Tema 1) Introducción

- Define qué es un sistema de información y describe brevemente cuáles son los principales tipos de sistemas de información en el ámbito empresarial. Por último, para cada uno de los tipos de sistemas indica su principal propósito y algún ejemplo de software comercial que se emplee en esos sistemas.

Un sistema de información son un conjunto de elementos interrelacionados que recogen datos (entrada), manipulan información (proceso), almacenan los datos, y los diseminan (salida); y además, proporcionan mecanismos correctores (feedback); todo ello con el propósito de alcanzar un determinado objetivo. Los principales tipos de SI en el ámbito empresarial son los siguientes:

1. TPS: Sistemas de procesado transaccional, que facilitan la operativa diaria de la empresa (ventas, compras, nuevos productos fabricados, etc).
2. MIS: Sistemas de información de gestión, que permiten definir procedimientos rutinarios como la planificación de rutas logísticas, gestión de nóminas, generación de facturas, etc.
3. DSS: Sistemas de apoyo a la toma de decisiones, que ayudan a tomar decisiones para un problema específico complejo.
4. EIS: DSS para ejecutivos, orientados a los altos directivos, que ayudan a conseguir los objetivos estratégicos de la empresa, aunque su uso principal es meramente informativo.

TPS -> OLTP & Oracle Retail
MIS -> OLTP & SAP (también tiene TPS)
DSS -> OLAP & Tableau
EIS: OLAP & Microsoft Power BI

- Explica brevemente en qué consiste un modelo de sistema de información de 3 capas con interfaz web. Escribe un ejemplo de este tipo de sistema.

Un sistema de información de 3 capas con interfaz web es un sistema de información de 3 capas, que además tiene interfaz web. Es decir, el sistema se divide en Interfaz (UI, vista...), Modelo (lógica de negocio) y Datos. La capa de datos gestiona las transacciones y la persistencia, el modelo toda la lógica de negocio, y la interfaz es la vista final para el usuario.

A diferencia de un sistema de 3 capas normal, uno con interfaz web sitúa la UI y el Modelo en un servidor de aplicaciones web, y la Base de Datos en otro servidor (un servidor de BD). El cliente accede a la aplicación desde un navegador web de propósito general. Esto tiene la gran ventaja de que cambiar el modelo, o cambiar la interfaz, no supone la re-compilación en los clientes, y únicamente supone la re-compilación en el servidor de aplicaciones web. Lo mismo para el modelo de datos.

Un ejemplo sería la web de reservas de un hotel. El usuario accede a él mediante Brave Browser o cualquier otro navegador, obteniendo las páginas mediante peticiones HTTP desde el servidor de aplicaciones, y realiza reservas escribiendo parámetros y mandándolos, mediante HTTP, al servidor de aplicaciones, que procesará la información y guardará lo que corresponde en la BD, y probablemente generará nuevas páginas de feedback.

Estos modelos surgieron con la Web 1.5 (web de transición, web dinámica) y son los más habituales hoy en día para cualquier página web.

- ¿Qué tipos de mecanismos de retroalimentación de sistemas de información existen? Enuméralos e indica un ejemplo de cada uno de ellos para el sistema de información "Anillo Digital Docente" de la Universidad de Zaragoza.

- Eficiencia: Mide la calidad del resultado obtenido (salida) con el número de recursos invertidos para conseguirla.
- Efectividad: Mide si el objetivo se ha cumplido o no, y en qué medida se ha cumplido.
- Medidas de rendimiento estándar propias del sistema.

Eficiencia -> Tiempo que se tarda en descargar una presentación de una asignatura
Efectividad -> ¿Se consigue descargar una presentación de una asignatura?
Rendimiento estándar -> ¿Se cumple el RGPD con la información almacenada de los estudiantes y profesores?

- Eres responsable TIC en una empresa, y tu jefe te pide que estudies las alternativas para implantar un sistema de gestión para la documentación interna de la empresa ¿Qué criterios o requisitos deberías tener en cuenta para poder decidir qué sistema implantar/comprar/desarrollar para esa empresa? (¡ojo!, no preguntamos qué sistema implantarías, sino qué información necesitas para poder tomar una decisión adecuada)

1. ¿Qué necesidad concreta se va a resolver? Necesito entrevistarme con el jefe y obtener una lista de **TODOS** los requisitos funcionales del sistema, para saber qué hace concretamente.

2. ¿Quién es el usuario final? ¿Cuántos usuarios va a tener? ¿Quién es el cliente?

3. ¿Qué información se va a gestionar, cómo y de dónde se va a obtener?

4. ¿Qué disponibilidad necesita el sistema? (24/7, laboral...)

5. ¿Qué tipo de persistencia de datos necesita?

6. ¿Cuánto presupuesto tenemos? ¿Cuántos somos en el equipo de desarrollo? ¿Para cuándo tiene que estar terminado el sistema?

7. ¿Existe ya alguna alternativa en el mercado que haga esto?

8. ¿Tiene ya la empresa un sistema legado que hace esto?

9. ¿Qué alternativas en el mercado hay que se puedan adaptar, con funcionalidades similares?

10. ¿Existen algunas condiciones específicas que me fuercen a utilizar una tecnología u otra?

