UD 6

FASES DE LOS MODELOS. VALORACIÓN. INTERPRETACIÓN. DESPLIEGUE.

1. Introducción

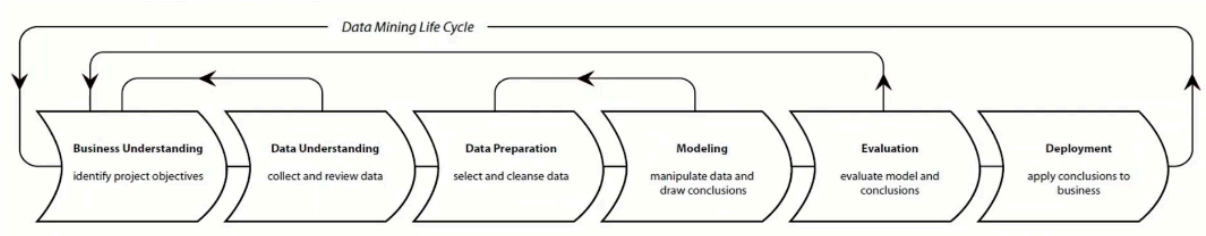
En este capítulo se presentan tres nuevas metodologías:

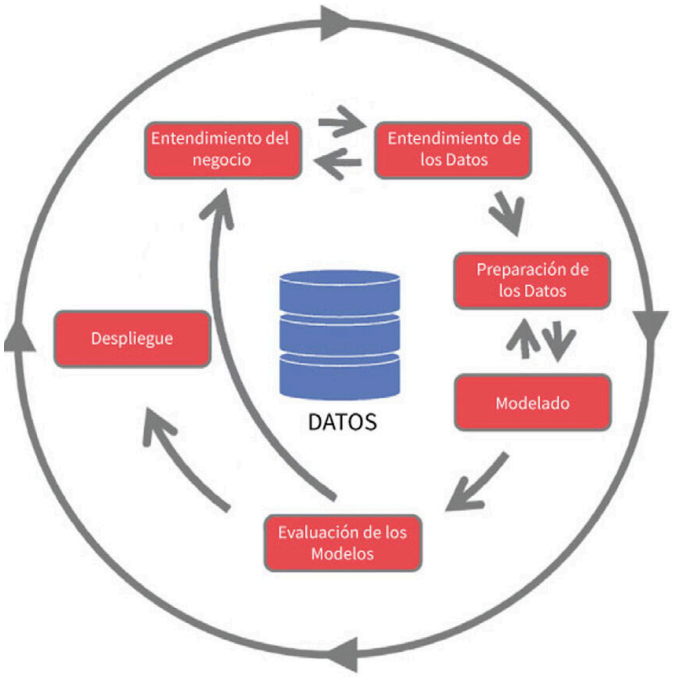
- **CRISP-DM (*Cross-Industry Standard Process for Data Mining*).** Quizá la más usada hoy en día en entornos empresariales (muy vinculada en su esencia al KDD). Se utiliza para **desarrollar proyectos de minería de datos** de forma estructurada, especialmente en entornos empresariales. Su objetivo es **guiar todas las etapas del análisis**, desde entender el problema de negocio hasta implementar y usar el modelo. Es una metodología **generalista**, adaptable a múltiples industrias y tipos de datos. Es ideal para proyectos de análisis de datos en empresas que quieren **extraer conocimiento útil** a partir de sus datos y **resolver problemas de negocio concretos**.
- **TDSP (*Team Data Science Process*).** Es una metodología desarrollada por Microsoft para **estructurar proyectos de ciencia de datos y aprendizaje automático** en equipos multidisciplinares. Integra herramientas de software, control de versiones, automatización de experimentos y colaboración. Está centrada en facilitar la **implementación de soluciones predictivas o de inteligencia artificial** en producción. Se utiliza principalmente en entornos de desarrollo profesional donde se requiere **desplegar modelos predictivos o de *machine learning*** en producción de forma colaborativa.
- **ASUM-DM (*Analytics Solutions Unified Method for Data Mining*).** Es una metodología propuesta por IBM que extiende a CRISP-DM, incluyendo aspectos de **gestión de proyectos analíticos y su alineación con los objetivos del negocio**. Está orientada a ofrecer un marco más detallado y empresarial para el desarrollo de soluciones analíticas complejas. Se aplica en proyectos donde se requiere una **gestión más formal, estructurada y alineada con las metas del negocio**, ideal para consultorías o grandes organizaciones.

