Tema 1

Introducción

# Glosario

**DBA (*Database Administrator*)**: Referencia tanto la función de administrar la base de datos como a la persona encargada de hacerlo. Instala, configura, tunea y monitoriza la BBDD, backups, recuperación…

Diferencias con DA (*Data Arquitect*): modelado conceptual y lógico (DA) versus modelado físico (DBA)

**DA**: control central sobre los datos y políticas de acceso y administración. El DBA crea la base de datos e implementa los controles técnicos necesarios para hacer cumplir las políticas del DA (C. J. Date)

El DA es el encargado de comprender los procesos del negocio, la arquitectura lógica y física del DW, comprender ETL, y diseñar, identificar las fuentes de datos…

Los **ingenieros de datos** (DE) suele tener responsabilidades parecidas a las de los DBA en organizaciones pequeñas. Integran datos de fuentes diversas, trabajan en entornos Big Data, implementan los procesos ETL, instalan soluciones DW, hacen modelado de datos, testean la arquitectura de la BBDD…

**Arquitecto BI**: enfocado en el diseño de las aplicaciones, generación de reportes, los cubos OLAP, la minería de datos, análisis descriptivo, KPIs, dashboards…

**Científico de Datos**: análisis de datos usando técnicas estadísticas. Análisis predictivo y prescriptivo, ML

# Relacionales vs. No relacionales

* Relacionales
  + ACID: Atomicidad, Consistencia, Aislamiento, Durabilidad
* No relacionales (NoSQL)
  + Sin esquema
  + Escalado horizontal
  + BASE: Basically Available, Soft state, Eventual consistency
    - Prima la disponibilidad frente a la exactitud

## Teorema CAP

En sistemas distribuidos no se pueden garantizar todas las propiedades a la vez:

* Consistencia (SQL)
* Disponibilidad (SQL, NoSQL)
* Tolerancia a fallos (NoSQL)

# Almacenes de Datos

Los almacenes de datos tratarán de proveer escala infinita, dejando la eficiencia en segundo plano.

* Las bases de datos se empiezan a usar en el ámbito de toma de decisiones (DSS), usando consultas complejas
* Las consultas analíticas (OLAP) colisionaban con operaciones de BBDD, por lo que se separó en consultas operacionales (OLTP). Nace los DW.
* Los DSS necesitan datos históricos que no están en bases de datos operacionales. Los DW los contienen, y están focalizados para consultas y análisis.
* La calidad de los datos en la fuente es un problema frecuente, ya que es el apoyo en la toma de decisiones de la empresa.
* En los DW habrá no datos crudos, sino sumarios o agregados sobre datos crudos de las BBDD.

## Generación de DW

* Interaccionar con los usuarios para documentar requisitos.
* Data wrangling (captura, prueba (staging), archivado)
* Minería con estadística o ML

Es necesario integrar los datos de fuentes heterogéneas de forma constante, y permitiendo identificarlos en la línea temporal. No es volátil: no se borran ni modifican los datos una vez añadidos.

## Uso de DW

Los sistemas operacionales se ejecutan en OLTP (pequeños, buen rendimiento, operaciones específicas, cambian), mientras que los DW se implementan para apoyar consultas y aplicaciones de toma de decisiones (grandes, crecen, consultas lentas).

* Captura de datos en fuente: requiere colaboración con DBAs en la fuente y recursos de procesado en producción
* Captura de datos después de extracción: mayor volumen que será necesario con una copia completa anterior para detectar los cambios.

Cualquier opción implica convertir formatos y dirigir datos a ficheros o tablas.

## Componentes

Fuentes de datos de producción, extracción y conversión, sistema gestor de la BBDD del DW, administración y herramientas de Business Intelligence.

Los esquemas están simplificados y orientados hacia el negocio para generar reportes fácilmente. Se suele usar un diseño estrella o copo de nieve. Al normalizar se pueden generar un alto número de tablas no redundantes.

Modelo dimensional: tablas hechos y tablas dimensiones.

Tema 2

AHP y OLAP

# Selección de BBDD, Servidor OLAP

Se debe ajustar a las especificaciones, considerando parámetros como tamaño, precio, rendimiento, herramientas y add-ons, escalabilidad, conectividad, etc., y comparándolo por lo general.