Tema 2) Evolución de la web

- Describe brevemente los cinco principios de la Web de datos enlazados e indica qué herramientas y tecnologías se han empleado para implementar y desarrollar esos principios. ¿Qué relación guardan estos principios con los tres pilares considerados en la Web Semántica?

Los principios de la Web de Datos enlazados fueron propuestos por Tim Berners Lee. Son los siguientes:

1. Definir una URI para cada recurso/identidad a representar. Una de las tecnologías que permite definir URIs para las entidades es RDFa. Este principio se relaciona con los pilares de anotación, que nos indica que todos los recursos deben estar anotados; y ontología, que define que se debe seguir un vocabulario estándar para definir vocabularios (en este caso, RDFa). OWL
2. Utilizar el protocolo estándar HTTP para mostrar información de la URI. Esto facilita la recuperación de información por parte de motores de búsqueda y razonadores, lo cual nos lleva a enlazarlo con el pilar de "uso de razonadores".
3. Definir relaciones entre los recursos, que permite establecer relaciones entre varios recursos, por ejemplo, de sinonimia, "es una instancia de", etc. Relacionado con las anotaciones (las relaciones se deben anotar), ontologías (tenemos que seguir un estándar para definir el vocabulario) y razonadores lógicos (ayuda a los agentes software a conocer las relaciones entre recursos). RDFa y las tripletas RDF nos permiten esto.
4. Uso de tripletas RDF para definir los datos y SPARQL para recuperarlos. Relacionado con la anotación y ontología, pues RDF y SPARQL siguen el formato estándar.
5. Ofrecer información útil, la información debe ser relevante y tener algún tipo de relación con el recurso buscado. Tecnologías son RDF y SPARQL, y se relaciona con el pilar de razonadores y motores de inferencia

- Indica brevemente qué es la Web Semántica y la Web de Datos Enlazados incidiendo en los principios y pilares en los que se basa. ¿Consideras que estas tecnologías y modelos pueden contribuir a la transparencia de las administraciones públicas? ¿Cuáles crees que son las barreras para la adopción de estos modelos en la administración?

web de lectura, escritura y ejecución

La Web semántica es una web orientada tanto a máquinas como a personas. Es una web en la cual los agentes Software pueden interactuar de forma automática y procesar, integrar y utilizar información de distintos sitios web para lograr sus objetivos. El término fue acuñado por Tim Berners Lee, Jim Hendler y Ora Lassila. Los pilares fundamentales de la web semántica son los siguientes:

1. Anotación: consiste en etiquetar todos los recursos con metadatos.
2. Ontologías: Consiste en utilizar vocabularios estandarizados para definir las anotaciones
3. Uso de reglas y razonadores lógicos (motores de inferencia) con el objetivo de deducir nuevos datos que ayuden a la toma de decisiones.

La web de datos enlazados es la aplicación de la web semántica, originada gracias a DBpedia. El término fue acuñado por Tim Berners - Lee, y se basa en tres reglas:

1. Definir una URI por cada entidad
2. Formato estándar para mostrar las URIs mediante HTTP.
3. Relacionar unos datos con otros, estableciendo relaciones entre URIS (recursos), por ejemplo, relaciones de sinonimia.

Sí, la web semántica/Linked Data web puede contribuir a la transparencia de las administraciones públicas, pues facilita el intercambio y comprensión de los datos de una forma estructurada y estandarizada, permitiendo que las relaciones entre los datos sean públicas.

Las barreras de su adopción seguramente estén relacionadas con la complejidad técnica que conllevaría cambiar los sistemas ya existentes por unos nuevos, y la pereza de los funcionarios públicos de anotar todos los datos siguiendo los estándares.

Tema 4) Tecnologías de la web dinámica:

- Define qué es una cookie y explica brevemente para qué se han empleado. Además, contesta a las siguientes preguntas: ¿Se deberían seguir empleando cookies en la actualidad? ¿Qué mecanismos alternativos a las cookies existen? Al realizar este ejercicio considera al menos los siguientes casos: 1) se desea construir un sistema de información Web desde cero, y 2) se trata de mejorar las funcionalidades proporcionadas por un sistema de información Web diseñado e implantado a finales de la década de los 90

Una Cookie es un par <"nombreAtributo", valorAtributo> donde ambos son Strings. Las Cookies se almacenan en el navegador web del cliente y su objetivo es mantener información para identificar al cliente en una sesión o entre varias sesiones. Una sesión se define como "tiempo que el usuario interactúa con el sistema hasta que lo abandona".

