Ciencia de Datos y Aprendizaje Automático

1. Introducción a la Minería de Datos y Ciencia de Datos

José Hernández-Orallo, DSIC, UPV, jorallo@dsic.upv.es

Esquema

- 1.1. Motivación
  - ¿Qué es todo esto? Sopa de letras
- 1.2. 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
- 1.3. El proceso de extracción de conocimiento
  - El proceso D2K
  - CRISP-DM
- 1.4. Tareas, técnicas y herramientas
  - Tareas
  - Técnicas
  - Herramientas

Introducción

- ¿Qué es todo esto?

Minería de datos

Análisis de datos (Predictivo)

Ciencia de datos

Inteligencia empresarial

Análisis de datos (Inteligente)

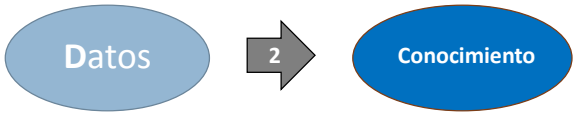
Extracción de conocimiento (desde bases de datos), KDD

Big Data

3

Introducción

- ¿Qué es todo esto?
  - SIMPLE…



4

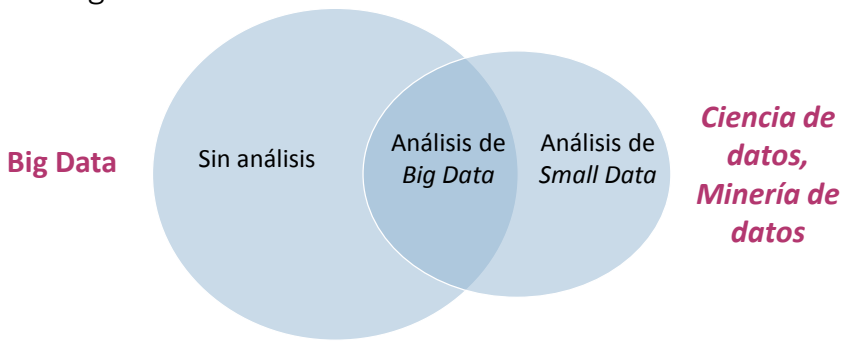
Introducción

- ¿Qué es todo esto?:
  - Minería de datos:
    - *La minería de datos está asociada a herramientas e inteligencia empresarial.*
  - Ciencia de datos:
    - *La ciencia de datos está asociada con una profesión inquisitiva.*
  - Análisis de datos (Inteligente)
    - *Parecido a la minería de datos, usado mayormente para estadística.*
  - Análisis predictivo (de datos)
    - *Un nombre más elegante para el análisis de datos.*
  - Big Data
    - *No todos los proyectos Big Data requieren análisis.*
    - *No todos los proyectos de ciencia de datos requieren una infraestructura para Big Data.*
  - Extracción de conocimiento (desde bases de datos), KDD
    - *El término clásico para enfatizar todo el proceso.*

5

Introducción

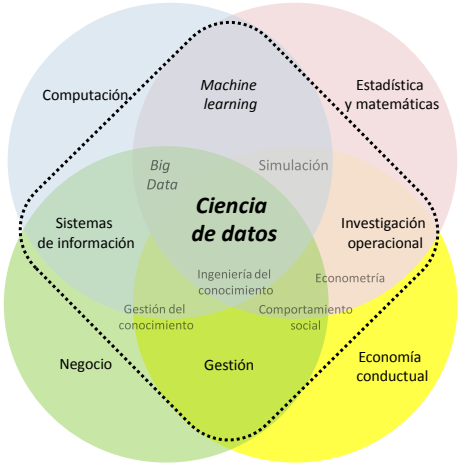
- No toda la ciencia de datos es Big Data.
- No todo el Big Data es ciencia de datos.



6

Introducción

- ¿Qué es todo esto?:



7

Minería de datos

- Ejemplos:
  - **AGENTE BANCARIO:** ¿Debo ofrecerle una hipoteca a este cliente?
  - **GERENTE DE SUPERMERCADO:** Cuando mis clientes compran huevos, ¿cogen también aceite?
  - **GERENTE DE PERSONAL:** ¿Qué tipo de empleados tengo?
  - **SUPERVISOR DE FÁBRICA:** ¿Cuántos fallos para X módulo esperamos cada mes?

8

Minería de datos

- **AGENTE BANCARIO:**

¿Debo ofrecerle una hipoteca a este cliente?

Histórico:

cId	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

9

Minería de datos

- **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	…
…	…	…	…	…	…	…	…	…	…

Minería de datos

Patrón / Modelo:

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

10

Minería de datos

- **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	sí	no	0	alquilado	no	7	15	M
2	20000	no	sí	1	alquilado	sí	3	3	F
3	15000	sí	sí	2	propietario	sí	5	10	M
4	30000	sí	sí	1	alquilado	no	15	7	F
5	10000	sí	sí	0	propietario	sí	1	6	M
6	40000	no	sí	0	alquilado	sí	3	16	F
7	25000	no	no	0	alquilado	sí	0	8	M
8	20000	no	sí	0	propietario	sí	2	6	F
15	8000	no	sí	0	alquilado	no	3	2	M
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Minería de datos

Patrón / Modelo:

- **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.