Se pueden usar métodos analíticos analizando los requerimientos, determinando su peso, calificando candidatos y seleccionando.

## Proceso Jerárquico Analítico (AHP, *Analyticial Hierarchical Process*)

Se basa en comparaciones a pares entre alternativas a cada uno de los criterios, determinando la importancia de cada criterio y obteniendo los pesos de cada uno.

Por ejemplo, para comparar tres alternativas de BBDD (PS, AE y CS), se usan los siguientes criterios:

* Control de Concurrencia
* Precio
* Optimización de Consultas
* GUI de Administración

Y se usa la siguiente escala:

* 1 Iguales
* 3 Moderadamente Preferida
* 5 Preferida
* 7 Muy Preferida
* 9 Extremadamente Preferida

Y se genera entonces la matriz comparando los criterios entre ellos a pares:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CC | Precio | OC | GUI |
| CC | 1 | 0.2 | 0.33 | 0.14 |
| Precio | 5 | 1 | 3 | 0.33 |
| OC | 3 | 0.33 | 1 | 0.2 |
| GUI | 7 | 3 | 5 | 1 |
| Suma Columna | 16 | 4.53 | 9.33 | 1.67 |

Se normaliza dividiendo cada valor en la columna por la suma de esta y se pondera por fila:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | CC | Precio | OC | GUI |  |
| CC | 0.0625 | 0.0441 | 0.0353 | 0.0838 |  |
| Precio | 0.3125 | 0.2207 | 0.3215 | 0.1976 |  |
| OC | 0.1875 | 0.0728 | 0.1071 | 0.1197 |  |
| GUI | 0.4375 | 0.6622 | 0.5359 | 0.5988 |  |

Finalmente se generaría la matriz por marca con los valores multiplicados por los criterios y se tendría la mejor opción.

# Benchmark

TPC es un estándar de rendimiento o precio/rendimiento y también teniendo en cuenta medidas energéticas. Hay varios tipos:

* TPC-C, TPC-E (OLTP): transacciones por minuto o por segundo
* TPC-H: consultas compuestas por hora
* TPC-DS, TPC-VMS, TPC-HS…

El precio incluye el sistema completo, incluyendo mantenimiento, software, hardware, etc.

# Diseño de DW

No requiere control de concurrencia, procesado de transiciones, recuperación… Su diseño busca que las consultas sean simples, rápidas, agregación de datos, uniones, etc.

Por lo general hay 3 capas: servidor de la BBDD al que alimentar con ETL, capa intermedia con el servidor OLAP y otra capa formada por las herramientas.

Hay diferentes arquitecturas:

* Virtual: vistas de la BBDD que operan en conjunto de BBDD separadas pero pueden ser consultadas a la vez. Datos muy grandes para ser movidos.
* Data mart: subconjunto de datos de la organización sobre un área funcional. Mejora consultas analíticas reduciendo el conjunto de datos a escanear.
* Empresarial: abarca toda la organización.

Hay dos medotologías:

* Inmon: arriba abajo. Crear un DW centralizado con toda la información normalizada. El DW es la única fuente de verdad. De ahí se crean data martas, que pueden estar desnormalizados. ETL es fácil porque hay poca redundancia, muchos joins y tablas.
* Kimball: abajo arriba. Se cran los data marts para combinar luego en el DW grande. Es no normalizado y fácil de entender. El proceso de creación del DW es más rápido, pero puede haber datos redundantes.

## Ciclo de vida de Kimpball

Planificación programa:

* Evalúa la preparación de la organización para una iniciativa DW/BI
* Justificación y alcance preliminar
* Localiza recursos
* Lanza programa/Proyecto

Definir los requisitos del negocio

* Tarea esencial para el éxito de la iniciativa
* Influye en las actividades siguientes de diseño e implementación

Diseño técnico de la arquitectura/selección instalación productos:

* Establece el marco de trabajo para integrar multiples tecnologías
* Evaluamos y seleccionamos los productos necesarios

Modelado dimensional/Diseño físico/ETL:

* Centrado en los datos
* Trasladamos los requitisos al modelo dimensional
* Migramos el modelo a una estructura física y lo tuneamos para optimizar el rendimiento
* Diseñamos e implementamos ETL

Diseño y Desarrollo de aplicaciones BI:

* Debe satisfacer la mayoría de las necesidades analíticas de los usuarios
* Despliegue, Mantenimiento y Crecimiento

## Modelos de Datos Multidimensional

Modelo Estrella: los datos se organizan en hechos y dimensiones. Cada dimensión tiene una tabla asociada a un conjunto de atributos.

