SBD: Resumen

Tema 5: KDD, SEMMA, P^{3}TQ

El Proceso KDD (Knowledge Discovery in Databases)

La **Minería de Datos** es una disciplina fundamental para descubrir patrones útiles en grandes volúmenes de información. Forma parte de un proceso más amplio y metodológico llamado **KDD** (Knowledge Discovery in Databases), que significa descubrimiento de conocimiento en bases de datos. El KDD es un proceso no trivial, interactivo e iterativo para identificar **patrones válidos, nuevos, potencialmente útiles y comprensibles** en datos, utilizando técnicas multidisciplinares. Guía el análisis de datos y es la base de la mayoría de metodologías y procesos en la analítica de Big Data.

Las fases del proceso KDD son:

- **SELECCIÓN**: Consiste en elegir un subconjunto de datos, variables o ejemplos que sean relevantes para el objetivo del proceso, creando una "**tarjeta de datos**" o datos objetivo. Se eliminan datos irrelevantes o sensibles. Es importante verificar la viabilidad de las hipótesis con los datos disponibles.
- **PREPROCESO**: Se **limpia y prepara** el conjunto de datos objetivo. Esto incluye eliminar ruido, tratar valores ausentes (rellenándolos con valores prototípicos o por defecto, valorando el coste del error), corregir valores incorrectos y manejar valores atípicos (outliers) o normales que requieren tratamiento especial (inliers). También se pueden agregar datos. El resultado son los "**datos preprocesados**".
- **TRANSFORMACIÓN**: Se modifican los datos preprocesados para optimizar su uso en la minería. Esto puede implicar reducir variables, reestructurar datos o agrupar elementos similares. Es común aplicar **clustering** en esta fase para agrupar elementos, como clientes, lo que puede convertir un análisis no supervisado en uno supervisado. Los datos resultantes son los "**datos transformados**".
- **MINERÍA DE DATOS**: Esta es la **fase central** del proceso KDD, donde se aplican algoritmos sobre los datos transformados para **descubrir patrones útiles**. Los algoritmos tienen cuatro componentes básicos: el modelo (funciones como clasificación, clustering, resumen, análisis de secuencias), la representación (forma de expresar el conocimiento, como árboles de decisión, reglas, redes neuronales), el criterio de preferencia (métricas cuantitativas como precisión, recall, F1-score o factores cualitativos para elegir el mejor modelo), y el algoritmo de búsqueda (procedimiento para ajustar parámetros del modelo, como descenso del gradiente o algoritmos evolutivos). El resultado de esta fase son los "**patrones**" o modelos.
- **INTERPRETACIÓN/EVALUACIÓN**: Se evalúa la calidad, validez y utilidad de los patrones descubiertos. Se interpretan los resultados en el contexto del problema original para asegurar que sean comprensibles y aplicables, traduciendo el conocimiento extraído en decisiones concretas. El resultado es el "**conocimiento**" listo para su aplicación.

El proceso KDD es iterativo; si los resultados no son satisfactorios, se puede regresar a fases anteriores.

La Metodología SEMMA

SEMMA es una metodología propuesta por SAS Institute, centrada en **construir modelos predictivos**. Ayuda a organizar el proceso de minería de datos, especialmente cuando se utiliza el software SAS Enterprise Miner. A diferencia de KDD, **SEMMA no es un proceso completo de minería de datos**, sino una forma de estructurar las herramientas de SAS para aplicar técnicas de minería de forma ordenada. Es una **hoja de ruta práctica** sobre cómo trabajar con los datos una vez listos.

El nombre SEMMA proviene de las iniciales de sus cinco pasos:

- **Sample (Muestra)**: Tomar una **muestra representativa** del conjunto de datos completo para trabajar con un subconjunto más pequeño y manejable, manteniendo las características generales de los datos. Suele ser aleatoria. Es comparable a la etapa de 'selección' de KDD.
- **Explore (Exploración)**: Analizar los datos de la muestra para entenderlos mejor, buscando tendencias, patrones, valores atípicos y relaciones entre variables. Se usan gráficos, tablas, estadísticas o técnicas avanzadas. Ayuda a identificar variables útiles y entender el comportamiento de los datos. Puede verse como equivalente a la 'prospección' en KDD.
- **Modify (Modificación)**: Preparar los datos explorados para el modelado. Esto incluye corregir errores, tratar valores faltantes, eliminar variables irrelevantes y crear nuevas variables. La calidad del modelo depende de la preparación de los datos. Es comparable al 'preproceso o limpieza' en KDD.
- **Model (Modelado)**: Construir **modelos predictivos** utilizando técnicas como árboles de decisión, redes neuronales, regresión, etc., para hacer inferencias o predicciones. El objetivo es predecir el valor de variables. Es equivalente a la 'transformación/data mining' en KDD.
- **Assess (Valoración)**: **Evaluar el modelo** creado para medir su precisión, capacidad de generalización y si cumple los objetivos. Se usan datos diferentes a los del entrenamiento. Es fundamental para ver si el modelo es útil.

La Metodología P³TQ

P³TQ es una metodología propuesta por Dorian Pyle en 2003, diseñada para proyectos de minería de datos y análisis de información en **entornos de negocio**. Su nombre proviene de **Product, Place, Price, Time, Quantity**, factores comunes en datos de negocio. P³TQ es una **guía estratégica y táctica** que asegura que los proyectos de minería de datos estén **alineados con las necesidades reales del negocio**. Busca entender el contexto de negocio, identificar el problema, preparar datos adecuados, extraer información útil y comunicar resultados para apoyar la toma de decisiones.

Sus principales aportaciones son:

- Conecta el análisis con el negocio.
- Estructura el proyecto de lo estratégico a lo técnico.
- Evita malinterpretaciones de requerimientos.
- Permite diferentes enfoques según el punto de partida (dato, problema, estrategia, etc.).

La metodología se divide en **dos modelos principales**:

- **Modelo de Negocio**: Ayuda a entender **de dónde se parte, qué se quiere resolver y qué se necesita**. Su objetivo es **formular correctamente el problema** y definir los **requerimientos reales** del proyecto. Puede partir de diferentes escenarios como dato, problema/oportunidad, prospectiva, modelo definido o estrategia. Utiliza recursos como entrevistas o casos de negocio para llegar a los **datos requeridos** y los **requerimientos reales**.
- **Modelo de Explotación de Información**: Describe **cómo se trabaja con los datos para obtener valor** y resolver el problema. Es la parte **práctica y técnica**. Incluye etapas como **Preparación de datos** (preproceso, transformación), **Selección de herramientas y modelado inicial** (elegir métodos), **Ejecución** (aplicar modelos), **Evaluación de resultados** (analizar efectividad) y **Comunicación de resultados** (presentar hallazgos).

En resumen, KDD es el proceso global de descubrimiento de conocimiento, SEMMA es una metodología práctica para construir modelos predictivos (especialmente asociada a SAS), y P³TQ es una metodología orientada al negocio que conecta el análisis de datos con los objetivos y necesidades organizacionales.

Tema 6: CRISP-DM, TDSP, ASUM-DM

La Metodología CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM es una metodología desarrollada en 1999 por un grupo de empresas europeas con el objetivo de convertirse en un **estándar libre para proyectos de Minería de Datos**. Actualmente, es la metodología más utilizada como **referencia a nivel empresarial** para el desarrollo de este tipo de proyectos. Está muy **vinculada en su esencia al proceso KDD**. Su objetivo principal es **guiar todas las etapas del análisis**, desde la comprensión del problema de negocio hasta la implementación y uso del modelo. Es una metodología **generalista y adaptable** a múltiples industrias y tipos de datos, ideal para empresas que buscan extraer conocimiento útil y resolver problemas de negocio concretos a partir de sus datos.

CRISP-DM estructura el ciclo de vida de un proyecto de Ciencia de Datos en **6 fases que interactúan de forma iterativa**:

1. **Comprensión del Negocio (Business Understanding)**: En esta fase, se identifican los objetivos del proyecto desde una **perspectiva empresarial**. El propósito es traducir las necesidades del negocio en objetivos técnicos.
    - **Tareas principales**: Establecer objetivos del negocio, evaluar la situación actual (recursos, limitaciones, riesgos), definir los objetivos de minería de datos (metas técnicas alineadas con el negocio), y elaborar el plan del proyecto (fases, herramientas, personal, cronograma).
    - **Ejemplo**: Aumentar ventas online, traduciendo esto al objetivo técnico de predecir qué clientes tienen más probabilidad de comprar.