11

Minería de datos

- **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:

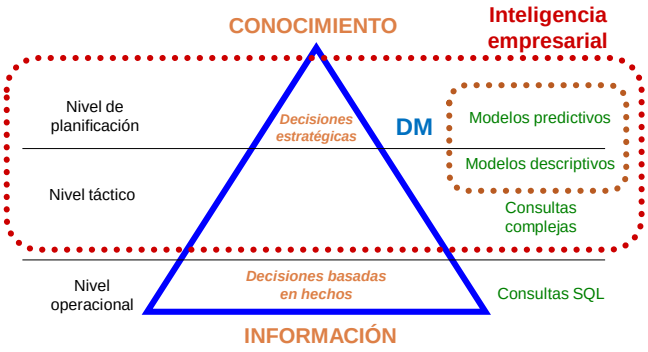
Modelo linear: Fallos de X para el siguiente mes:

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

12

Minería de datos

- El papel en la inteligencia empresarial



13

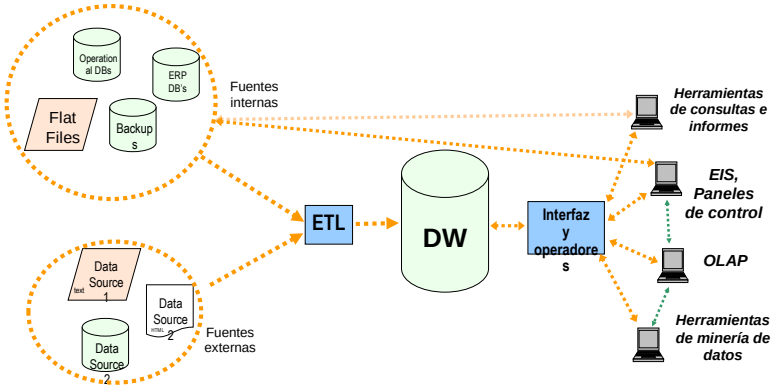
Minería de datos

- Inteligencia empresarial
  - La colección de tecnologías de la información que pueden proveer a una organización de los conocimientos necesarios para tomar decisiones estratégicas.
  - Compuesto de:
    - Herramientas de almacenaje de datos y OLAP.
    - Paneles de control y otras herramientas de supervisión.
    - Herramientas de minería de datos.
  - La minería de datos comienza con un objetivo empresarial
    - Orientado a objetivos

14

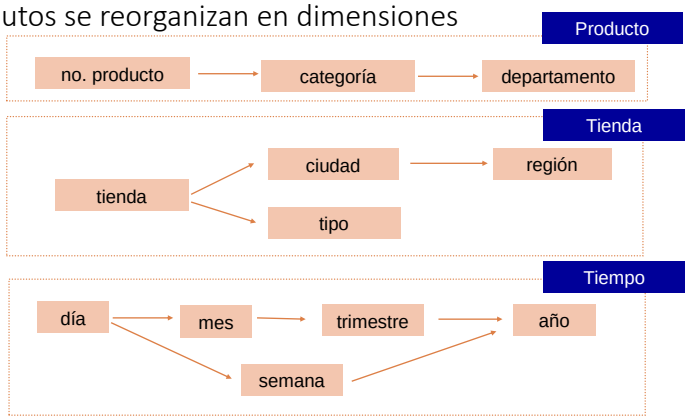
Minería de datos

- Inteligencia empresarial
  - Arquitectura clásica



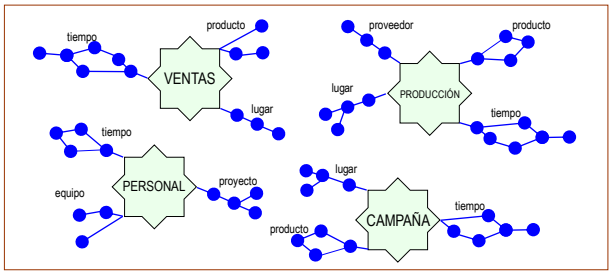
Minería de datos

- Almacenaje de datos multidimensional
  - Los atributos se reorganizan en dimensiones



Minería de datos

- Almacenaje de datos multidimensional
  - Cada diagrama se llama *datamart*:



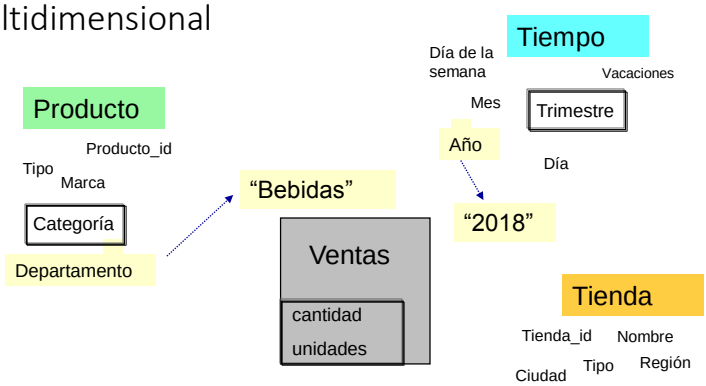
17

Minería de datos

- Almacenaje de datos multidimensional

Ejemplo de consulta OLAP:

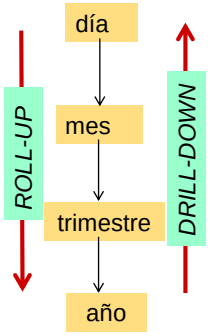
"Ventas totales en el año 2018 del departamento de "Bebidas", por categoría y trimestre".