Fases:

* Identificar los procesos a cubrir
* Identificar los datos a almacenar
* Identificar las dimensiones y atributos asociados
* Identificar los hechos
* Crear el esquema

El esquema está desnormalizado, y se normaliza en las tablas de dimensiones. Medidas en hechos: aditivas (pueden ser sumadas en cualquier dimensión asociada a los hechos), semi aditivas (solo a lo largo de alguna dimensión) o no aditivas (ninguna dimensión).

Puede haber valores de medidas nulos en las tablas hechos, pero se deberán evitar en claves foráneas.

Modelo Estrella: tipo de ejemplo de datos de venta de una compañía. Redundancia en país provincia, consultas simples, generar reportes fáciles, cubos OLAP.

Modelo copo de nieve: normalizar las tablas dimensiones con respecto al modelo de estrella. Menos espacio ocupado, pero menor rendimiento.

Modelo constelación de hecho: múltiples tablas de hechos compartiendo tablas de dimensiones.

# Servidor OLAP

Permite realizar operaciones en el DW, expandiendo la utilidad analítica. Estructuras multidimensionales para consultar datos de forma organizada. Los cubos permiten representar los datos en varias dimensiones.

Los cubos OLAP agregan datos durante los periodos de baja actividad de las BBDD. Se pueden crear en ficheros, CSV, etc. Conjuntos de datos diferentes requieren cubos diferentes. Se usa MDX en vez de SQL.

Las herramientas son:

* MOLAP: consultas rápidas. Herramientas especializadas. Se precomputa.
* ROLAP: BDDD relacionales, tiempo real pero consultas lentas.
* HOLAP: híbrido de las anteriores.

## Operaciones OLAP

Operadores multidimensionales:

* Roll-up: agregación, se sube la jerarquía. Reduce el número de dimensiones.
* Drill-down: se baja una jerarquía, se pueden generar más dimensiones.
* Slice: seleccionar una dimensión, pasa de cubo a 2D.
* Dice: proyectar, como slice pero sin reducir el número de dimensiones (subcubo)
* Pivot: rotar, representar diferentes planos del cubo, reordena dimensiones.

# Minería de Datos

OLAP es una colección de métodos para consultar conjuntos de datos multidimensionales.

Minería de datos emplea técnicas estadísticas, de IA y gestión de datos para extraer patrones de grandes cantidades de datos. Consultas más elaboradas que OLAP. Puede proporcionar información sobre “el futuro”.

OLAP: porcentaje de alumnos que han estudiado en colegios privaos que no han superado el primer curso del año pasado.

DM: usar información de los ingresos de los padres, lugar en residencia, aficiones para predecir resultados académicos.

Tema 3

ETL

# Extraer, Transformar y Cargar

Se extraen los datos de los sistemas operacionales. Se transforman limpiándolos y se cargan en el DW.

Conexión entre los sistemas operacionales “fuente” y el DW. Suele representar un porcentaje muy grande. Hay un gran número de herramientas, pero se puede programar.

## Requisitos

* De negocio
* De conformidad
* De calidad de datos
* De seguridad
* De interfaces BI
* De integración de datos
* De latencia
* De archivado y linaje
* De habilidades disponibles
* Licencias de sistemas legados

# Planificación e implementación

* Diseño lógico
* Comprender arquitectura DW/BI
* Mapear fuente

## Planificación

1. Esbozar plan de alto nivel: identificar fuentes y destinos
2. Seleccionar herramientas ETL
3. Desarrollar las estrategias para las actividades más comunes
4. Mapear el flujo de datos
5. Cargar las tablas de dimensiones con datos históricos
6. Cargar las tablas de hechos con datos históricos
7. Procesado incremental de dimensiones
8. Procesado incremental de hechos
9. Agregar carga de tablas y OLAP
10. Operaciones del sistema ETL y automatización

# Data Lakes

Los data lakes están asociados a big data y capturan los datos en crudo sin pasar por ETL. Se almacena tal cual y se transforman según las necesidades en las aplicaciones (schema-on-read).