2. La metodología CRISP-DM

Como ya hemos visto en el tema anterior, existen metodologías de minería de datos muy conocidas y usadas, como SEMMA o P^3TQ, además del proceso KDD. Sin embargo, en 1999, un grupo de empresas europeas, NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), desarrollaron una nueva metodología con el objetivo de convertirse en un estándar, y de libre distribución, denominada CRISP-DM, *Cross-Industry Standard Process for Data Mining*, que estructura el ciclo de vida de un proyecto de Ciencia de Datos en 6 fases que interactúan entre sí de forma iterativa durante el desarrollo del proyecto.

CRISP-DM en la actualidad es la metodología que más se utiliza como referencia a nivel empresarial para el desarrollo de proyectos de Minería de Datos. Las fases son las siguientes (ver las siguientes dos figuras).





Las flechas internas de la figura 2 indican las **relaciones más frecuentes** entre las fases, aunque **se pueden establecer relaciones entre cualquier fase** o tarea del proceso, variando de acuerdo con los **objetivos**, el **contexto** o por el **interés del usuario sobre los datos**. El círculo exterior simboliza la naturaleza cíclica del proceso de modelado.

Vamos a explicar las etapas del ciclo siguiendo un ejemplo:

1. **Comprensión del negocio**. En esta fase se identifican los objetivos del proyecto desde una perspectiva empresarial, no técnica. El propósito es traducir las necesidades del negocio en objetivos técnicos que guíen el resto del proyecto.

   Tareas principales:

   - **Establecimiento de los objetivos del negocio.**
     Analizar el contexto, antecedentes, objetivos y criterios de éxito del proyecto.

   - **Evaluación de la situación actual.**
     Recopilar recursos disponibles, limitaciones, riesgos y beneficios.

- **Definición de los objetivos de minería de datos.**
  Establecer metas técnicas específicas alineadas con los objetivos del negocio.

- **Elaboración del plan del proyecto.**
  Definir las fases del trabajo, herramientas, personal y cronograma.

Ejemplo: una tienda quiere aumentar sus ventas online. El objetivo del negocio será predecir qué clientes tienen más probabilidad de comprar.

2. **Comprensión de los datos**. Una vez tenemos claros los objetivos, esta fase 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:

   - **Obtención de los datos iniciales.**
     Acceso a las fuentes de datos y documentación del proceso.

   - **Descripción de los datos.**
     Resumen estadístico de variables y estructuras de datos.

   - **Exploración de los datos.**
     Identificación de patrones iniciales, relaciones y valores atípicos.

   - **Verificación de la calidad de los datos.**
     Evaluar dimensiones como exactitud, completitud, consistencia, oportunidad y relevancia.

   Ejemplo: vemos que los datos de edad de los clientes tienen muchos valores faltantes o erróneos.

3. **Preparación de los datos**. Esta fase se centra en dejar los datos listos para el modelado. Incluye tareas de selección, limpieza, transformación y combinación de distintas fuentes de datos.

   Tareas principales:

   - **Selección de los datos relevantes.**
     Justificación de los datos incluidos o excluidos.

   - **Limpieza de datos.**
     Corrección de errores, eliminación de duplicados, tratamiento de valores faltantes.

   - **Construcción de atributos.**
     Generación de nuevas variables derivadas a partir de las existentes.

- **Integración de datos.**
  Combinación de diversas fuentes de datos.

- **Formateo de datos.**
  Conversión de formatos y estructuras para facilitar el modelado.

Ejemplo: creamos una variable “cliente frecuente” a partir del historial de compras.