¿Se deberían seguir empleando cookies en la actualidad? Sí, las cookies son útiles para preservar información del usuario y brindarle una experiencia más agradable, permiten que el usuario guarde parámetros de personalización etc. No obstante, debe tenerse cuidado de cumplir las reglas de protección de datos, y de no ser invasivo con la información que recopilamos. Sobre todo, en los sistemas legados, donde la GDPR no existía y había otras leyes más leves que han sido endurecidas. Es posible que el legado no tenga soporte a cookies, por lo que deberían emplearse otros métodos (o encapsular el legado lo cual quizá no es permisible)

Mecanismos alternativos a las cookies:

1. Guardar atributos en la URL, solo sirve para una sesión y no permite guardar datos entre sesiones. Probablemente útil para el sistema legado, pero muy ofuscado para el nuevo.
2. Guardar atributos en los formularios, atributos ocultos. No es seguro, pues el usuario puede modificarlos a su gusto.
3. Local Storage. Consta en almacenar los datos en el navegador, pero no como una cookie. El navegador no lo manda al servidor con sus get/post, sino que lo interpreta él mismo con javascript. Es menos invasivo que las cookies.
4. Uso de tokens para identificar al cliente.
5. Uso de login y guardar la información respectiva al usuario en el servidor de BD de la webapp. Probablemente útil si el sistema legado no soporta cookies. El usuario se identifica, y se realizan peticiones a la BD con los atributos que había guardados sobre él. Conlleva un coste de la BD.

Tema 5) Tecnologías de la web semántica y de la web de datos

- Cuáles son las principales semejanzas y diferencias entre un modelo de datos basado en el modelo relacional y un modelo de datos basado en el modelo de tripletas RDF. ¿Qué lenguaje de interrogación usarías para una fuente de datos que emplee el modelo relacional? ¿Y para una que emplee RDF?

Ambos modelos de datos generan fuentes de datos estructuradas, que permiten definir entidades (o recursos, datos, instancias) y relaciones entre ellos. Además, permiten definir metadatos y ambos poseen un lenguaje de consulta.
Sin embargo, RDF utiliza tripletas <S,P,O>, compuestas por un Sujeto (recurso), un Predicado (propiedad del recurso) y un Objeto (valor de la propiedad para el recurso), mientras que en el modelo relacional se emplean relaciones (tablas SQL).
En el modelo relacional para insertar nuevos valores, debemos emplear la sentencia INSERT, mientras que en RDF basta con crear nuevas tripletas, aunque también existe la sentencia CONSTRUCT. RDF se puede expresar de forma gráfica mediante grafos donde S y O (recursos) son nodos y P son aristas que va de un nodo a otro, y expresan que el nodo de partida cumple la propiedad "P" con el valor del nodo de llegada.
Aunque ambos tienen un lenguaje declarativo para consultar datos, son distintos. RDF emplea SPARQL, mientras que el relacional comúnmente utiliza SQL. Las consultas SQL y las SPARQL son semejantes, pero tienen algunas distinciones.
En SPARQL tenemos dos tipos de consultas, las SELECT y las ASK. Las ASK devuelven true o false, dependiendo de si existe algún sujeto / objeto que cumpla las propiedades que le hemos indicado, mientras que la SELECT devolverá el valor solicitado.
En SPARQL no existe un "FROM", esto va implícito en las ontologías al especificar la URI que las define. Esta todo "junto" en la misma base de datos, no hay tablas.

Teoría mejor

Tema 6) Recuperación de la información

- Un departamento de la Universidad cuenta con una colección de documentos internos. Muchos de estos documentos contienen metadatos con los autores del documento, la fecha de publicación, y palabras clave que resumen su contenido. El número de autores distintos está alrededor de 500, escritos posiblemente de distintas formas, y se cuenta con documentos desde la creación del Depto. en el año 1995. Debes diseñar un sistema de Recuperación de Información (RI) que indexe los documentos y puede implementar diferentes tipos de búsqueda. En concreto, debes describir qué tipos de índices considerarías para tu sistema de RI y que tipos de búsqueda permitirías.

Consideraría tres índices invertidos distintos:
1. **Índice por metadatos "autor":** con el cual indexaría los documentos a partir de los autores que están etiquetados. Ej: Autor1 -> Doc1, Doc77, Doc88...
2. **Índice por metadatos "fecha":** con el cual se indexarían los documentos a partir de la fecha que tienen etiquetada.
3. **Índice por metadatos "palabras clave":** con el cual los documentos serían indexados en base a sus metadatos que son palabras clave.
4. **Índice general:** que es un índice tradicional de RI, sin aprovecharse de las anotaciones (ya que dice que MUCHOS contienen metadatos, no TODOS, y si no los tenemos en cuenta, algunos quedarían ocultos). Previa indexación en este índice, se deberán procesar los documentos (eliminación de *stop words*, lematización, *stemming*, etc.).

A partir de los índices desarrollados, se podrían considerar tres tipos de búsqueda distintos:
1. **Búsqueda por autor:** en la que el usuario puede buscar un autor, y se utilizaría el índice 1. Los documentos recuperados por el índice 1 tendrían un ranking alto. También se podría utilizar el índice 4, por si algún autor estuviese escrito de otras formas o no existiesen metadatos sobre él, pero el ranking sería menor.
2. **Búsqueda por fecha:** análoga a la 1 pero con la fecha, se utilizaría principalmente el índice 2 (ranking muy alto) y el 4 (ranking más bajo).
3. **Por palabra clave (búsqueda "normal"):** que se utilizaría el índice 3 pero también el 4. Es análoga a las anteriores.

- En el contexto de recuperación de información, en qué consisten los algoritmos de PageRank y HITS. Enumera las ventajas e inconvenientes de cada uno de ellos y describe en qué contextos es más adecuado usar uno u otro.