Si no se monitoriza el contenido de los DL, sin catalorgar ni organizar ni control de acceso, el DL no aporta valor a la empresa. Hay que evitar que se convierta en un Data Swamp (pantano).

Tema 4

Rendimiento

# Rendimiento

* Carga de trabajo: combinación de transacciones online, predecible vs impredecible
* Caudal de datos: capacidad del sistema para procesar datos
* Recursos: kernel, discos, memoria
* Optimización: en las consultas y parámetros de la BD
* Contención: más contención, menor caudal. Varios componenentes intentan acceder al mismo recurso.

El objetivo es optimizar el uso de recursos para minimizar la contención y maximimzar el caudal de datos. Eliminar cuellos de botella. Regla de Pareto: 80% de los efectos provienen del 20% de las causas. 70% de carga de trabajo: desaconsejado.

## Niveles de Ajuste

* Hardware: más discos, mejores, más memoria, mejor procesador, SAN NAS, RAID, escalado horizontal o vertical…
* SO: filesystem y kernel
* Parámetros y diseño de la BD: tamaño de buffer apropiado, ajuste automático, localización de logs en otro disco
* Middleware: drivers, conexiones, caching…
* Aplicaciones

## Ajustes del Software

Mal diseño, SQL malo, malas prácticas. Vistas materializadas, configuración de la BBDD, modelos mal estructurados: transformarlos desnormalizando (minimza los joins), dividir tablas, aplicar clustering…

Índices: para consultas importantes, estudiar si es un plan mejor es posible con nuevos índices. Seleccionar atributos con alta selectividad (pocas filas con mismo valor). No indexar columnas que se modifican frecuentemente. No usarlos en tablas pequeñas, con muchos valores NULL, columnas que se manipulan mucho.

Clustering de index: orden físico de la tabla corresponde a la secuencia del índice. Implica modificar la tabla: se genera el índice, se reordena la tabla, se copia sobre la original. Indica consultas de rango. Existen los índices compuestos, y el orden importa.

Optimización de SQL: evitar HAVING, grandes operaciones de ordenación, seleccionar columnas mínimas. Utilizar transacciones combinando muchas consultas. Usar procedimientos almacenados.

## Test de carga

Hay que asegurar que se puede gestionar con un rendimiento razonable. Interaccionar con el sistema y evaluar su respuesta en función de la carga de trabajo.

# Nube vs Local

* Dinámico/horizontal (nube) vs estático/vertical (local)
* Variedad de soluciones y economía de escala (nube) vs gestión completa (local)
* Proporcional a los servicios contratados (nube) vs infraestructura (local)
* Control por parte del usuario: Moderado (IaaS), Bajo (DaaS) vs Alto (local)
* Costes: pago por almacenaje, cpu, trafico (nube) vs gran inversión inicial (local)
* Integración rápida (nube) vs lento (local)

# Tiempo real

Hay necesidades de disponer de los datos en tiempo real. Cumplir requisitos puede condicionar el presupuesto. Cuello de botella en ETL. Diferentes opciones:

Microbatching: “batch processing” automático a intervalos cortos o conjuntos pequeños de datos. No hay tiempo real, se procesan con retraso (pocas horas) y se acumulan los tiempos de retraso de cada etapa. Casos en los que los requisitos de latencia no son muy estructos.

Alimentación contínua (direct trickle feed): propagar los cambios en BBDD operacionales a DW. Problemas de rendimiento si las tablas están siendo consultadas por proceso analíticos. Usar tablas de hechos en una partición separada. En alto volumen de datos el problema puede existir.

Trickle Flip (alimentar y dar la vuelta): pasar los datos a tablas intermedias (staging). Consultas usan las tablas de hechos en DW. Periódicamente se copian las tablas intermedias y se renombran para sustituir las tablas de hechos. Si las tablas son grandes puede llevar mucho tiempo.