4. **Modelado**. Se aplican técnicas de minería de datos para crear modelos predictivos o descriptivos. Es común probar varios modelos y ajustar sus parámetros para mejorar el rendimiento.

   Tareas principales:

   - **Selección de la técnica de modelado.**
     Escoger el algoritmo adecuado según el problema y los datos disponibles.

   - **Diseño de pruebas y evaluación.**
     Definir cómo se validarán los modelos (por ejemplo, con validación cruzada).

   - **Construcción del modelo.**
     Configuración de parámetros y entrenamiento del modelo.

   - **Evaluación del modelo.**
     Medición del rendimiento con métricas adecuadas (precisión, recall, etc.).

   Ejemplo: usamos un modelo que predice con un 85% de acierto qué clientes volverán a comprar.

5. **Evaluación**. Antes de implementar el modelo, se evalúa si cumple con los objetivos del negocio. Se analizan los resultados obtenidos y se decide si el modelo está listo para el despliegue o si es necesario repetir alguna fase.

   Tareas principales:

   - **Evaluación de resultados.**
     Comparar los resultados con los objetivos de negocio definidos al inicio.

   - **Revisión del proceso.**
     Identificar posibles errores o mejoras.

   - **Definición de los siguientes pasos.**
     Tomar decisiones sobre el uso del modelo y acciones futuras.

   Ejemplo: aunque el modelo es preciso, no sirve si predice clientes que ya están fidelizados. Hay que ajustar el enfoque.

6. **Despliegue**. Consiste en poner el modelo en producción y asegurar su uso por parte de los usuarios o sistemas que tomarán decisiones a partir de sus resultados. Puede incluir desde informes simples hasta sistemas automáticos en tiempo real.

   Tareas principales:

   - **Planificación del despliegue.**
     Definir cómo y dónde se implementará el modelo.

   - **Plan de monitorización y mantenimiento.**
     Asegurar que el modelo siga funcionando correctamente con el tiempo.

   - **Generación del informe final.**
     Documentar los resultados y el proceso.

   - **Revisión del proyecto.**
     Extraer lecciones aprendidas y buenas prácticas para futuros proyectos.

   Ejemplo: se crea un informe mensual que indica qué clientes deberían recibir promociones personalizadas.

Ejemplo aplicación metodología CRISP-DM	
Enunciado: una empresa de manufactura quiere evitar paradas inesperadas en su línea de producción. Para eso, decide usar datos de sensores instalados en las máquinas para **predecir fallos antes de que ocurran**.	
1. Comprensión del negocio	**Objetivo del negocio.** Reducir el tiempo de inactividad por fallos de maquinaria. **Preguntas clave.** ● ¿Podemos anticiparnos a las fallas? ● ¿Qué variables indican desgaste o mal funcionamiento? **Plan inicial.** Crear un modelo que detecte patrones de fallas inminentes para activar mantenimiento preventivo solo cuando sea necesario.
2. Comprensión de los datos	**Datos disponibles.** ● Registros de sensores (temperatura, vibración, presión, horas de uso). ● Fechas y tipos de fallos ocurridos anteriormente. ● Información técnica de cada máquina. **Exploración inicial.** ● Se observa que el aumento de temperatura y vibración precede algunos fallos. ● Algunas máquinas no tienen historial completo, lo que limita el análisis.

3. Preparación de los datos	**Limpieza.** ● Se eliminan registros incompletos o corruptos de sensores. ● Se alinean los datos en función del tiempo para analizarlos como series temporales. **Transformaciones.** ● Se crean variables como: "media móvil de vibración", "incremento de temperatura en 24h". ● Se etiqueta cada registro como "fallo/no fallo" para el entrenamiento del modelo. **Datos listos.** Un dataset cronológico con las condiciones de cada máquina y si falló o no.
4. Modelado	**Modelo elegido.** *Random forest*, por su capacidad para manejar datos complejos y ruidosos. **Entrenamiento.** Se usa un historial de datos con fallos conocidos para que el modelo aprenda los patrones previos a un fallo. **Resultado.** El modelo predice con un 85% de precisión cuándo una máquina fallará en las próximas 48 horas.
5. Evaluación	**Evaluación del modelo.** ● El modelo detecta correctamente la mayoría de fallos reales. ● Se evalúa el impacto económico: menos paradas, mejor planificación de mantenimiento. **Decisión.** Se aprueba para uso piloto en una línea de producción.
6. Despliegue	**Despliegue práctico.** El modelo se integra en el sistema de control de planta. Si detecta riesgo alto de fallo, notifica al jefe de mantenimiento. **Plan de mantenimiento.** El mantenimiento se realiza solo cuando el modelo lo recomienda (ya no en fechas fijas), ahorrando recursos. **Monitorización.** Se revisan las predicciones semanalmente y se recalibra el modelo si es necesario.