2. **Comprensión de los Datos (Data Understanding)**: Consiste en familiarizarse con los datos disponibles, **evaluar su calidad** y explorar su estructura para obtener una visión preliminar del problema y las posibles soluciones.
    - **Tareas principales**: Obtener los datos iniciales (acceso a fuentes, documentación), describir los datos (estadísticas, estructura), explorar los datos (patrones, relaciones, atípicos), y verificar la calidad de los datos (exactitud, completitud, consistencia).
    - **Ejemplo**: Observar que los datos de edad de los clientes tienen muchos valores faltantes o erróneos.
3. **Preparación de los Datos (Data Preparation)**: Esta fase se enfoca en **dejar los datos listos para el modelado**, incluyendo tareas de selección, limpieza, transformación y combinación de fuentes.
    - **Tareas principales**: Seleccionar los datos relevantes, limpiar datos (errores, duplicados, faltantes), construir atributos (generar nuevas variables), integrar datos (combinar fuentes), y formatear datos.
    - **Ejemplo**: Crear una variable "cliente frecuente" a partir del historial de compras.
4. **Modelado (Modeling)**: Se aplican **técnicas de minería de datos** para crear modelos predictivos o descriptivos, probando varios modelos y ajustando parámetros.
    - **Tareas principales**: Seleccionar la técnica de modelado (algoritmo), diseñar pruebas y evaluación (validación, por ejemplo, cruzada), construir el modelo (configuración, entrenamiento), y evaluar el modelo (medir rendimiento con métricas).
    - **Ejemplo**: Usar un modelo que predice con un 85% de acierto qué clientes volverán a comprar.
5. **Evaluación (Evaluation)**: Antes de implementar, se **evalúa si el modelo cumple los objetivos del negocio** y si está listo para el despliegue, o si es necesario repetir fases.
    - **Tareas principales**: Evaluar resultados (comparar con objetivos de negocio), revisar el proceso (identificar errores/mejoras), y definir los siguientes pasos (decisiones sobre uso, acciones futuras).
    - **Ejemplo**: El modelo es preciso, pero predice clientes ya fidelizados; requiere ajustar el enfoque para cumplir el objetivo real del negocio.
6. **Despliegue (Deployment)**: Consiste en **poner el modelo en producción** y asegurar su uso por parte de usuarios o sistemas que toman decisiones basadas en sus resultados.
    - **Tareas principales**: Planificar el despliegue (cómo y dónde implementar), plan de monitorización y mantenimiento (asegurar funcionamiento continuo), generar el informe final (documentar resultados y proceso), y revisión del proyecto (lecciones aprendidas).
    - **Ejemplo**: Crear un informe mensual que identifique clientes para promociones personalizadas.

La metodología CRISP-DM es **iterativa**; las flechas en sus diagramas indican que se puede volver a fases anteriores según sea necesario.

La Metodología TDSP (Team Data Science Process)

El **TDSP** (Team Data Science Process) es una metodología ágil creada por **Microsoft** para **organizar proyectos de ciencia de datos en equipo** de forma estructurada y eficiente. Su objetivo es ayudar a los equipos a desarrollar **soluciones predictivas e inteligentes que funcionen bien en producción**. Se centra en facilitar la **implementación de soluciones predictivas o de machine learning en producción de forma colaborativa**.

TDSP ofrece un **ciclo de vida estándar** para proyectos, una **estructura organizada de carpetas y documentación**, recomendaciones de **infraestructura y herramientas**, y buenas prácticas tomadas de **metodologías ágiles como Scrum**.