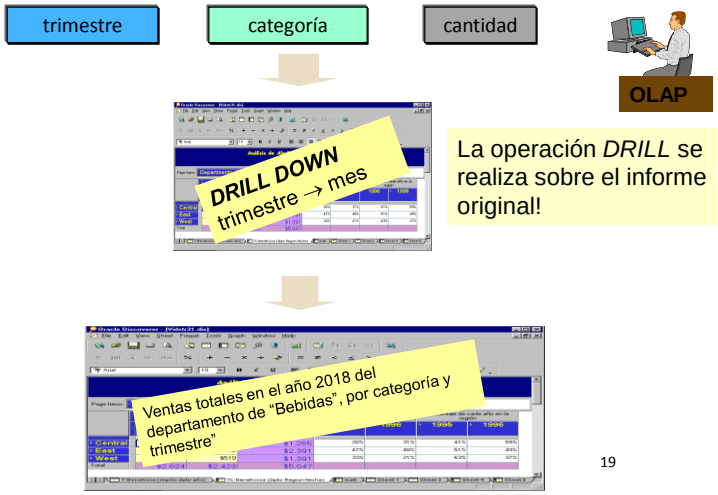


18

Minería de datos

- Operadores OLAP





La operación *DRILL* se realiza sobre el informe original!

19

Minería de datos

- OLAP vs Minería de datos

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?

20

Ciencia de datos

- Ciencia de datos: aparece como una actitud…
  - Director de sistemas de la información (*Chief information officer*, CIO):
    - Término tradicional para los altos ejecutivos de Sistemas de la información y Tecnologías de la información.
  - *Data Manager*
    - Término tradicional para el responsable de la gestión de bases de datos.
  - Director de datos (*Chief Data Officer*, CDO)
    - "[…] *responsable de la dirección de la empresa y la utilización de la información como un activo, a través del procesamiento de datos, el análisis, la extracción de datos, el intercambio de información y otros medios*."
  - Científico de datos
    - "[…] *un conjunto integrado de habilidades que abarca matemáticas, aprendizaje automático, inteligencia artificial, estadísticas, bases de datos y optimización, junto con un profundo conocimiento de la elaboración de problemas para diseñar soluciones efectivas*".

21

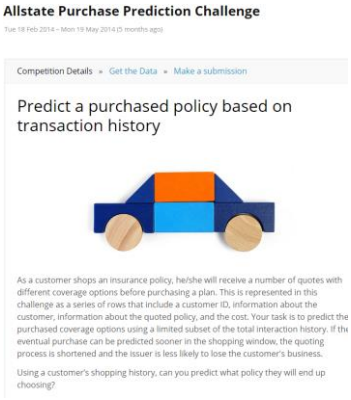
Ciencia de datos

- Mis datos son valiosos para mi (in → in).
  - Datos internos útiles para la organización.
  - Inteligencia empresarial clásica… Muchas oportunidades todavía.
- Esos datos son valiosos para mi (out → in).
  - Datos externos útiles para la organización.
  - Medios sociales, Internet, datos abiertos, … Muchas oportunidades nuevas.
- Mis datos son valiosos para otros (in → out).
  - Datos internos útiles para otras organizaciones.
  - Mis datos tienen utilidad para otros, … Muchas oportunidades nuevas.
- Esos datos son valiosos para otros (out → out).
  - Datos externos útiles para otras organizaciones.
  - Estos datos tienen utilidad para otros, … ¡Científico de datos *freelancer*!
- Creando datos (∅ → out).
  - Coleccionar datos que pueden tener valor. ¡Emprendedor de datos!

22

Ciencia de datos

- Ejemplos de productos basados en datos (in → in):
  - Una compañía de seguros de automóviles, Allstate, quiere predecir la póliza que se comprará dado el historial de transacciones.



https://www.kaggle.com/c/allstate-purchase-prediction-challenge

Ciencia de datos

- Ejemplos de productos basados en datos (in → out):
  - *Smart Steps* son datos en tiempo real recolectados por subsidiarios de Telefónica (Movistar, O2, …).
  - Venden los datos, las herramientas y la habilidad de analizar y representarlos a otras compañías.