Ejercicio 6.1

Analiza los factores que influyen en la entrega tardía de pedidos

Supuesto real:
Eres parte del equipo de datos de una empresa de e-commerce. El gerente de operaciones está preocupado porque muchos pedidos llegan tarde a los clientes. Quiere entender por qué ocurre esto y qué se puede hacer para prevenirlo.

Aplica la metodología CRISP-DM explicando qué realizarías en cada una de las fases. Te ofrezco esta guía para que sigas sus pasos:

Fase 1: Comprensión del negocio

- ¿Cuál es el objetivo principal del negocio?
- Tarea 1.1: redacta en una frase el objetivo.

Fase 2: Comprensión de los datos

- Tarea 2.1: crea una hoja de cálculo con 10 registros de datos ficticios sobre:
    - Fecha del pedido
    - Ciudad de destino
    - Tiempo de entrega (días)
    - Medio de envío (normal / urgente)
    - ¿Llegó tarde? (Sí / No)
- Tarea 2.2: analiza qué variables podrían estar relacionadas con la tardanza.

Fase 3: Preparación de datos

- Tarea 3.1:
    - Elimina columnas irrelevantes (si las hay).
    - Clasifica los pedidos como “rápidos”, “normales” o “lentos” según los días de entrega.
    - Identifica registros con datos faltantes o inconsistentes.

Fase 4: Modelado

- Tarea 4.1:
    - ¿Qué variables usarías para predecir si un pedido llegará tarde?
    - Si tuvieras que crear una simple regla (“si… entonces…”), ¿cómo sería?. Escríbela.

Fase 5: Evaluación

- Tarea 5.1:
    - Explica si tu regla o modelo funciona bien para todos los casos.
    - ¿Cumple con el objetivo del negocio?¿Qué errores podrías estar cometiendo? ¿Qué mejorarías? ¿Faltan datos?