Caché externa para datos en tiempo real (RTDC): datos en tiempo real se almacenan fuera del DW, en servidor dedicado para almacenar y procesar datos en tiempo real. Las consultas que requieren datos en tiempo real acceden a esa base de datos y no a la histórica. Regularmente se vuelcan los datos al DW. Escala bien, pero procesar consultas es complejo.

# Consultas analíticas

Las consultas analíticas pueden llevar mucho tiempo y realizar varios pases sobre los datos. Tradicionalmente se asume que los datos no cambian y no se toman precauciones para prevenir inconsistencias.

RTDC previene que pase esto último, ya que mantiene los datos en tiempo real e históricos separados.

Tema 5

Valores Atípicos

# Big Data

Grandes conjuntos de datos caraterizados por las 5 V’s que demandan formas innovadoras y económicas de procesado de información que permiten mejorar el análisis, toma de decisiones y automatización de procesos.

## 5 V’s

* Volumen
* Velocidad
* Variedad
* Veracidad
* Valor

# Business Intelligence

Conjunto de métodos, procesos, arquitecturas, herramientas y tecnologías que permiten convertir los datos en información que soporte de forma estratégica la toma informada de decisiones.

Reportes descriptivos, OLAP, Analítica de negocio, Minería de datos…

# Minería de Datos

Descubrir de forma analítica patrones en grandes conjuntos de datos previamente desconocidos. Interacciona con IA, ML, estadística, HPC y BBDD. Detección de anomalías, reglas de asociación, agrupamiento, clasificación, regresión.

Se puede beneficiar de un DW, reduciendo la fase de preprocesamiento.

## Proceso Genérico

* Enunciado del problema y formulación de una o varias hipótesis.
* Recoger los datos
* Preprocesar los datos (eliminar valores extremos, nulos, etc.), escalar valores
* Generar el modelo
* Interpretar el modelo

## CRISP-DM

Cross Industry Standard Process for Data Mining. Es un modelo ideal de naturaleza cíclica.

Business understanding -> Data understanding -> Data preparation -> Modeling -> Evaluation -> BU & Deployment

# Valores Atípicos

Datos estadísticamente no consistentes con los demás. Pueden dar información valiosa, se trata de eliminar ruido. Múltiples aplicaciones: fraude tarjetas, IDS, eventos extremos, crimen, terrorismo.

## Detección

En datos espaciales las dimensiones espaciotemporales se deberán tratar por separado. Los valores extremos son un tipo de valor atípico, pero no todos los valores atípicos son valores extremos.

Los valores extremos se usan para alimentar modelos de fenómenos extremos.

Métodos supervisados y no-supervisados: se tiene información a priori sobre los valores extremos del conjunto de datos.

Métodos univariados y multivariados: cuantas variables a considerar en el conjunto de datos.

Métodos estadísticos y no paramétricos: se usa boxplots para representarlos.

# Estrategias

* Eliminarlos: grandes cantidades de datos sin valores atípicos que representan todas las clases. Origen dudoso. No hay tiempo para analizarlos.
* Eliminar la variable que presente más valores atípicos.
* Tratarlos por separado respecto al resto de datos. Se quieren saber las causas.
* Imputación o estimación de valores “missing”: asignar un valor que puede ser la media u otro valor arbitrario.

Tema 6

Agrupamiento

Metas fundamentales: predecir y describir. Agrupamiento: método no supervisado.

# Agrupamiento

Buscar grupos de puntos que están relacionados entre si y que son lo suficientemente diferentes de puntos asignados a otros grupos. Similitud inter-grupo baja e intra-grupo alta.

Los puntos pueden pertenecer a varios grupos. No todos los puntos tienen por que pertener a algún grupo. Los grupos pueden tener tamaños parecidos. Detección de diferencias.

Aspectos para tener en cuenta sobre el método de agrupamiento:

* Escala bien sobre grandes conjuntos
* Datos categóricos, numéricos o binarios
* Tipo de grupos a cubrir, esféricos
* Grupos vacíos
* Capacidad de interpretar y usabilidad

Se debe realizar una selección de características antes del agrupamiento. Se deberá luchar contra la maldición de la dimensionalidad: obligaría a tener gran cantidad de datos a medida que se aumentan las dimensiones del conjunto. Eliminar redundancia y atributos irrelevantes.

