Data Mining

- Un conjunto de principios, conceptos y técnicas que estructuran el pensamiento y análisis de datos

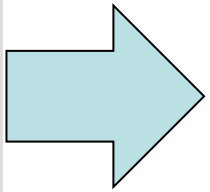
- - Extrae información y conocimientos útiles de grandes volúmenes de datos siguiendo un proceso con pasos razonablemente bien definidos

- - Cambia la forma de pensar sobre los datos y su papel en los negocios.

Nuevas Oportunidades en la Toma de Decisiones a partir de datos (*Data Driven Decisions* – DDD)

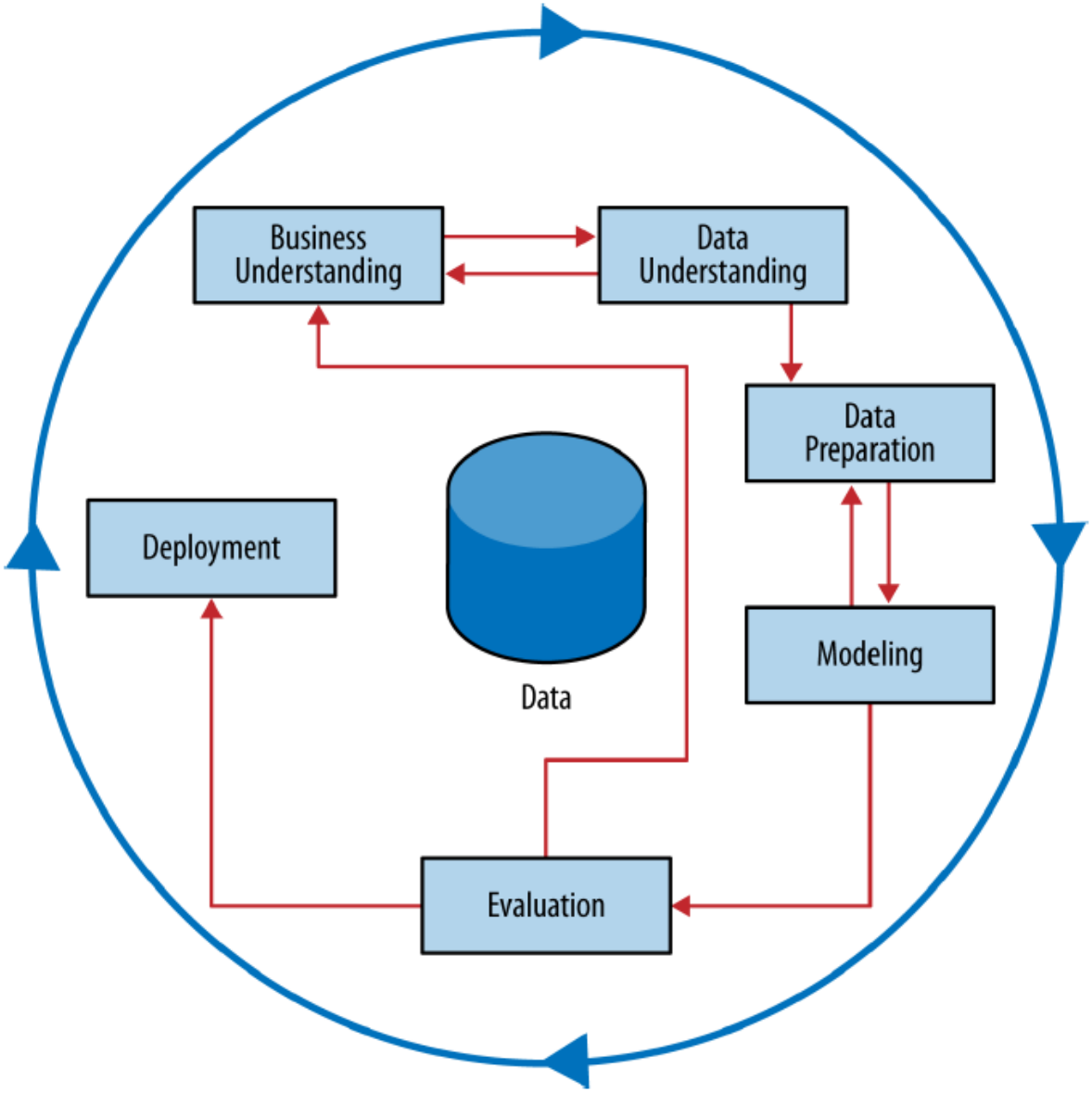
- Volumen de datos
- Variedad de datos
- Mejor Hardware
- Mejores Software y Algoritmos