Fase 6: Despliegue

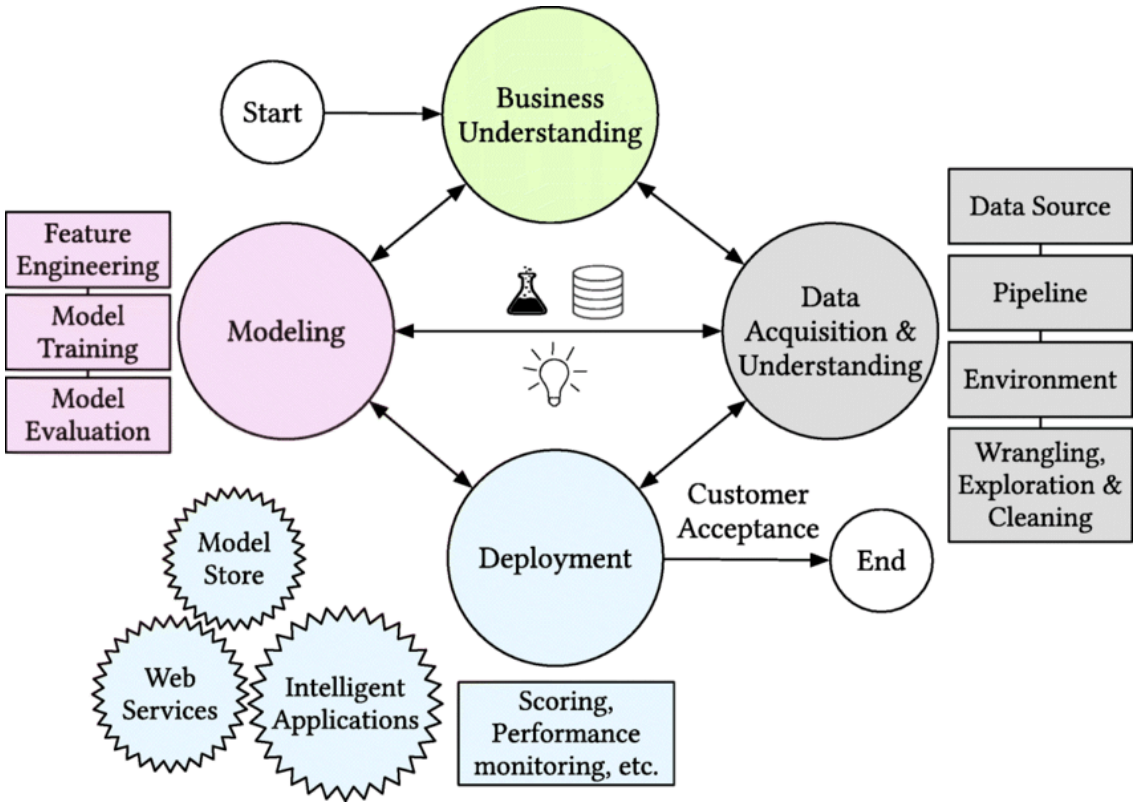
- Tarea 6.1:
    - Imagina que debes dar una recomendación al gerente. ¿Qué le dirías que haga a partir de tus hallazgos?
    - ¿Cómo podrías automatizar esta solución en la empresa?

3. TDSP

El ***Team Data Science Process*** **(TDSP)** es una metodología ágil (se trabaja en **ciclos cortos** (iteraciones), en los que se construye una parte funcional del producto, se prueba, se mejora y se continúa) creada por **Microsoft** para **organizar proyectos de ciencia de datos en equipo**, de forma estructurada y eficiente. Su objetivo es ayudar a equipos a desarrollar soluciones predictivas e inteligentes que funcionen bien en producción. TDSP ofrece:

1. Un **ciclo de vida estándar** para proyectos de ciencia de datos.
2. Una **estructura de carpetas y documentación** organizada.
3. Recomendaciones de **infraestructura y herramientas** para facilitar la colaboración.
4. **Buenas prácticas** tomadas de **metodologías ágiles** como Scrum.

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



1. Comprensión del problema empresarial

- Se define el objetivo del proyecto en términos de negocio: ¿qué se quiere predecir o resolver?
- Se identifican las preguntas clave, como:
    - ¿Cuánto? → regresión
    - ¿Qué categoría? → clasificación
    - ¿Qué grupo? → clustering
    - ¿Es raro? → detección de anomalías
    - ¿Qué decisión tomar? → recomendación
- Se definen roles del equipo, métricas de éxito y posibles fuentes de datos.

2. Adquisición y comprensión de los datos

- Se obtienen y limpian los datos necesarios.
- Se exploran las variables más relevantes.
- Se diseña una arquitectura para procesarlos y prepararlos.

3. Modelado

- Se seleccionan las variables más importantes.
- Se entrenan varios modelos de *machine learning*.
- Se comparan y evalúan los modelos para elegir el mejor.

4. Despliegue

- El modelo elegido se publica en un entorno de producción (como una API).
- Se configura para que pueda ser usado por aplicaciones reales.

5. Aceptación del cliente

- El cliente revisa el sistema y valida que cumple sus necesidades.
- Se entrega un informe final con la documentación del proyecto.
- Se entrena al cliente para que pueda operar el sistema por sí mismo.

Los **roles** que existen en un equipo TDSP son:

Rol	Función principal
Administrador de grupo	Supervisa el área de ciencia de datos.
Líder de equipo	Coordina a los científicos de datos.
Líder de proyecto	Gestiona las tareas diarias del proyecto.
Colaboradores	Ejecutan tareas técnicas o analíticas.