# Selección de características

* Métodos de filtrado: tests estadísticos sobre las propiedades del conjunto.
* Métodos envolventes: en función de los resultados del algoritmo, usando una función objetivo para evaluar los resultados

# Evaluación de resultados

La validación del agrupamiento suele ser difícil. Los resultados pueden ser bonitos y nada útiles, y viceversa. Se pueden evaluar de dos formas:

* Usando los datos del agrupamiento: indicadores de evaluación internos
* Usando datos externos ya etiquetados

Usando la correlación, se parte de 2 matrices: una fila y punto por cada punto.

* Matriz de proximidad (con las distancias)
* Matriz de incidencia con 1 o 0 según pertenezcan al mismo clúster

Se calculan las correlaciones entre las 2 matrices. Si la correlación es alta, los puntos están en el mismo clúster.

# Basados en…

## Particiones

Asignar cada punto a uno de los k-agrupamientos a través de un proceso iterativo. Cada grupo contiene al menos un punto. Se minimiza una medida de no similaridad entre los puntos dentro del agrupamiento.

## Localización

Grupos de objetos vecinos basados en condiciones locales.

## Grid

Se divide el espacio en hiper celdas eliminando las casi vacías. Se combinan las celdas adyacentes para formar agrupamientos.

## Jerárquico

No se determinan el número de agrupamientos. Basta una matriz de distancias para calcularlo. Grupos anidados organizados en una estructura de árbol denominada dendograma.

* Divisivos: se comienza con un agrupamiento y se va diviendo. Se divide hasta que todos los objetos están separados o se alcanzan los K grupos, y se añaden objetos que son similares al grupo.
* Aglomerativos: cada punto es un clúster y se va mezclando en pasos sucesivos.

## Densidad

Se busca un conjunto máximo de puntos conectados por densidad. Los grupos son regiones densas, separados por regiones de baja densidad.

# K-Means

Se comienza con un conjunto de puntos y se debe determinar el número de grupos que se quieren. Se define al inicio. Se localizan centroides de forma aleatoria. Cada grupo está representado por su centro. Se trata de minimizar la distancia media de cada miembro del grupo al centro. Se repite.

Fácil de implementar y rápido. No funciona bien con datos categóricos. Escalar resultados puede ser la diferencia entre buenos resultados y malos.

# Mean shift

Se inicializa el punto inicial y una ventana de tamaño S. Se calcula la media de S. Se desplaza la ventana hacia la media. Se vuelve a calcular el centro de gravedad hasta que converge.

No asume una forma de clúster, pero depende de S (se selecciona de manera arbitraria).

# DBSCAN

Trabaja con el concepto de un punto que debe ser alcanzable por densidad. Tres tipos de puntos:

* Core: interior del clúster. Debe haber puntos suficientes en su vecindario.
* Borde: no tienen puntos suficientes, pero es vecino de un punto core.
* Ruido: el resto.

# MAFIA

Basado en Grid. El algoritmo busca regiones de alta densidad particionando el espacio en hiper celdas. Los clústeres se crean uniendo esas celdas de alta densidad adyacentes.

# GMM

Ajustar los datos a un modelo matemático que son las superposiciones de varias distribuciones. Estimación de los parámetros a través de un algoritmo de optimización EM. El número de clústeres se obtiene usando una aproximación bayesiana. Los clústeres pueden tener cualquier forma elíptica. K-Means es un caso particular de GMM.

Se selecciona el número de clústeres y se inicializan los parámetros de la distribución gaussiana. Calcular la probabilidad de que cada punto pertenezca a cada clúster. Recalcular los parámetros para maximizar las probabilidades de tener puntos dentro de cada clúster. Se repite hasta converger.

# SOM

Self Organizing Maps, basada en redes neuronales. Compiten para representar mejor los datos. El número de nodos en la red viene dado por los usuarios, junto con el nivel de detallo. Los pesos se inicializan aleatoriamente. Cada entrada se compara cada nodo de la red con el de menor diferencia, y se ajustan los pesos de nuevo.