La ciencia de Datos aplicada a los negocios es un proceso:



científico + artesanal + creativo + de sentido común

Esquema



- Entendimiento de negocios
- Análisis de los datos
- Preparación de datos
- Modelado
- Evaluación
- Despliegue

Data Mining vs:

- **Almacenamiento de Data warehousing/ Almacenamiento-**
  - Los Data warehouses reúnen datos de toda una empresa, a menudo de múltiples sistemas transaccionales

- **Consultas / Informes (SQL, Excel, MySQL y otras consultas basadas en GUI)**
  - Interfaz muy flexible para hacer preguntas objetivas sobre los datos-
  - Sin modelado ni búsqueda de patrones sofisticados

- **OLAP - Procesamiento analítico en línea-**
  - proporciona una interfaz gráfica de usuario fácil de usar para explorar grandes conjuntos de datos.
  - La exploración es manual; no hay modelado
  - Dimensiones de análisis pre-programadas en el sistema

- **Análisis estadístico tradicional**
  - Se basa principalmente en la comprobación de hipótesis o en la estimación / cuantificación de la incertidumbre

- **Modelado estadístico automatizado (por ejemplo, regresión múltiple)**
  - Esto es minería de datos de un tipo - usualmente basado en modelos lineales
  - Las bases de datos masivas permiten alternativas no lineales

Respondiendo preguntas de negocios con Modelos de DM

• ¿Quiénes son los clientes más rentables?

Database querying

• ¿Existe realmente una diferencia entre los clientes rentables y la cliente medio?

Ensayos de Hipótesis estadísticas

• ¿Pero quiénes son realmente estos clientes? ¿Puedo caracterizarlos?

OLAP (búsqueda manual)**, Minería de datos** (búsqueda de patrones automatizada)

• ¿Será rentable algún nuevo cliente en particular? ¿Cuánto ingresos debo esperar que genere este cliente?

Minería de datos (modelado predictivo)

Ejemplo: Predicción de baja de clientes

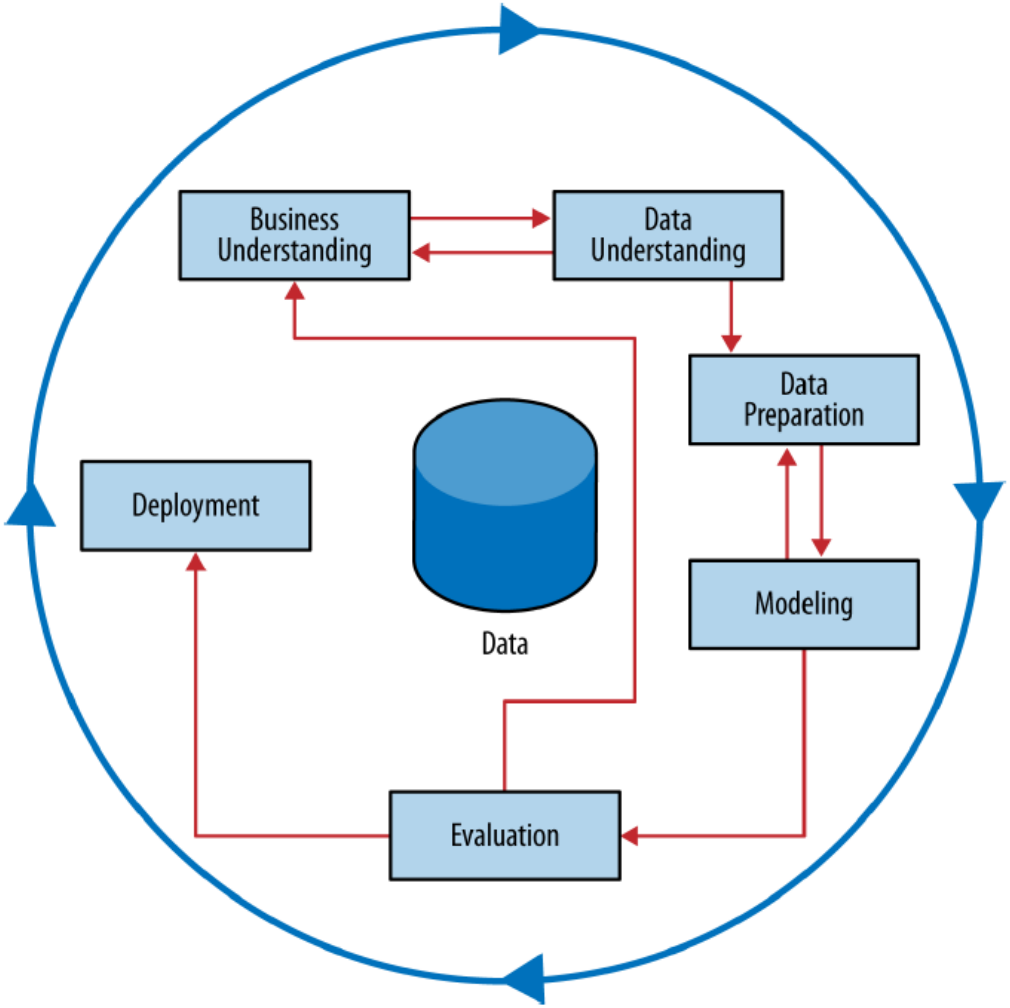
- Acabas de conseguir un gran trabajo en el área Ciencia de Datos en MegaTelCo, una de las las mayores empresas de telecomunicaciones del mercado.