Ambos son algoritmos que permiten ordenar los documentos devueltos por un sistema de RI según su relevancia. PageRank está basado en HITS.
HITS: HITS es un algoritmo iterativo sobre los enlaces entre documentos. Requiere realizar una consulta inicial para que funcione. Cada documento tiene un peso de autoridad y un peso de hub. Una página con un peso de hub alto significa que apunta a muchas autoridades, y una página con alto peso de autoridad implica que es apuntada por muchos hubs. En cada iteración, el peso de autoridad se calcula como la suma del peso de hub de las páginas que lo apuntan en la iteración anterior, mientras que el de hub se calcula como la suma del peso de autoridad de todos los nodos a los que apunta.
PageRank: Únicamente considera un peso (el PageRank). No requiere una búsqueda inicial. Considera los enlaces como citas de otros documentos. Una página cuenta con un PageRank alto si la apuntan muchas páginas o la apuntan páginas con alto PageRank (hubs). PageRank asocia el texto de los enlaces con la página destino.

PageRank es más eficiente si tenemos un gran corpus documental y no depende de una búsqueda inicial como HITS. Sin embargo, la relevancia de los documentos recuperados por PageRank puede ser menor ya que estamos considerando un único peso. HITS nos aporta esta información extra: nos dice si una página es un hub o una autoridad, pero es más complejo computacionalmente y menos escalable.

PageRank se utiliza normalmente en motores de búsqueda o contextos en los que necesitamos documentos de forma rápida y no pasa nada si se cuela alguno menos relevante.

HITS viene bien cuando necesitamos diferenciar hubs de autoridades, y la eficiencia computacional no nos importa tanto. Por ejemplo, para buscar artículos científicos o libros.

- ¿En qué consiste un índice invertido? Nombra alguna aplicación donde es común que se empleen índices invertidos e indica brevemente para qué se emplean en ese caso y que ventajas y desventajas proporcionan en ese contexto.

Un índice invertido es una estructura de datos que, a cada uno de los términos de cada uno de los documentos presentes en el corpus documental, le asocia una lista de documentos en los que está presente.
Ej: caballo -> Documento1, Documento7 ; hueso -> Documento2, Documento7; Jesucristo -> Documento777, Documento12

Una aplicación donde es común es en los motores de búsqueda, en el módulo de indexación. Se emplean con el objetivo de, dada una consulta, obtener la lista de documentos relevantes a ella.

Las ventajas de un índice invertido son que encontrar el documento es casi inmediato, además de que se pueden eliminar stopwords antes de incluirlo al índice, o enriquecer la consulta para tratar los problemas de sinonimia, polisemia, etc. Cuesta muchísimo menos buscar en un índice que irse recorriendo de forma secuencial millones de documentos a ver si son relevantes o no.

La desventaja principal es el coste de mantener el índice: Debemos añadir y clasificar todo documento que se añada, y mantener el índice siempre actualizado y funcional. Además, un índice ocupa espacio extra.

Tiene sentido utilizar índices cuando el tamaño de documentos empieza a ser considerable (>200MB), o si sabemos a ciencia cierta que no se van a añadir más documentos (es un histórico).

Tema 7) Bases de datos distribuidas:

- Explica en qué se diferencian los modelos de bases de datos distribuidas y federadas, y en qué consisten los procesos de fragmentación y asignación de fragmentos.

Mejor desde el resumen

En las bases de datos federadas, ya existían múltiples esquemas individuales que deciden integrarse en uno solo. Surgen, por ejemplo, tras la fusión de varias empresas. Por ello, el diseño de estas es "bottom-up". Sin embargo, en las bases de datos propiamente distribuidas, se parte de "cero". El diseño es *top-down*, es decir, se realiza un esquema E/R y se fragmenta, asociando cada fragmento a cada uno de los posibles nodos. En las federadas, no existe fragmentación. Además, en las distribuidas debemos elegir nosotros qué elementos serán redundantes (mismo elemento en varios nodos), mientras que en las federadas la redundancia existía previamente a su creación, inherente a ellas.

Un fragmento es la unidad a distribuir entre distintos nodos: puede ser parte de una tabla, una tabla o un conjunto de tablas. Existe la fragmentación horizontal, en la que se separan filas a partir de condiciones de selección (SELECT * WHERE ciudad="Zaragoza"); y la vertical, basada en separar columnas a partir de conjuntos de atributos a proyectar (SELECT id, ciudad, país FROM tabla). En la horizontal, la intersección ha de ser vacía, mientras que, en la vertical, los fragmentos deben tener las claves primarias. También existe la mixta (primero horizontal y luego vertical).
Todo fragmento debe asociarse a un nodo como mínimo y la relación inicial debe ser reconstruible con operaciones tipo JOIN, UNION, etc.
Para asignar fragmentos a los esquemas locales, tenemos que decidir en base a la cercanía de los nodos (en el caso de las particiones horizontales) y en base a la utilidad de los nodos (en el caso de la vertical).
Además, habrá que tomar decisiones de replicación:
- Sin replicación: Un fragmento está en un solo nodo. Esto es muy positivo para actualizaciones, pero negativo para consultas. Es como tener menos cachés.
- Replicación total: Un fragmento está en todos los nodos, con lo que mejoramos las consultas al tener todos los datos, pero las actualizaciones son costosas. Es algo así como ACID → debe estar en todos.
- Replicación parcial: Algunos fragmentos están en todos los nodos, otros solo en aquellos en los que se van a utilizar. Es la opción más común. Por ejemplo, si tenemos datos de clientes de Francia y de España, en la sede española nos guardamos los clientes de España y no los de Francia, y viceversa; pero en ambas sedes nos guardamos las políticas de crédito, ya que estas probablemente serán las mismas.