Se inicializan los pesos de cada nodo. Se selecciona una entrada de los datos de entrada. Se evalúa cada nodo para ver cual tiene los pesos más parecidos. Se calcula el vecindario del nodo más parecido. Se actualizan los pesos.

Existe un factor de aprendizaje que determina lo rápido que converge la red. Reduce la dimensionalidad de los datos a un mapa. Se aplica en muchas áreas.

Tema 7

Clasificación y Reglas de Asociación

# Clasificación

Técnica de aprendizaje supervisada: se incluye información sobre las clases en los datos de entrada. Se parte de un conjunto de entrenamiento sobre el que construir un modelo para el atributo clase en función de los demás atributos.

La clasificación puede ser binaria o multiclase. Se llama regresión cuando se predice una variable continua. Se crean dos sets: uno de entrenamiento y otro de test.

Técnicas: árboles de decisión, SVM, ANN, Random Forest… Se debe considerar exactitud, rapidez, interpretabilidad y tiempo de aprendizaje.

# Calidad

* Positivos verdaderos: modelo predice correctamente
* Negativos verdaderos: muestras reconocidas correctamente que no pertenecen a la clase
* Falsos positivos: muestras incorrectamente asignadas a la clase
* Falsos negativos: muestras que no fueron correctamente asignados a la clase

Matriz de confusión, KxK siendo K el número de clases. Precisión es la frecuencia en la que se clasifican las clases positivas (P=PV/(PV+FP)).

Cuando se genera un modelo, se aproxima una función verdadera y por eso se tiene sesgo (variación sobre la realidad) y varianza (variación entre predicciones). Si domina mucho el sesgo, no se ha ajustado el modelo lo suficiente. SI domina mucho la varianza, se está en un sobreajuste. Se desea bajo sesgo y baja varianza.

# Árboles de decisión

Nodos (evaluación de atributos), ramas (salidas del nodo) y hojas (etiquetas de clase). La mejor raíz es la que más reduce la incertidumbre.

Múltiples condiciones “si … entonces …”. Se pueden entender al analizar el camino de raíz a hojas. El proceso es inductivo, se construyen de forma fácil. Trabajan tanto con números como con categorías.

Pueden ser inestables, pequeños cambios en la entrada pueden aportar cambios importantes en la estructura del árbol. Puede producirse sobre entrenamiento, para evitarlo:

* Deteniendo el crecimiento (ganancia de información es poca)
* Realizando la poda (eliminar ramas y evaluar usando un conjunto distinto)

# Método Naive Bayes (Bayesiano Simple)

Clasificador probabilístico para clasificación binaria y multiclase. Se calculan las probabilidades de cada atributo, y se usa la regla del producto para obtener la probabilidad condicional para los atributos.

Parte de que cada predictor es independiente y eso no pasa siempre. Los datos de entrenamiento deben ser representativos de cada clase. Parte de una distribución normal, lo cual puede ser inexacto.

# SVM

Máquinas de Soporte Vectorial. Máquinas de Vectores de Soporte. Puede usarse con regresión. Se quiere encontrar el hiperplano óptimo (superficie de decisión). Encontrar la solución óptima.

Se maximiza el margen en torno al hiperplano. El margen viene definido por los vectores de soporte. Protege contra errores a la hora de elegir el mejor separador. Problema simple. En una dimensión es un punto.

Es una clasificación one-vs-one, si hay X clases hacen falta X(X-1)/2 clasificadores. Es posible que haya ruido que interfiere en el proceso de separar las clases.

Maximizar el margen y clasificar los puntos de forma correcta. Hace falta un kernel, un factor C, que penaliza por clasificar de forma incorrecta.

Es fácil de usar, conjuntos de datos pequeños. Escala mal, datos ruidosos dan problemas.

Función kernel: no lineal, realiza de forma eficiente el mapeo de puntos. Una dimensión adicional sobre las originales.

# K-NN

Clasificador sencillo. La elección de la métrica juega un papel fundamental. En el caso del vecino más próximo, se computa la clase de salida en base al vecino más próximo. En el caso de los k-vecinos más próximos, se asigna la clase más frecuente de sus k-vecinos más próximos.