- Están teniendo un gran problema con la retención de clientes.

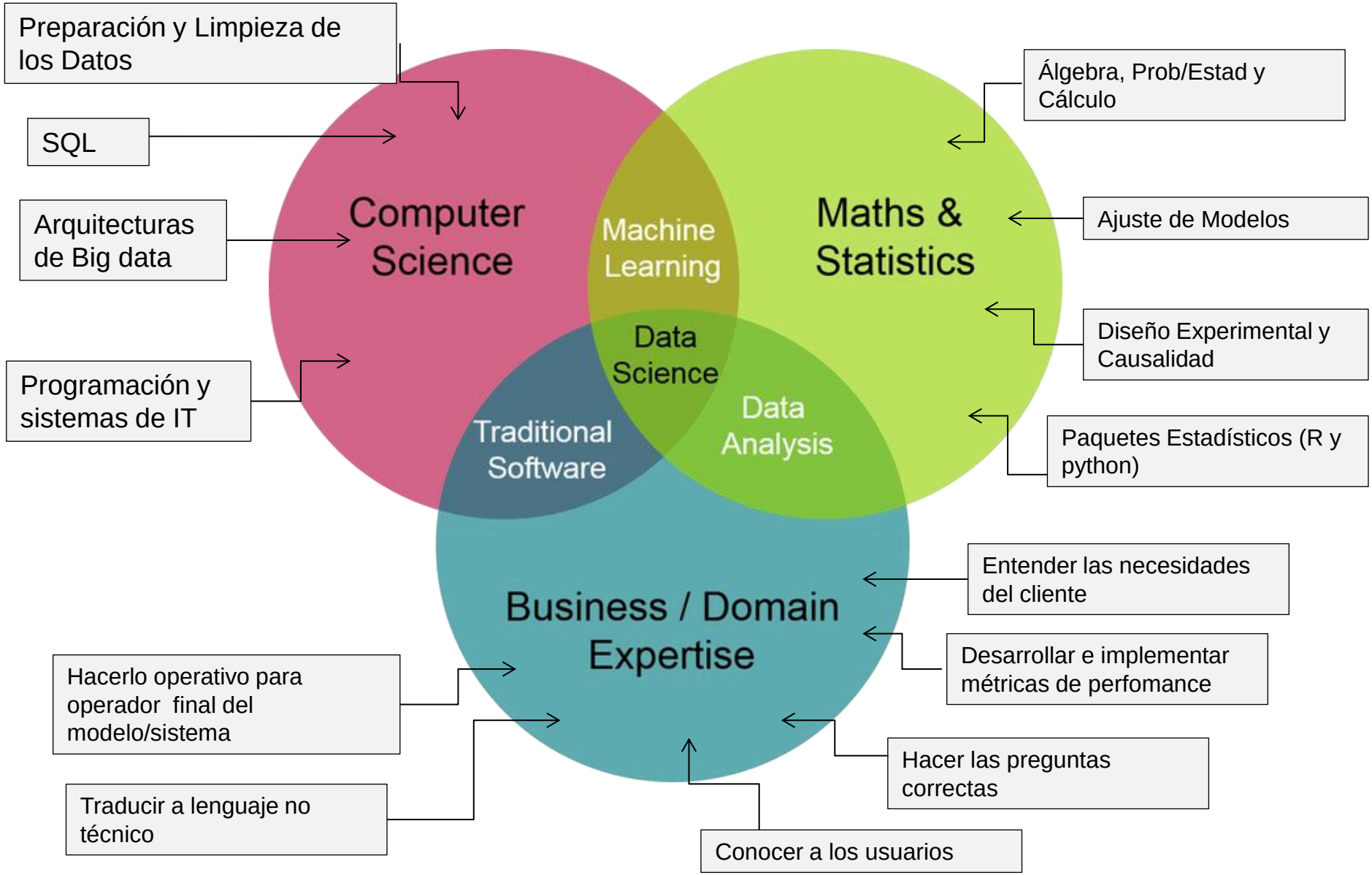
- El 20% de los clientes de teléfonos celulares se dan de baja cuando sus contratos expiran. Las empresas de comunicaciones son muy competitivas y agresivas a la hora para atraer a los clientes que cambian de compañía y conservar los suyos propios.

