictivo

- Busca predecir lo que ocurrirá en el futuro, haciendo uso de modelos predictivos → Por ejemplo, Machine Learning.
- Podemos usarlo para detectar tanto riesgos como oportunidades.
- Ejemplos de uso:
  - ¿Con qué probabilidad un cliente contratará un producto si se lo ofrecemos?
  - Si un cliente ha comprado un producto, ¿qué otros productos le pueden interesar?

Análisis prescriptivo

- A partir de los resultados del análisis predictivo, prueba diferentes acciones de forma automática para prescribir la mejor acción a tomar.
- Busca encontrar la mejor acción, pero también proporcionar información de por qué es la mejor.
- Responde a preguntas como:
  - De una lista de productos, ¿cuál se venderá mejor en Galicia?
  - ¿Qué mes es mejor para lanzar un nuevo producto concreto?

Predictivo vs Prescriptivo

- Mediante el análisis predictivo conseguimos realizar predicciones de lo que ocurrirá en un futuro.
- Sin embargo, el análisis prescriptivo busca entender por qué va a suceder un hecho para poder encontrar las palancas que hagan que ocurra lo que nosotros queremos.
- El análisis prescriptivo parte del predictivo, y tiene mayor complejidad.

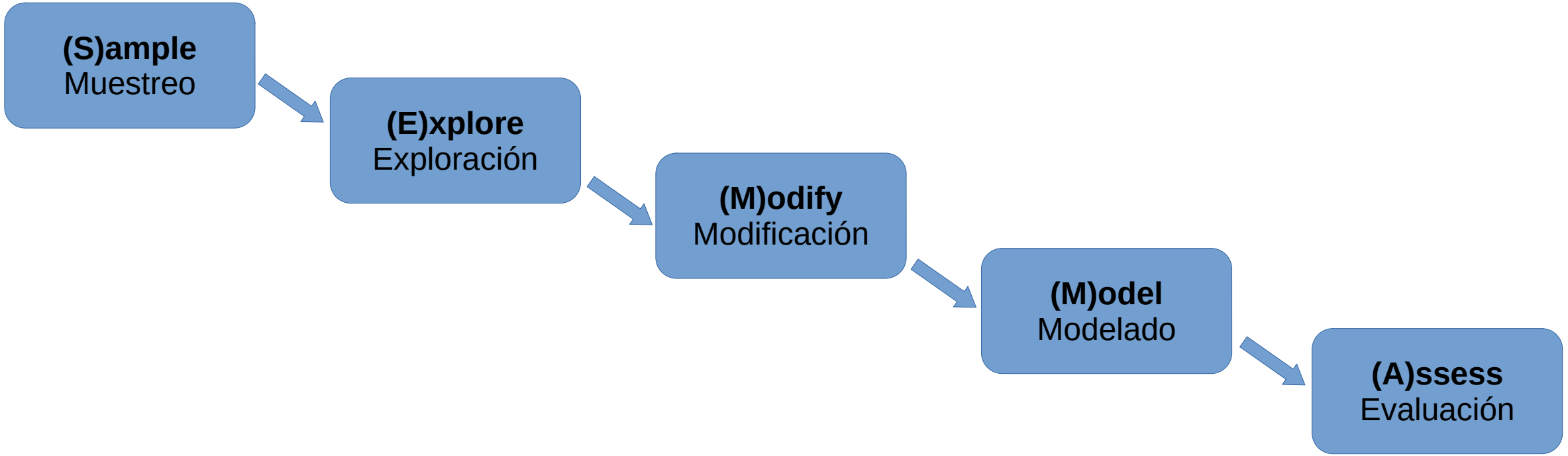
Metodologías en minería de datos

- Con la minería de datos, utilizamos modelos predictivos.
- Se emplea tanto para análisis predictivo como para un posible análisis prescriptivo posterior.
- Para implantar un proceso de minería de datos existen diferentes metodologías. Veremos brevemente dos de ellas:
  - SEMMA: Sample, Explore, Modify, Model, Assess.
  - CRISP-DM: Cross-Industry standard process for data mining.
- Otras metodologías: DATLAS, KDD, ASUM-DM

SEMMA (I)

- Definido por el SAS Institute → **S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess.
- De trata de una empresa que se dedica al desarrollo de software de inteligencia de negocio.
- https://www.sas.com/es_es/home.html
- Esta metodología define una lista de pasos que se deben ejecutar secuencialmente.
- Aunque se acepta como metodología, realmente nace como la organización lógica del conjunto de herramientas de esta empresa para la minería de datos.