http://dynamicinsights.telefonica.com/488/smart-steps

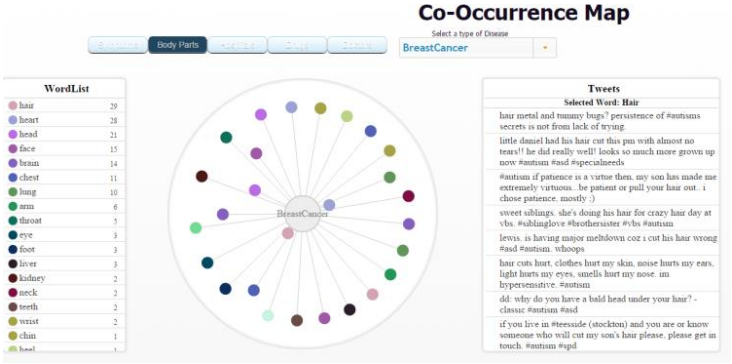
Ciencia de datos

- Ejemplos de productos basados en datos (out → in):
  - La Comisión de Crimen de Detroit (DCC) reconoció que muchos delincuentes publicaban sobre sus crímenes en varias plataformas de redes sociales, anunciando potenciales planes, alardeando de drogas y armas en Facebook, Twitter e Instagram, y organizando su próximo movimiento. Sin embargo, al hacer tal información transparente al público, el DCC decidió aprovechar esta información abierta al asociarse con Semantria para introducir análisis de texto que le permitiría al equipo rastrear criminales, actividades y consecuencias.

  - http://www.1to1media.com/weblog/2014/04/using_social_cues_to_combat_cr.html#sthash.BAlMAaOs.dpuf

Ciencia de datos

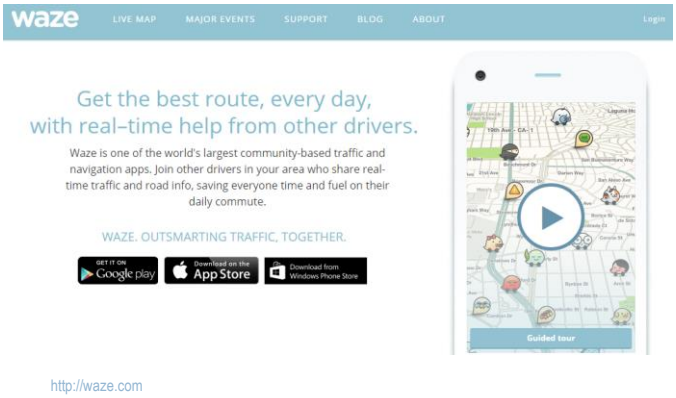
- Ejemplos de productos basados en datos (out → out):
  - *Healthcaredataanalysis.org* fue un experimento para mostrar que los tweets podrían brindar información valiosa sobre el efecto de las enfermedades y la relación entre los síntomas y las drogas.



http://www.healthcaredataanalysis.org/demos/index.html

Ciencia de datos

- Ejemplos de productos basados en datos (∅ → out):
  - Al recopilar y compartir información de los controladores, se creó una aplicación para brindar información y consejos en tiempo real sobre las rutas.



http://waze.com

Ciencia de datos

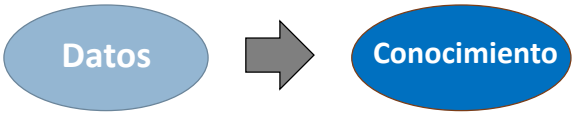
- Datos de telecomunicaciones
  - Valioso para comerciantes, tráfico, ayuntamiento, policía…
- Otros datos de geolocalización(*Flickr*, Instagram, *Wikiloc*, …)
  - Valioso para agencias de viaje…
- Datos en consumo de energía
  - Valioso para anuncios de televisión…
- Datos del transporte público (bus, metro, tren, taxi, tráfico, …)
  - Valioso para turismo, consumo, contaminación, comercio…
- Datos de redes sociales.
  - Valioso para casi todo…
- Datos de uso de tarjetas de crédito
  - Valioso para comercios, ayuntamientos, …
- Datos de policía
  - Valioso para aseguradoras, agentes inmobiliarios, …
- Datos comerciales (*Amazon*, *Ebay*, *segundamano.es*, …)
  - Valioso para salud, demografía, sociología…
- Datos climatológicos
  - Valioso para comercios.
- Datos de búsquedas web.
  - Valioso para casi todo.

El proceso D2K

- Siempre ha estado ahí
  - *Extracción de conocimiento desde bases de datos.*

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

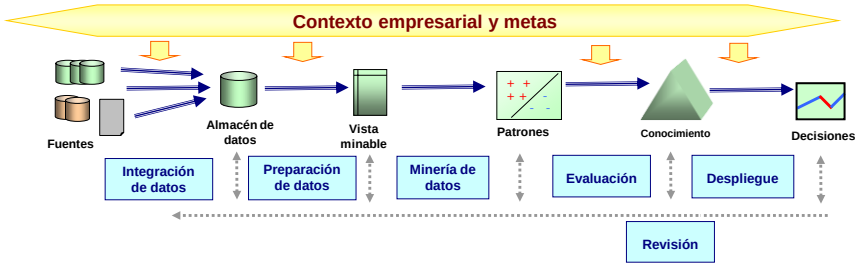
(Fayyad et al. 1996)