Ejemplo aplicación metodología TDSP	
Enunciado	Una empresa de manufactura quiere evitar paradas inesperadas en su línea de producción. Para eso, decide usar datos de sensores instalados en las máquinas para **predecir fallos antes de que ocurran**.
1. Comprensión del problema empresarial	**Objetivo del negocio:** Evitar paradas inesperadas reduciendo el tiempo de inactividad. **Preguntas clave:** ¿Se pueden predecir los fallos? ¿Qué variables anticipan una falla? **Plan inicial:** Desarrollar un modelo predictivo con datos de sensores para activar mantenimiento preventivo solo cuando sea necesario.
2. Adquisición y comprensión de los datos	**Datos disponibles:** Históricos de sensores (temperatura, vibración, presión), historial de fallos, especificaciones técnicas de máquinas. **Exploración de datos:** Identificación de correlaciones entre fallos y aumentos en vibración o temperatura. **Problemas iniciales:** Datos incompletos o sensores con lecturas anómalas.
3. Modelado	**Modelo elegido:** *Random Forest* (por su eficacia con datos complejos y ruidosos). **Entrenamiento:** Con registros etiquetados como fallo/no fallo. **Resultado:** El modelo predice con 85% de precisión cuándo fallará una máquina en las próximas 48 horas.
4. Despliegue	**Integración:** El modelo se conecta al sistema de monitoreo de planta. **Acciones automáticas:** Alerta al jefe de mantenimiento si se detecta un alto riesgo. **Plan de mantenimiento dinámico:** Activación solo cuando el modelo lo recomienda.
5. Aceptación del cliente	**Evaluación técnica y económica:** Se detectan la mayoría de fallos reales, con reducción de tiempos muertos y mejora en planificación. **Decisión del cliente:** Se aprueba su uso en una línea piloto. **Revisión periódica:** Se recalibra el modelo semanalmente con nuevos datos.

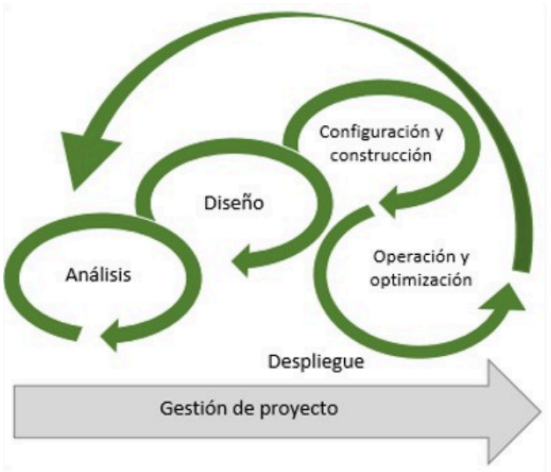
4. ASUM DM

ASUM-DM (*Analytics Solutions Unified Method for Data Mining*) es una metodología desarrollada por **IBM** como evolución de **CRISP-DM**, adaptada a los nuevos desafíos de la Ciencia de Datos moderna, como el uso de **Big Data**, **análisis de texto**, **modelado predictivo** y **automatización de procesos**.

Es un enfoque **iterativo y estructurado** que busca:

- **Acelerar el tiempo** de obtención de valor.
- **Reducir riesgos** en la implementación.
- Aportar **consistencia y eficiencia** mediante pasos, roles, plantillas y buenas prácticas definidas.

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



1. **Análisis**
    - **Objetivo**. Entender el problema de negocio, los objetivos del cliente y las restricciones del entorno.
    - **Actividades.**
        - Identificar los actores clave y sus necesidades.
        - Definir claramente los entregables esperados.
        - Comprender el contexto: ¿Qué datos existen? ¿Qué sistemas los generan?
        - Proponer ideas preliminares sobre la posible solución analítica.
    - **Resultado esperado**: un **documento de requisitos** y un plan de trabajo inicial.