Altos valores hacen que el algoritmo sea menos sensible al ruido. Valores pequeños de k permiten capturar estructura a pequeña escala. La dimensionalidad puede provocar una degradación al crecer el número de atributos.

# Random Forest

Se combinan resultados de varios modelos predictivos. Puede hacerse a partir de diferentes algoritmos de clasificación.

Bootstrap: método fundamental en ML. Re muestrear los datos de manera que éstos puedan repetirse en la muestra.

Bagging: usa muchas versiones de un conjunto de entrenamiento haciendo Bootstrap. Las salidas se combinan promediando (regresión) o por mayoría (clasificación).

Ensamblado de múltiples árboles de decisión. Al dividir un nodo, RF busca la mejor división considerando solo un subconjunto aleatorio de atributos. Dos componentes aleatorios: bootstrapping y los atributos al dividir un nodo.

## Algoritmo

1. Se toma un conjunto aleatorio de tamaño N usando bootstrapping
2. Se toma una muestra aleatoria de los predictores (sin reemplazo)
3. Se divide usando predictores del paso anterior
4. Se repiten los pasos 2 y 3 para cada división hasta completar el árbol
5. Se evalúa el árbol con observaciones no usadas. Almacenar la clase asociada a cada observación así como los valores de sus precictores.
6. Repetir pasos 1-5 muchas veces
7. Parac ada observación, se cuenta el número de árboles clasificadas en cada categoría
8. Se asigna la observación a la categoría mayoritaria

Los árboles son más independientes debido al Bootstrap y la variabilidad de los precictores. Son buenos clasificadores. Pueden trabajar con muchas variables y valores nulos. Pueden usarse en agrupamiento, regresión y detección de valores atípicos.

Son una especie de caja negra.

# ANN

Intentar una estructura cerebral de forma artificial. Perceptrón: tiene unas entradas y unos pesos para cada entrada. Produce una clasificación lineal del espacio de entradas. Inicialmente aleatorio.

Las redes neuronales artificiales están formadas por capas de neuronas. Capa inicial es entrada, y la final es salida. Las demás capas están escondidas, la capa de entrada se limita a alimentar a la siguiente capa con los atributos necesarios para la clasificación.

El número de nodos de entrada coincide con la dimensionalidad de los datos. La salida de una neurona puede ser la entrada de otra neurona en la misma capa o incluso capas previas (recurrentes).

Si las operaciones en las neuronas fueran solo lineales, se pierde la capacidad de generalización de las ANN. Hay que entrenar las neuronas, minimizando la función de coste. Validaciones con muestras independientes.

Una ANN se representa como una caja negra con 2 métodos: aprender y predecir. Al aprender se modifica el estado interno de la red mientras que al predecir se usa ese estado para generar salidas a partir de las entradas.

Para ir hacia delante se aplica la función en cada neurona y para ir hacia atrás su derivada. Es impracticable chequear el efecto que la variación en cada peso tiene en el error. Se debe evitar los mínimos locales.

Los cambios sobre los pesos deben ser pequeños para evitar comportamientos caóticos. Si oscilan, se debe reducir el factor aprendizaje. Difíciles de interpretar y computar. Buena opción cuando no se puede formalizar el problema.

## CNN

Redes neuronales convolucionales, transformar imágenes en vectores. Detección de patrones, función de activación. Para el análisis de series temporales también se usan redes neuronales recurrentes.

# Reglas de Asociación

Encontrar relaciones en un conjunto de datos. No predecir un único atributo como en la clasificación, sino la ocurrencia de un elemento en función e la ocurrencia de otros (leche y cacao da galletas).

Cantidad de ocurrencias: frecuencia de la ocurrencia de un itemset. Soporte es la fracción del conjunto que contiene un item set determinado. Confianza es la medida de la frecuencia que Y aparece en las transacciones que contienen X.

Descubrirlas: generar las reglas con soporte y confianza mayores que unos umbrales. Fuerza bruta no escala bien. Reducir el número de candidatos, reducir el conjunto de entrada y reducir el número de comparaciones a realizar.

Si un itemset es frecuente, sus subconjuntos también son frecuentes. Calcular itemsets k+1, podar itemsets candidatos y eliminar candidatos infrecuentes, repetir hasta identificar itemsets frecuentes.