SEMMA (II)


(S)ample
Muestreo
(E)xplore
Exploración
(M)odify
Modificación
(M)odel
Modelado
(A)ssess
Evaluación

SEMMA (III)

- **Sample**: Muestreo de los datos.
  - Debemos escoger un subconjunto de los datos:
    - Lo más pequeño posible para poder usarlo de forma eficiente.
    - Con suficiente información para representar a todo el conjunto.
- **Explore**: Exploración de los datos
  - Buscar tendencias y anomalías que ayuden a comprender los datos mejor y encontrar sus relaciones.
  - Exploración, tanto gráfica como numérica de los principales estadísticos.

SEMMA (IV)

- **Modify**: Modificación de los datos.
  - Identificación de variables existentes.
  - Creación de nuevas variables.
  - Transformar variables existentes.
  - Tratar valores atípicos.
  - Eliminar variables que no aporten valor.

SEMMA (V)

- **Model**: Construcción de modelos.
  - Aplicación de técnicas de minería para producir modelos válidos:
    - Redes neuronales.
    - Árboles de decisión.
    - Modelos logísticos.
    - Máquinas de vectores de soporte.
    - Etc.

SEMMA (VI)

- **Assess**: Exploración de los datos.
  - Evaluación de la utilidad y confiabilidad de los resultados.
  - Mediante diferentes conjuntos de datos para poder evaluar su utilidad práctica.
- Principal crítica a esta metodología: no tiene en cuenta el conocimiento de los datos ni los requisitos de negocio dentro de ninguna de las fases.
  - SAS argumenta que en ningún momento nació con el objetivo de ser una metodología genérica.

CRISP-DM (I)

- CRISP-DM: **C**Ross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining.
- Nace dentro de un proyecto de la UE, bajo la iniciativa ESPIRIT (programa de financiación para la investigación y desarrollo tecnológico en el campo de las TI).
- En este casí sí es un estándar abierto, que ha sido adoptado por diferentes fabricantes (SPSS, IBM, Daimler, Teradata, ...).
- Esta metodología sí tiene en cuenta los requisitos de negocio y el despliegue en producción dentro de sus fases.
- Es la metodología más usada.

CRISP-DM (II)

- Define un proceso de minería de datos en 6 fases, que se ejecutan de forma cíclica:
  - Comprensión del negocio.
  - Comprensión de los datos.
  - Preparación de los datos.
  - Modelado.
  - Evaluación.
  - Despliegue.

Fuente: Kenneth Jensen (CC BY-SA)

CRISP-DM (III)

- En esta metodología las fases no se definen como una secuencia, sino que se ejecutan de forma cíclica.
- En cada iteración se aplican lecciones aprendidas de la anterior para mejorar el proceso.
- Además, existen transiciones entre las distintas fases, como se puede observar en la imagen del slide anterior.
- Por ejemplo, una vez comprendidos los requisitos de negocio, avanzamos al análisis de los datos para comprenderlos.
  - Pero si encontramos información en esta exploración, podemos tener que volver a hablar con negocio para refinar los requisitos.

CRISP-DM (IV)

- **Comprensión del negocio**: debemos entender qué se busca.
  - se determinan los objetivos a nivel de negocio
  - análisis de la situación de partida.
  - inventario de recursos.
  - análisis de coste-beneficio.
  - realización del plan de proyecto.
- **Comprensión de los datos**: entendemos los datos disponibles.
  - recolección de los datos.
  - descripción y exploración de los datos.
  - análisis de la calidad de los datos.

CRISP-DM (V)

- **Preparación de los datos**: preparamos los datos para ser usados.
  - selección de los datos útiles.
  - limpieza de datos.
  - creación de variables.
  - integración de datos.
  - transformación de datos.

CRISP-DM (VI)

- **Modelado**: equivalente a la misma fase en SEMMA, definimos modelos.
  - experimentamos con distintos modelos.
  - definimos parámetros.
  - probamos los modelos.
- **Evaluación**: analizamos los resultados obtenidos.
  - Se revisa si los resultados son positivos o negativos de acuerdo a los objetivos buscados.
  - Se toman decisiones de los siguientes pasos (continuar, volver a empezar).

CRISP-DM (VII)

- **Despliegue**: se activa el proyecto.
  - se genera un plan de desarrollo.
  - se genera plan de mantenimiento.
  - elaboración de informa final.
  - presentación de resultados obtenidos.

Comparativa de ambos modelos

SEMMA	CRISP-DM
	Comprensión del negocio
Muestreo	Comprensión del dato
Exploración	
Modificación	Preparación de datos
Modelado	Modelado
Evaluación	Evaluación
	Despliegue