- Marketing ya ha diseñado una oferta especial de retención

- La tarea asignada es diseñar un plan preciso, paso a paso, de cómo el equipo de ciencia de datos debe utilizar los vastos recursos de datos de MegaTelCo para resolver el problema.



- ¿Qué datos puede utilizar?
- ¿Cómo se utilizarían?
- ¿Cómo debería MegaTelCo elegir un conjunto de clientes para recibir su oferta con el fin de reducir de manera óptima la pérdida de clientes para un determinado presupuesto de incentivos?


Preparación y Limpieza de los Datos
SQL
Arquitecturas de Big data
Programación y sistemas de IT
Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise
Álgebra, Prob/Estad y Cálculo
Ajuste de Modelos
Diseño Experimental y Causalidad
Paquetes Estadísticos (R y python)
Entender las necesidades del cliente
Desarrollar e implementar métricas de perfomance
Hacer las preguntas correctas
Conocer a los usuarios
Traducir a lenguaje no técnico
Hacerlo operativo para operador final del modelo/sistema

Terminología

- Modelo: Una representación simplificada de la realidad creada para servir a un propósito

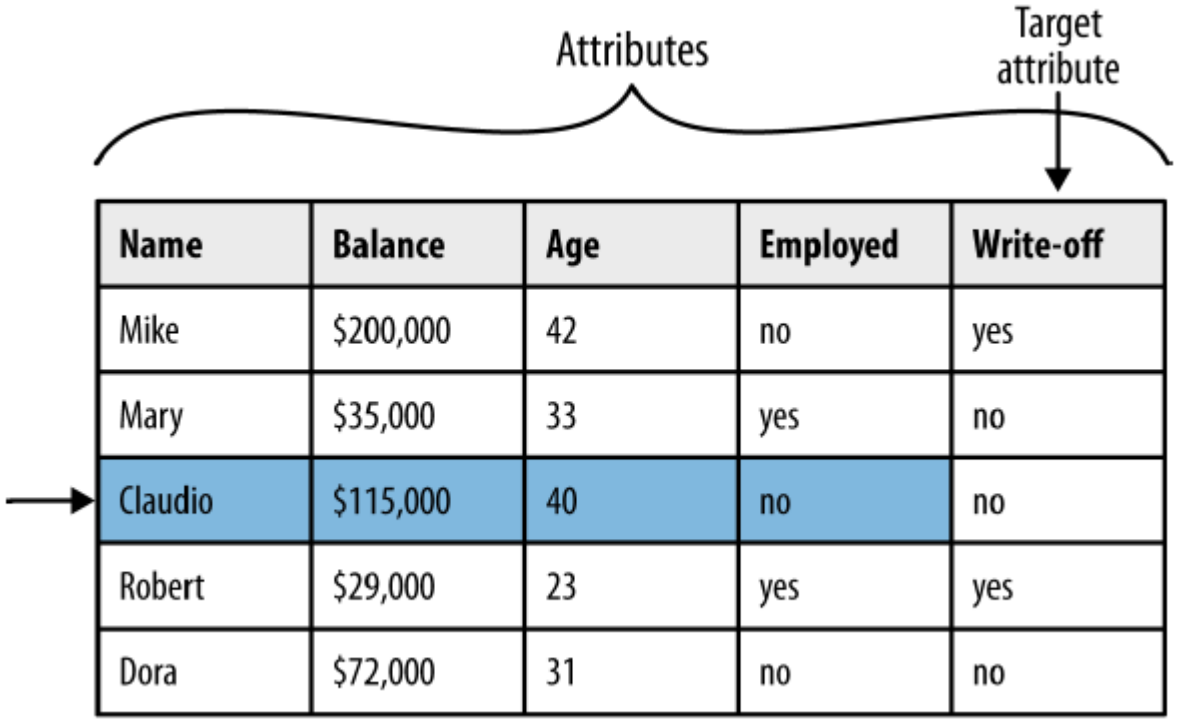
- Modelo Predictivo: Una fórmula para estimar el valor desconocido de interés: el objetivo La fórmula puede ser matemática, lógica (por ejemplo, regla), etc.