- Explica brevemente qué es una base de datos federada. Asimismo, enumera las fases o pasos que se toman a la hora de construir una base de datos de este tipo. Además, explica brevemente alguna de sus fases.

Una **base de datos federada** es un conjunto de bases de datos autónomas operando sobre un esquema global. Este esquema global se obtiene "bottom-up", es decir, son bases de datos previamente existentes que se integran con el objetivo de ofrecer un esquema global. A diferencia de las bases de datos propiamente distribuidas, no hay que fragmentar y la redundancia (replicación) probablemente ya existe. La heterogeneidad es inherente a estas bases de datos. Distintos Sistemas operativos (SO), Sistemas de gestión de bases de datos (SGBD), Modelos de datos o Heterogeneidad semántica (sinonimia, polisemia, homonimia, hiperonimia, herencia, etc.)
Para construir una **BDF (Base de Datos Federada)**, debemos construir un esquema global a partir de **N esquemas globales**. Este proceso consta de dos etapas:
1. **Traducción:** Cada uno de los esquemas locales se convierte a un modelo canónico. Se identifican entidades, relaciones y atributos, siendo posible la aparición de nuevas entidades y el cambio de nombre de algunas.
2. **Integración:** Los esquemas locales en modelo canónico se unifican bajo un único esquema global. Es complejo porque hay que tener en cuenta la sinonimia, herencia, uniones (algunos datos pueden estar en dos esquemas locales con distinto nombre, por ejemplo), etc.

- Explica brevemente qué es una base de datos distribuida. Asimismo, enumera las fases o pasos que se toman a la hora de construir una base de datos de este tipo. Además, explica brevemente alguna de sus fases.

Una Base de Datos Distribuida (BDD de ahora en adelante) se define como un conjunto de bases de datos autónomas que funcionan bajo un mismo esquema global, de forma transparente al usuario. El usuario no conoce ni que hay distribución, ni que hay fragmentación, ni dónde están los fragmentos. El solo ve una única BD.
El diseño de una BDD se realiza de forma "top-down" (de arriba hacia abajo). Es decir, se crea un esquema entidad relación y relacional, y posteriormente se fragmenta, asignando los fragmentos a cada uno de los nodos. Por tanto, el proceso consta de tres fases:
1. Creación del esquema global
2. Fragmentación del esquema
3. Asignación de los fragmentos

La fase 1 es idéntica a la del diseño de una BD centralizada.

En la fase 2 hay que decidir qué fragmentar y cómo fragmentarlo. Un fragmento puede ser parte de una tabla, una tabla o incluso un conjunto de tablas. Existen dos tipos de fragmentación:
Horizontal: basada en encontrar condiciones de selección
Vertical: basada en encontrar subconjuntos de atributos a proyectar (en español: partimos la tabla verticalmente y nos quedamos solo con los atributos que queremos).
También puede ser híbrida (primero horizontal y luego vertical), que es lo más común. Todo fragmento debe estar asignado a, al menos, un nodo (la fragmentación ha de ser completa).

El esquema sin fragmentar debe ser reconstruible a partir de los fragmentos haciendo operaciones de JOIN, UNION, etc. La intersección entre los fragmentos ha de ser vacía, exceptuando las claves primarias en la fragmentación vertical.

En lo que respecta a la fase 3 (asignación), debemos decidir qué fragmentos añadir a qué nodos y si vamos a replicar o no. Si replicamos poco, es decir, los fragmentos están en un solo nodo, es positivo para el control de concurrencia (actualizaciones), pero negativo para consultas. Si replicamos mucho, es decir, los fragmentos están en muchos nodos, es positivo para consultas pero negativo para las actualizaciones, y también aumenta la disponibilidad de los datos. Normalmente se llega a un punto medio, lo cual empeora todavía más el control de la concurrencia, pero es lo que se suele hacer.

Por ejemplo: tenemos un fragmento con información de los clientes de Francia y otro con los de España. A su vez, existe otro fragmento con las condiciones para dar hipotecas a los clientes. Necesitamos a cada uno en su nodo, pero las condiciones no variarían: Francia estará en el nodo de París, España en el nodo de Madrid, pero ambos nodos tendrán las condiciones de la hipoteca.

Tema 8) Minería de datos:

- En el contexto de la minería de datos, define el concepto de regla de asociación. Asimismo, define los conceptos de soporte, confianza y lift de una regla de asociación y explica la relación existente entre esas medidas. Indica además en qué contextos se suele aplicar esta técnica (minería de datos descriptiva y/o predictiva) y pon algún ejemplo de aplicación que consideres de interés en el ámbito de gestión médica.