29

El proceso D2K

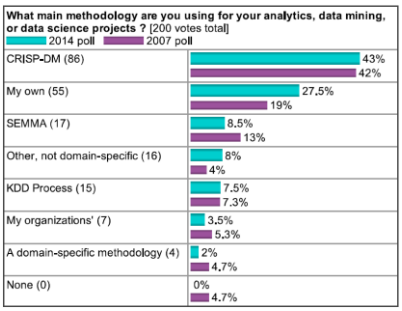
- El proceso



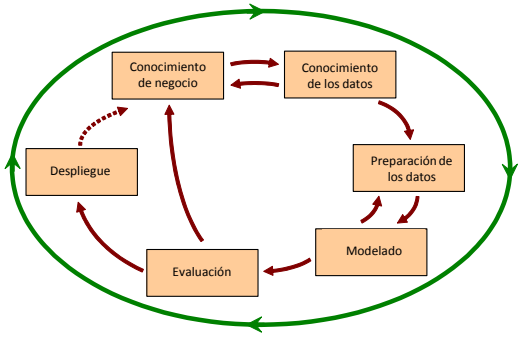
30

El proceso D2K: CRISP-DM

- La metodología
  - CRISP-DM sigue siendo la metodología más común:



* Fuente: kdnuggets.com



31

El proceso D2K: CRISP-DM

- Comprensión de negocio:
  - Comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial. Subfases:
    - **establecimiento de objetivos** (contexto inicial, objetivos y criterios de éxito)
    - **evaluación de la situación** (recursos, requerimientos, suposiciones, restricciones, riesgos y contingencias, terminología y costes y beneficios)
    - **establecimiento de los objetivos de la minería de datos** (planteamiento de los objetivos y criterios de éxito de la minería de datos)
    - **creación del plan de proyecto** (plan de proyecto y evaluación inicial de herramientas y técnicas).

El proceso D2K: CRISP-DM

- Comprensión de los datos:
  - Recopilar y familiarizarse con los datos, identificar los problemas de calidad y discernir datos potenciales o subconjuntos que pueden ser interesantes de analizar (de acuerdo con los objetivos empresariales de la fase anterior) Subfases:
    - **recopilación inicial de datos** (informe de recopilación),
    - **descripción de los datos** (informe de descripción),
    - **exploración de los datos** (informe de exploración)
    - **verificación de la calidad de los datos** (informe de calidad).

33

El proceso D2K: CRISP-DM

- Preparación de los datos:
  - El objetivo de esta fase es obtener la "vista minable". Aquí encontramos: integración, selección, limpieza y transformación. Subfases:
    - **selección de datos** (motivos de inclusión/exclusión),
    - **limpieza de datos** (informe de limpieza de datos),
    - **construcción de datos** (atributos derivados, archivos generados),
    - **integración de datos** (mezcla de datos)
    - **formateo de datos** (datos reformados).

34

El proceso D2K: CRISP-DM

- Modelado de los datos:
  - Es la aplicación de técnicas de modelado o minería de datos a las vistas minables anteriores. Subfases:
    - **selección de la técnica de modelado** (técnica de modelado, suposición de modelado),
    - **diseño de evaluación** (diseño de prueba),
    - **construcción del modelo** (parámetros elegidos, modelos, descripción del modelo)
    - **evaluación del modelo** (medidas del modelo, revisión de los parámetros escogidos).

35

El proceso D2K: CRISP-DM

- Evaluación:
  - Es necesario evaluar (desde el punto de vista de la meta) los modelos de la fase anterior. En otras palabras, si el modelo es útil para responder algunos de los requisitos comerciales. Subfases:
    - **evaluación del resultado** (evaluación de los resultados de la minería de datos, modelos aprobados),
    - **revisar el proceso** (proceso de revisión)
    - **establecimiento de los siguientes pasos** (lista de posibles acciones, decisiones).

36

El proceso D2K: CRISP-DM

- Despliegue:
  - La idea es explotar el potencias de los modelos extraídos, integrarlos en los procesos de toma de decisiones de la organización, repartir informes sobre el conocimiento extraído, etc. Subfases:
    - **planificación del despliegue** (plan de despliegue),
    - **planificación del mantenimiento y monitorización** (plan de monitorización y mantenimiento),
    - **creación del informe final** (informe final, presentación final),
    - **revisión del proyecto** (documentación de la experiencia).

37

Tareas, técnicas y herramientas

- Tarea
  - Predictiva: (tenemos una variable de salida)
    - *Clasificación/Categorización:* la variable de salida es nominal.
    - *Regresión:* la variable de salida es numérica.
  - Descriptiva: (no hay variable de salida)
    - *Clustering:* el objetivo es descubrir grupos en los datos.
    - *Análisis exploratorio:*
      - *Reglas de asociación, dependencias funcionales:* las variables son nominales.
      - *Análisis factorial/de correlación, análisis de dispersión, análisis multivariable:* las variables son numéricas.