Tiene **5 fases principales** que se repiten de forma iterativa y son **similares a las de CRISP-DM, pero con un enfoque más técnico y orientado al trabajo en equipo**:

1. **Comprensión del problema empresarial (Business Understanding)**: Se define el objetivo del proyecto en términos de negocio (qué predecir o resolver) y se identifican las preguntas clave (regresión, clasificación, clustering, etc.). También se definen roles del equipo, métricas de éxito y fuentes de datos.
2. **Adquisición y comprensión de los datos (Data Acquisition & Understanding)**: Se obtienen y limpian los datos necesarios, se exploran las variables relevantes y se diseña una arquitectura para procesarlos y prepararlos.
3. **Modelado (Modeling)**: Se seleccionan las variables más importantes, se entrenan varios modelos de *machine learning*, y se comparan y evalúan para elegir el mejor.
4. **Despliegue (Deployment)**: El modelo elegido se publica en un entorno de producción (como una API) y se configura para ser usado por aplicaciones reales.
5. **Aceptación del cliente (Customer Acceptance)**: El cliente revisa y valida que el sistema cumple sus necesidades, se entrega un informe final y se entrena al cliente para operar el sistema.

TDSP define **roles específicos** para el equipo: Administrador de grupo, Líder de equipo, Líder de proyecto y Colaboradores. Su fortaleza radica en la **automatización, integración continua, colaboración y producción**. Es creciente su uso en entornos técnicos y de *machine learning*.

La Metodología ASUM-DM (Analytics Solutions Unified Method for Data Mining)

ASUM-DM (Analytics Solutions Unified Method for Data Mining) es una metodología propuesta por **IBM** como una **evolución de CRISP-DM**, adaptada a los desafíos de la Ciencia de Datos moderna, incluyendo **Big Data, análisis de texto, modelado predictivo y automatización**. Está orientada a ofrecer un **marco más detallado y empresarial** para soluciones analíticas complejas.

Es un enfoque **iterativo y estructurado** que busca acelerar el tiempo de obtención de valor, reducir riesgos e aportar consistencia y eficiencia mediante pasos, roles, plantillas y buenas prácticas definidas. Se aplica en proyectos que requieren una **gestión más formal, estructurada y alineada con las metas del negocio**, siendo ideal para consultorías o grandes organizaciones.

ASUM-DM propone **6 fases principales**: cinco secuenciales y una transversal:

1. **Análisis (Analysis)**: El objetivo es entender el problema de negocio, los objetivos del cliente y las restricciones del entorno. Las actividades incluyen identificar actores, definir entregables, comprender el contexto (datos, sistemas) y proponer ideas de solución. El resultado esperado es un **documento de requisitos** y un plan de trabajo inicial.
2. **Diseño (Design)**: Se planifica cómo se desarrollará la solución, incluyendo seleccionar técnicas analíticas/herramientas, diseñar la arquitectura técnica (cloud/on-premise, tecnologías), planificar recursos y establecer criterios de éxito. El resultado esperado es un **diseño funcional y técnico validado**.
3. **Configuración y Construcción (Configuration & Construction)**: Se desarrolla la solución mediante **ciclos iterativos**, incluyendo limpieza/transformación de datos, construcción/entrenamiento/validación de modelos, integración con sistemas y realización de pruebas. El resultado es una versión funcional del modelo o sistema, lista para pruebas piloto.
4. **Despliegue (Deployment)**: Consiste en implementar la solución en el entorno productivo, migrando la solución, documentando, capacitando a usuarios y estableciendo soporte. El resultado es un **sistema productivo funcionando** con usuarios reales.
5. **Operación y Optimización (Operation & Optimization)**: Se monitorea y mejora la solución a lo largo del tiempo, verificando el rendimiento (monitoreo del modelo), realizando ajustes (reentrenamiento si cambia el contexto) y ampliando la solución si es exitosa. El resultado es una **solución optimizada y sostenible** con impacto medible.
6. **Gestión de Proyecto (Project Management)**: Esta es una fase **transversal** que ocurre en paralelo a las demás, asegurando el avance coordinado. Incluye establecer cronogramas, gestionar recursos, supervisar riesgos/cambios y gestionar la comunicación. El rol clave es el Project Manager o Scrum Master.

ASUM-DM es sólido para **entornos corporativos**, con gestión de riesgos y alineación organizativa, utilizando plantillas, *checklists* y buenas prácticas estandarizadas.

En resumen, estas tres metodologías (CRISP-DM, TDSP, ASUM-DM), junto con las previamente discutidas (KDD, SEMMA, P³TQ), ofrecen diferentes marcos para abordar proyectos de análisis y minería de datos, cada una con sus particularidades y enfoques (generalista y empresarial, técnico y de equipo para producción, y empresarial estructurado para grandes organizaciones, respectivamente).