Una regla de asociación es una relación matemática X->Y que significa que "si sucede el evento X, entonces también sucederá el evento Y", donde X es el conjunto de elementos del antecedente e Y el conjunto de elementos del consecuente. Su objetivo es identificar relaciones interesantes y no triviales en grandes conjuntos de datos, normalmente correlación o causalidad. {Cerveza} -> {Pañales} indica que, si un cliente compra cerveza, también comprará pañales.

patrones frecuentes en los datos

1. El **soporte** es la probabilidad de que una transacción contenga los eventos X e Y. $Soporte(X \rightarrow Y) = \frac{N_{XUY}}{N}$; Nxuy es el número de instancias que contienen X e Y, N es el número total de instancias. Un soporte alto (cercano a 1), implica que hay muchas transacciones con los eventos X e Y juntos.

soporte/soporte.

2. La **confianza** se define como la probabilidad de que, si una transacción contiene X, también contenga Y. $Confianza(X \rightarrow Y) = \frac{N_{XUY}}{N_X}$; Nxuy es el número de instancias que contienen X e Y, Nx es el número de instancias que contienen X. Una confianza alta (cercana a 1), nos informa de que existe correlación entre X e Y, es decir, si sucede X, suele suceder Y.

LIFT > 1 c.p
LIFT = 1 indep
LIFT <1 c.n

ratio entre el soporte y el producto de las probabilidades de cada conjunto por separado

3. **Lift**: Relación entre la confianza y la probabilidad de que Y aparezca solo. Es decir, mide la correlación entre X e Y. Tiene en cuenta la probabilidad de que Y aparezca solo porque si Y sale muchas veces solo, Y no es relevante y no existe correlación. Si el lift es positivo, existe correlación. $Lift(X \rightarrow Y) = \frac{Confianza(X \rightarrow Y)}{N_Y/N}$

Las tres métricas sirven para medir la calidad de una regla de asociación descubierta. Lo más interesante es que las tres salgan bien. Por ejemplo, una regla con confianza y soporte moderados, pero con un lift positivo es mejor que una regla con soporte muy cercano a 1 pero confianza y lift muy bajos. Es decir, por si solas apenas proporcionan información relevante, deben estudiarse en conjunto para sacar conclusiones sobre los datos.

Minería de datos **descriptiva**: La técnica se utiliza para descubrir patrones y relaciones ocultas en los datos, sin un objetivo predictivo. Por ejemplo, análisis de patrones de compra.

Minería de datos **predictiva**: Aunque menos frecuente, también puede emplearse para predecir comportamientos futuros basándose en asociaciones previas.

Un ejemplo de aplicación médica podría servir para detectar síntomas que provocan una enfermedad y predecir enfermedades a partir de un conjunto de síntomas. También puede ser utilizado para describir síntomas que a menudo suelen ir juntos. Por ejemplo, supongamos que existe la siguiente relación {Tos, Fiebre} -> {Gripe} y que tiene un soporte de 0.2 y una confianza de 0.8. Esto significa que el 20% de los datos presentan {Tos, Fiebre} y {Gripe}, y que el 80% de los que tienen {Tos, Fiebre} también tienen {Gripe}

- Explica brevemente los principales tipos de análisis de minería de datos que se suelen emplear. Para cada uno de ellos, indica las técnicas que se suelen aplicar explicando brevemente en qué consisten.

La minería de datos puede ser con fin descriptivo (describir y encontrar patrones en un conjunto de datos existente), o predictivo (a partir de un conjunto de datos, realizar predicciones con nuevos datos similares). Existen varias técnicas para ellos, utilizables la mayoría en ambos.

predictivo

1. **Regresión**: Dado un conjunto de datos bidimensional (datos con dos propiedades, por ejemplo, área, el peso y precio), consiste en encontrar el polinomio que mejor se ajusta a estos. De este modo, si en el futuro tenemos un nuevo dato del cual desconocemos una de sus propiedades, podemos hacer una estimación de cuál será el valor de la otra en base a ese polinomio. La más sencilla es la regresión lineal, que busca una recta (polinomio de grado 1).

mas en el descriptivo

patrones frecuentes en las reglas de asociación, además de correlación

2. **Reglas de asociación**: Consisten en encontrar relaciones de tipo X→Y, donde "si ocurre X, entonces también ocurre Y". Su objetivo es buscar causalidad y correlación. X es el conjunto de elementos del antecedente, y el conjunto de elementos consecuente es Y. A es un evento de X y B un evento de Y. Se trata de buscar A.intersect(B), que son los eventos que cumplen la regla X→Y.

Hablar brevemente de soporte, confianza y lift.

mas en el descriptivo

su objetivo principal es descubrir estructuras ocultas o grupos naturales en los datos sin tener información previa sobre las categorías.