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
…	…	…	…	…	…	…	…

38

Tareas, técnicas y herramientas

- Relación tarea/técnica:

TÉCNICA	PREDICITVA / SUPERVISADA		DESCRIPTIVA / NO SUPERVISADA		
	Clasificación	Regresión	*Clustering*	Reglas de asociación	Otros (factorial, correlación…)
Redes neuronales	✓	✓	✓*		
Árboles de decisión	✓(c4.5)	✓(CART)	✓		
Kohonen			✓		
Regresión linear (local, global), exp..		✓			
Regresión logística	✓				
K-means	✓*		✓		
A Priori (associaciones)				✓	
Análisis factorial, análisis multivariable					✓
CN2	✓				
K-NN (vecinos más próximos)	✓		✓		
FBR	✓				
Clasificadores básicos	✓	✓			

39

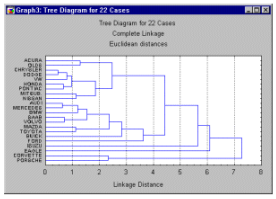
Tareas, técnicas y herramientas

- Técnicas descriptivas:
  - Correlación y asociaciones (análisis exploratorio, *link analysis*):
    - Coeficiente de correlación (cuando los atributos son numéricos):
      - Ejemplo: la desigualdad de distribución de la riqueza y el índice de criminalidad están positivamente relacionados.
    - Asociaciones (cuando los atributos son nominales).
      - Ejemplo: el tabaco y el alcohol están relacionados.
    - Dependencias funcionales: Asociación unidireccional.
      - Ejemplo: el nivel de riesgo de enfermedades cardiovasculares depende de el tabaco y el alcohol (entre otras cosas).

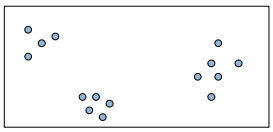
40

Tareas, técnicas y herramientas

- Técnicas descriptivas:
  - *Clustering*:
    - Jerárquico: los datos se agrupan en forma de árbol (por ejemplo, el reino animal)



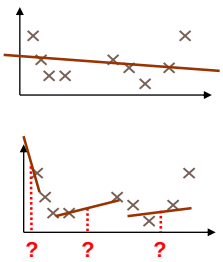
    - No jerárquico: los datos se agrupan en un mismo nivel jerárquico.



41

Tareas, técnicas y herramientas

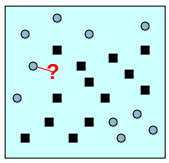
- Técnicas predictivas:
  - Regresión linear
  - Regresión logística
  - Regresión no linear (local)



42

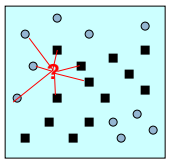
Tareas, técnicas y herramientas

- Técnicas predictivas:
  - Vecinos más próximos



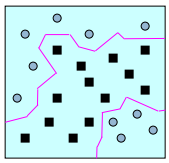
Círculo

1-vecino más próximo



Cuadrado

7-vecino más próximo



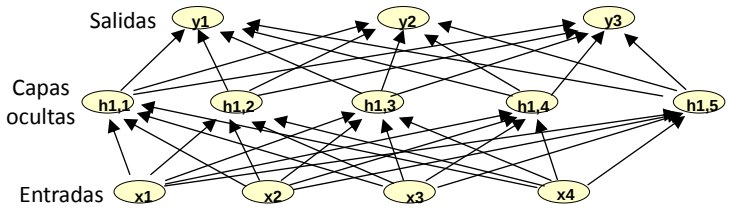
1-partición vecino más próximo

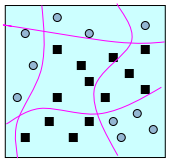
(Poliédrico o Voronoi)

43

Tareas, técnicas y herramientas

- Técnicas predictivas:
  - Redes neuronales



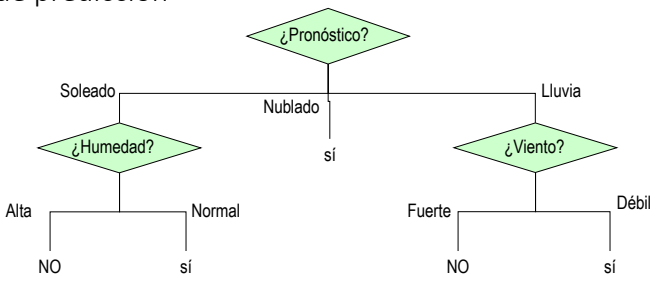


MÚLTIPLES PARTICIONES NO LINEARES POSIBLES CON CUATRO UNIDADES INTERNAS

44

Tareas, técnicas y herramientas

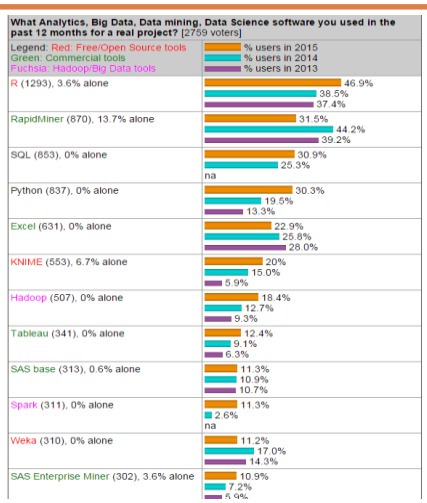
- Técnicas predictivas:
  - Árboles de predicción



45

Tareas, técnicas y herramientas

- Herramientas:
  - Muchas son libres y de código abierto.