- Predicción: Estimar un valor desconocido (es decir, el objetivo)

- Instancia / ejemplo: Representa un hecho o un punto de dato, Descripto por un conjunto de atributos (campos, columnas, variables o características)

- Modelos o algoritmos inductivos: La creación de modelos a partir de datos

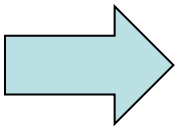
- Datos de entrenamiento:- Los datos de entrada para el algoritmo inductivo



Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

¿Dimensionalidad de los datos?

La dimensionalidad de un conjunto de datos (dataset) es la suma de las dimensiones de las características



La suma del número de características numéricas y el número de valores de las características categóricas.

Name	Balance	Age	Default
Mike	$123,000	30	Yes
Mary	$51,100	40	Yes
Bill	$68,000	55	No
Jim	$74,000	46	No
Mark	$23,000	47	Yes
Anne	$100,000	49	No

Tareas comunes de minería de datos

- **Clasificación y estimación de probabilidad de clase**
  ¿Cuán probable es que este consumidor responda a nuestra campaña?

- **Regresión**
  ¿Cuánto usará el servicio?

- **Búsqueda de similitudes**
  ¿Podemos encontrar consumidores similares a mis mejores clientes?

- **Agrupación**
  ¿Mis clientes forman grupos?

- **Agrupación de co-ocurrencias**
  - ¿Qué artículos se compran comúnmente juntos?

- **Perfilado (descripción del comportamiento)**
  - ¿Cómo es el "comportamiento normal"? (por ejemplo, como línea de base para detectar el fraude)

- **Reducción de dimensiones**
  ¿Qué variables latentes describen las preferencias del consumidor?

- **Predicción de enlaces**
  Ya que Juan y Juana comparten dos amigos, ¿debería Juan convertirse en amigo de Juana?

- **Modelado Causal**
  ¿Por qué se van mis clientes?

Minería de Datos Supervisada y Modelos Predictivos

- Existe un objetivo específico y cuantificable que nos interese i que estemos tratando de predecir?

- ¿Tenemos datos sobre este objetivo? ¿Tenemos suficientes datos sobre este objetivo?

- El resultado de la minería de datos supervisada es un modelo que predice alguna cantidad o permite entender un contexto o sutuación

Subclases de Minería de Datos Supervisada

- **Clasificación**
  Objetivo categórico (a menudo binario)

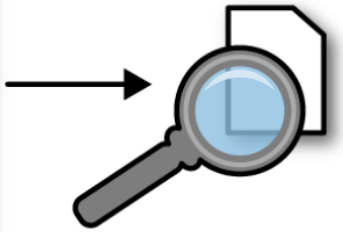
- **Regresión**
  Objetivo numérico

Minería de Datos vs Implementación y Operación del Modelo

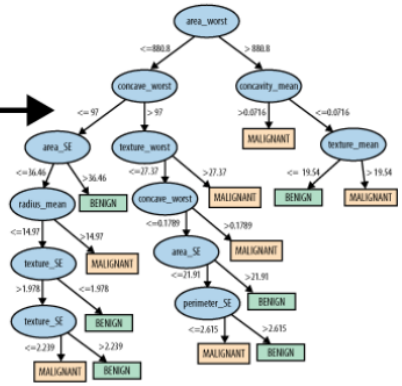
Datos Históricos

x	y	z	class
14	True	Red	accepted
6	True	Blue	rejected
...			
50.3	False	Red	accepted

Minería de Datos



Entrenamiento del Modelo



Datos de entrenamiento con todos las variables

Modelo entrenado y probado

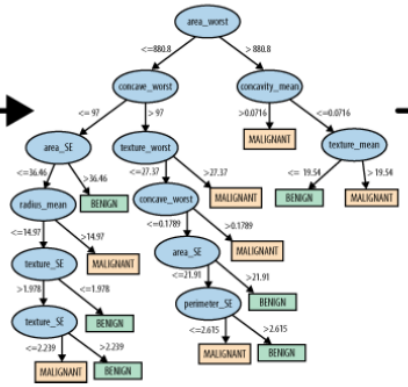
Entrenamiento

Uso

Nueva instancia

x	y	z	class
30	false	Red	?

Modelo



Clase: acepta con 80% de probablidad

Nuevo dato: la clase es desconocida (p.e. ¿el cliente aceptará la propuesta?

Obstáculos clásicos en la configuración e implementación de un modelo de DM

- Los datos del entrenamiento NO son consistentes con el uso
- Muestra insuficiente
- Malas características