3. **Agrupamiento** (**clustering**): Se basa en encontrar conjuntos de elementos de modo que exista mucha similitud entre elementos del mismo conjunto y mucha diferenciación entre elementos de diferentes conjuntos. No existen los conjuntos previamente definidos, los decide la máquina.

predictivo

4. **Clasificación** (aprendizaje supervisado): Dado un conjunto de elementos, y N categorías, se quiere clasificar los elementos del conjunto en cada una de las N categorías. Para ello, se separa el conjunto en 2 en una proporción 80-20 aproximadamente. El primer conjunto es de entrenamiento y el segundo de test. Al de entrenamiento se le van pasando los datos y se le va diciendo de qué clase son, y posteriormente se prueba con el de test, midiendo precisión, recall, etc., y decidiendo si el entrenamiento ha sido bueno.

matriz de confusion
precisión
recall
accuracy
f-measure

Para evaluar hay que separar el conjunto de entrenamiento y el conjunto de validación.
No evaluar sobre el mismo conjunto utilizado para entrenar.
Evaluación cruzada de k (k-fold-cross-validation)
muestra inicial en k muestras igual tam
1 test k-1 train
se hace k veces
se promedian las k evaluaciones
k = numero de muestras => leave one out

Tema 9) Almacenes de datos:

- Explica brevemente en qué consisten las distintas fases del proceso ETL para almacenes de datos. ¿Qué es el staging area?

Las fases de construcción de un almacén de datos son ETL → **Extraction, Transformation & Load.**
1. **Extracción:** Los datos se obtienen de diversas fuentes de datos, que pueden ser desde bases de datos relacionales con operativa transaccional (OLTP), hasta ficheros de texto plano u otros documentos sin estructurar (webs, PDFs...). Es la fase más costosa, y se lleva a cabo con el objetivo de crear una nueva imagen inicial o actualizar una ya existente.
2. **Transformación:** Debido a que los datos provienen de fuentes de datos muy distintas, y el DW requiere unos datos estructurados, estos deben someterse a un proceso de transformación, que consiste en limpiar, traducir, fusionar, dividir y validar los datos.
3. **Carga:** Finalmente, los datos se cargan en el almacén. Esto puede realizarse de dos formas: una única carga completa o cargas incrementales (*Streaming:* volúmenes pequeños, *Lotes:* volúmenes grandes). Además, gracias a esta etapa, se puede llevar a cabo un mantenimiento de históricos (se puede hacer el seguimiento temporal de un dato).

gracias a esto podemos hacer un seguimiento temporal de los datos

Las **staging areas** facilitan los procesos de extracción y transformación de los datos antes de incluirlos en el DW. Es un almacén intermedio donde se almacenan antes de ser transformados y antes de ser cargados. De este modo, no hay que hacerlo todo de golpe: extraemos, lo guardamos en la *staging area*, transformamos, lo guardamos en la *staging area* y luego lo cargamos en el DW desde la última *staging area*.

- Explica brevemente el proceso de Data Warehouse: qué fases tiene, cuál es su propósito, qué usuarios participan, etc. Además, escribe tres ejemplos de tipos fuentes de datos que podrían formar parte del proceso.

El proceso de Data Warehouse tiene como propósito construir un repositorio de datos estructurados no volátiles (almacén de datos), tanto históricos como actuales con fines analíticos, para facilitar la toma de decisiones.

El proceso se compone de tres etapas:

Extracción: En la que se obtienen datos de diversas fuentes de datos (BD relacionales OLTP, ficheros de texto, CSVs, Excels, noticias, páginas webs, etc.), es un proceso costoso y su objetivo puede ser crear una nueva imagen o reforzar una ya existente.
Transformación: En la que los datos obtenidos de la fase anterior se limpian, estructuran, detallan, resumen, etc., con el objetivo de tenerlos en el formato adecuado.
Carga: Finalmente los datos se cargan en el almacén de datos, pudiendo hacerse de forma completa (en una carga) o incremental (varias cargas), ya sea por *streaming* (pequeños volúmenes de datos en cada carga) o por lotes (grandes volúmenes). En esta fase se pueden mantener históricos haciendo el seguimiento temporal de los datos.

gracias a esto podemos hacer un seguimiento temporal de los datos

Entre las etapas se puede definir una staging area, que es un almacén intermedio y facilita la extracción y la transformación. Se extraen de las fuentes y se llevan al staging area, luego se transforman y antes de ser cargados se guardan en otra staging area, y finalmente de la staging area se cargan en el DW.

Los usuarios que participan son principalmente directivos y analistas de datos de la empresa, que son los destinatarios finales. No obstante, también participan informáticos técnicos y administradores de sistemas.

Fuentes de datos que forman parte del proceso ya las he dicho antes, pero serían: BD relacional, fichero Excel, página web, CSV, documentos Word, entre otras. En general, cualquier fuente de datos imaginable.

Tema 10) Sistemas legados:

- Explica cuatro de los derechos que proporciona a los ciudadanos el Reglamento General de Protección de Datos (RGPD).

El RGPD proporciona los denominados derechos ARCO, además del derecho al olvido y el derecho a portabilidad.

1. Derecho de Acceso: Todo individuo tiene derecho a conocer si sus datos están siendo tratados y, en caso afirmativo, obtener acceso a dichos datos, la finalidad del tratamiento de estos, quién es el responsable de dicho tratamiento, los destinatarios de los datos, sus derechos (rectificación, cancelación, oposición), etc.