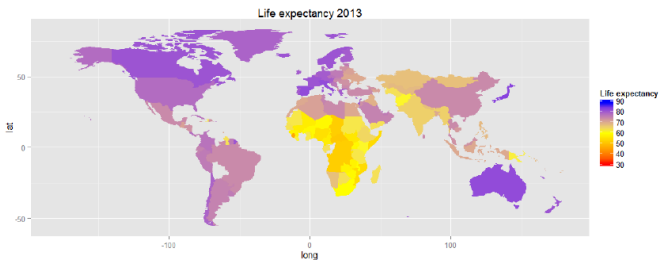
46

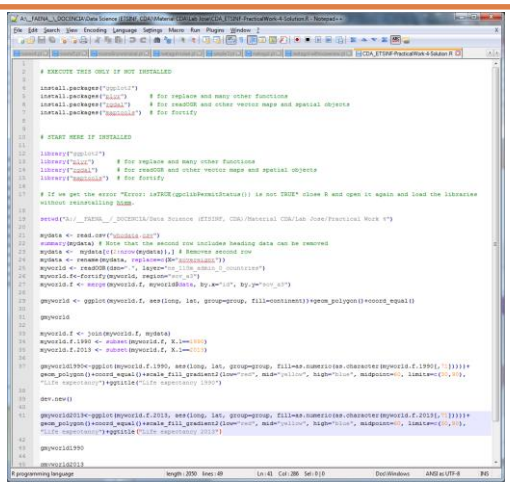
Tareas, técnicas y herramientas

- Lenguajes
  - Lenguajes más usados
    - R
    - Python (con librerías)

Tareas, técnicas y herramientas

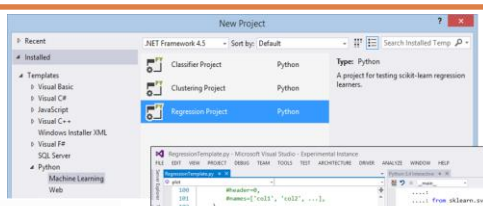
- Lenguajes
  - R

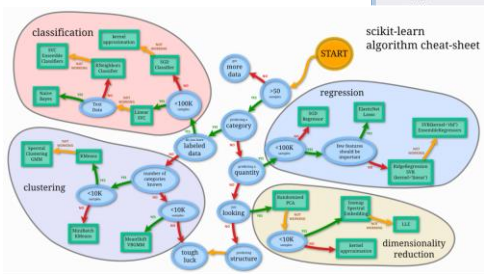


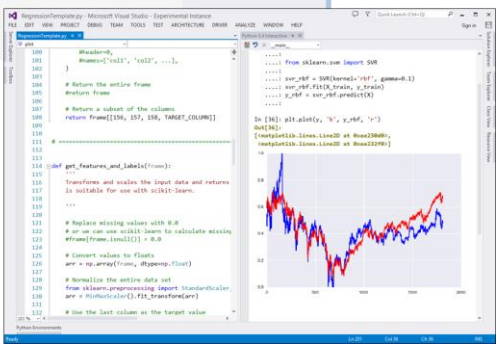


Tareas, técnicas y herramientas

- Lenguajes
  - Python + scikit-learn





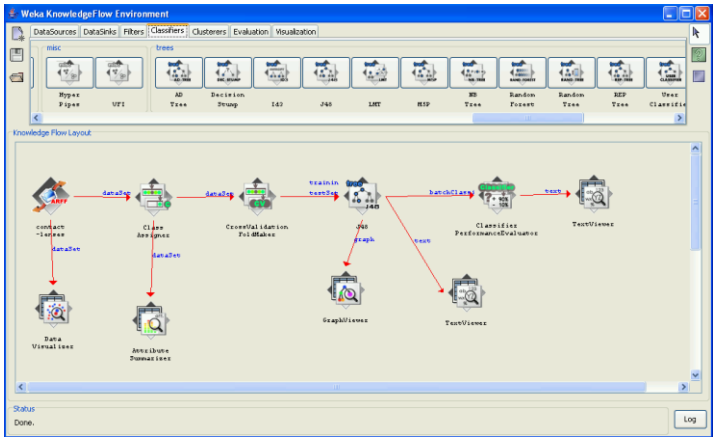


Tareas, técnicas y herramientas

- Suites
  - *Open source* (código abierto):
    - Weka
    - Rapid Miner
  - Comerciales:
    - IBM Modeler
    - SAS Enterprise Miner
    - Oracle BI Data Miner