2. Derecho de Rectificación: Este derecho se puede ejercitar si los datos incluidos en una actividad de tratamiento de datos son erróneos, inexactos o incompletos. También se recoge que la modificación (rectificación) de los datos debe ser realizada sin hacer esperar al interesado, es decir, tan pronto como sea posible una vez ha ejercido su derecho.

3. Derecho de cancelación, muy relacionado con el derecho al olvido: Es el derecho de un individuo a solicitar la eliminación de sus datos personales cuando estos ya no sean necesarios para la finalidad, cuando se retira el consentimiento del tratamiento de estos (haciendo uso de su derecho a oposición), o cuando se hayan tratado los datos personales de forma ilícita.

4. Derecho de oposición: Un individuo puede oponerse al procesamiento de sus datos personales en ciertas circunstancias, como cuando estos se usan con fines de mercadotecnia o una misión de interés público.

- Enumera y explica los elementos que se consideran en el análisis de riesgos siguiendo la metodología MAGERIT.

Activos: Son los elementos del sistema de información, o estrechamente relacionados con él, que soportan la misión de la organización. Entre ellos, se encuentran información, servicios, datos, aplicaciones informáticas, equipos, instalaciones, personas, redes, etc.

Amenazas: Son eventos que pueden ocurrir a los activos y pueden causar un perjuicio en la organización. Las amenazas deben identificarse (origen natural, del entorno, de las aplicaciones, causadas malintencionadamente por personas). Los activos se degradan cuando les afecta una amenaza, por lo que se debe calcular el impacto y el riesgo para cada par <amenaza, activo>. El impacto se define como valor del activo * degradación causada por la amenaza, y el riesgo como impacto * probabilidad de que la amenaza se materialice.

Salvaguardas: Son medidas de protección desplegadas para reducir el potencial daño producido por las amenazas. Su objetivo es reducir la probabilidad de materialización de una amenaza, así como tratar de mitigar los daños causados por esta.

- Cuando te matriculas en un grado en la Universidad de Zaragoza se pide que proporciones una serie de datos personales. Indica qué tipo de información sería sensible de ser protegida por la legislación vigente en lo que a protección de datos se refiere.

El RGPD especifica que se considera información sensible de ser protegida aquella que es referente a datos personales. Y define datos personales como aquellos relacionados con personas físicas identificadas o identificables. Por lo tanto, todo lo que nos permita identificar a una persona física será información sensible de ser protegida, es decir, prácticamente todos los datos de la matrícula. Entre ellos:
Nombre y apellidos, DNI, Correo electrónico, Nacionalidad y país de nacimiento, NIP (lo piden para iniciar sesión), Cuenta bancaria, Ciudad de residencia, Condición de familia numerosa y Condición de si solicita o no beca.

Y, en menor medida, pero también está amparada por el RGPD:
Asignaturas matriculadas, sobre todo las segundas matrículas o terceras (pueden identificar si repites, etc.), cuánto pagas de matrícula en total, carrera que cursas...

- Describe brevemente los tipos de ficheros que se definen en la LOPD y enumera tres medidas que deben aplicarse a todos y cada uno de los tipos de ficheros según la LOPD y el RD donde se especifica su desarrollo.

(no hay respuesta, pone que está obsoleto)

- Derecho de acceso: se puede ejercitar para conocer si el responsable está tratando o no los datos de carácter personal del solicitante, y en caso de que se esté realizando dicho tratamiento, obtener información como: una copia, los fines del tratamiento, los destinatarios, el plazo previsto de conservación, …
- Derecho de rectificación: se puede ejercitar si los datos personales incluidos en una actividad de tratamiento son inexactos, y deben ser rectificados sin dilación indebida del responsable. Teniendo en cuenta los fines del tratamiento, mediante este derecho se puede solicitar que se completen los datos personales que sean incompletos, inclusive mediante una declaración adicional.
- Derecho de oposición: se puede ejercitar para oponerse a que el responsable realice un tratamiento de los datos personales en los siguientes supuestos: cuando sean objeto de tratamiento basado en una misión de interés público o en el interés legítimo o cuando el tratamiento tenga como finalidad la mercadotecnia directa.
- Derecho de supresión (“al olvido”): Es el derecho de un individuo a solicitar la eliminación de sus datos personales cuando estos ya no sean necesarios para la finalidad, cuando se retira el consentimiento del tratamiento de estos (haciendo uso de su derecho a oposición), o cuando se hayan tratado los datos personales de forma ilícita

Activos: Son los elementos clave del sistema de información o relacionados con él que son fundamentales para cumplir con la misión de la organización. Incluyen información, servicios, datos, aplicaciones, equipos, instalaciones, personas y redes, entre otros.

Amenazas: Son eventos que pueden dañar los activos de la organización. Pueden tener orígenes naturales (desastres), tecnológicos (fallos de sistemas) o humanos (errores o ataques intencionados). Se analiza el impacto que tendrían sobre los activos (valor del activo * degradación) y el riesgo asociado (impacto * probabilidad de ocurrencia).

Salvaguardas: Son medidas diseñadas para proteger los activos, reduciendo la probabilidad de que las amenazas se materialicen o mitigando sus efectos.Pueden ser técnicas (firewalls), organizativas (políticas de seguridad) o procedimentales (planes de respuesta ante incidentes).