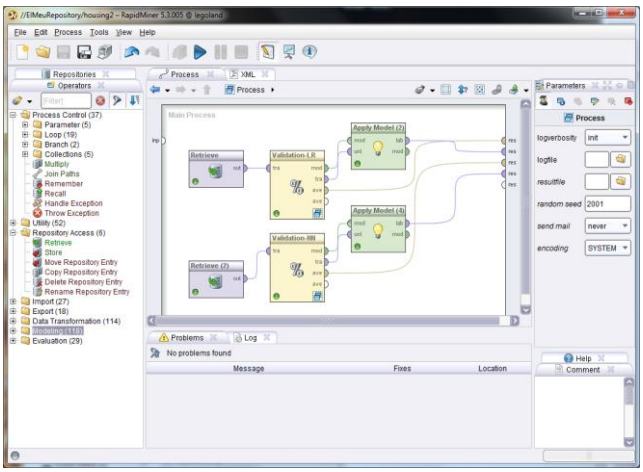
Tareas, técnicas y herramientas

- Suites
  - Weka



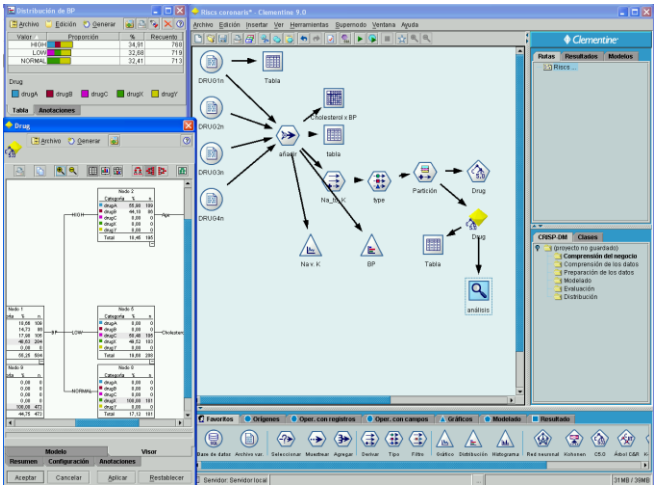
Tareas, técnicas y herramientas

- Suites
  - RapidMiner



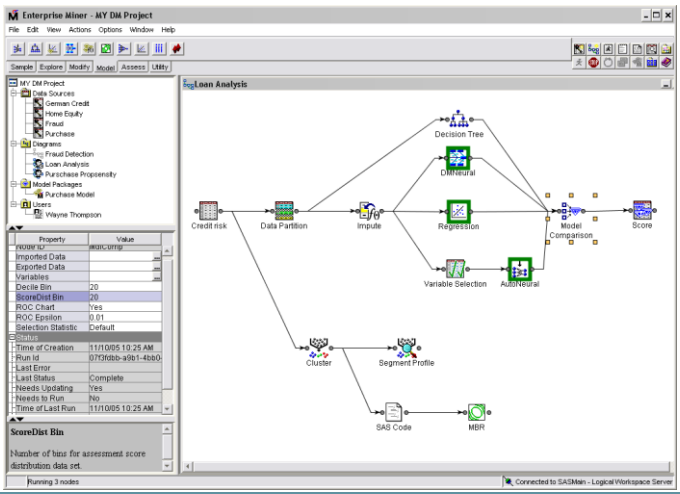
Tareas, técnicas y herramientas

- Suites
  - IBM Modeler



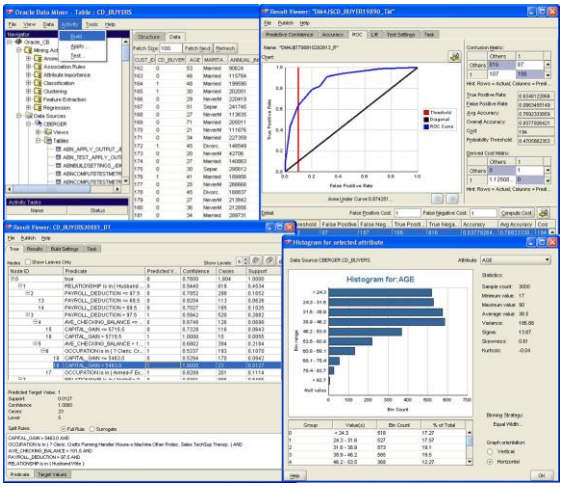
Tareas, técnicas y herramientas

- Suites
  - SAS Enterprise Modeler



Tareas, técnicas y herramientas

- Suites
  - Oracle BI Data Miner



55

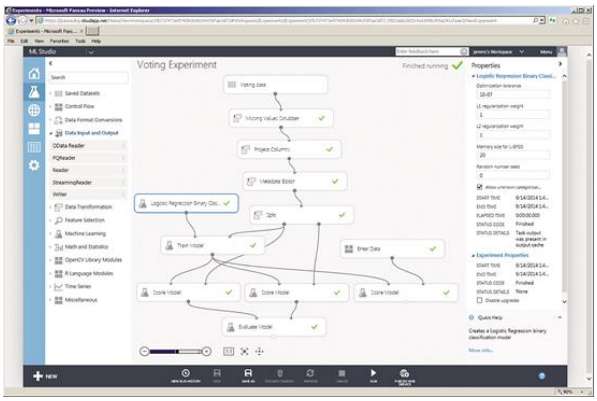
Tareas, técnicas y herramientas

- Nube:
  - AzureML (Microsoft)
  - BigML
  - …

56

Tareas, técnicas y herramientas

- AzureML (Microsoft)



Tareas, técnicas y herramientas

- BigML