2. **Diseño**
    - **Objetivo.** Planificar cómo se desarrollará la solución.
    - **Actividades:**
        - Seleccionar técnicas analíticas y herramientas (machine learning, visualización, NLP, etc.).
        - Diseñar la arquitectura técnica (¿cloud o on-premise?, ¿Python, R, Spark?).

        - Planificar los recursos necesarios: roles, herramientas, infraestructura.
        - Establecer los criterios de éxito del proyecto.
    - **Resultado esperado**: un diseño funcional y técnico validado con los stakeholders.

3. **Configuración y construcción**
    - **Objetivo.** Desarrollar la solución mediante ciclos iterativos.
    - **Actividades:**
        - Limpieza, transformación y exploración de los datos.
        - Construcción, entrenamiento y validación de modelos.
        - Integración con sistemas internos o externos (ej.: APIs, dashboards).
        - Realización de pruebas en entornos controlados (pruebas unitarias y de sistema).
    - **Resultado esperado**: una versión funcional del modelo o sistema, lista para pruebas piloto.

4. **Despliegue**
    - **Objetivo**. Implementar la solución en el entorno productivo.
    - **Actividades:**
        - Migrar la solución al entorno final sin afectar operaciones.
        - Documentar completamente el sistema.
        - Capacitar a los usuarios o stakeholders clave.
        - Establecer soporte técnico inicial.
    - **Resultado esperado**: un sistema productivo funcionando con usuarios reales.

5. **Operación y optimización**
    - **Objetivo**. Monitorear y mejorar la solución a lo largo del tiempo.
    - **Actividades:**
        - Verificar si el sistema mantiene su rendimiento (monitoreo del modelo).
        - Realizar ajustes si el contexto cambia (reentrenamiento, nuevos datos).
        - Ampliar la solución a otras áreas o usuarios si ha sido exitosa.
    - **Resultado esperado**: una solución optimizada y sostenible, con impacto medible.

6. **Gestión de proyecto (fase transversal)**. Esta fase ocurre en paralelo a todas las anteriores.
    - **Objetivo**. Asegurar el avance coordinado del proyecto.
    - **Actividades:**
        - Establecer cronogramas, hitos y prioridades.
        - Asignar tareas y gestionar recursos humanos.
        - Supervisar riesgos, cambios y bloqueos.
        - Gestionar la comunicación con stakeholders.
    - **Rol clave**: el Project Manager o Scrum Master, quien vela por el cumplimiento del plan y la resolución de problemas.

COMPARATIVA DE METODOLOGÍAS			
Aspecto	**CRISP-DM**	**TDSP**	**ASUM-DM**
Origen	Desarrollada por empresas europeas en 1999.	Desarrollada por Microsoft.	Desarrollada por IBM.
Enfoque	Generalista, orientado a negocio, adaptable a muchos contextos.	Ágil, colaborativo, centrado en proyectos técnicos y productivos	Empresarial, estructurado, orientado a grandes organizaciones y soluciones escalables.
Iteración	Iterativo, permite volver a fases anteriores según necesidades.	Iterativo, con ciclos cortos y mejora continua (influenciado por Scrum).	Iterativo en ciclos, con gestión paralela del proyecto.
Colaboración en equipo	No especifica roles explícitos, más centrado en el proceso.	Define roles específicos en el equipo: líder técnico, administrador, colaboradores.	Enfatiza gestión, roles y control del proyecto de forma más estructurada.
Uso actual	Muy utilizado como referencia base en la industria.	Creciente uso en entornos técnicos y de *machine learning*.	Más adoptado en contextos empresariales complejos, con requisitos de calidad y soporte.
Documentación	Flexible y conceptual.	Estructura de carpetas, documentación compartida y código versionado.	Plantillas, *checklist* y buenas prácticas estandarizadas.
Fortalezas	Simplicidad, enfoque al negocio, adaptable.	Fuerte en automatización, integración continua, colaboración y producción.	Sólido para entornos corporativos, con gestión de riesgos y alineación organizativa.
Debilidades	Poca guía sobre infraestructura o herramientas.	Menos amigable para principiantes o pequeños equipos.	Complejidad y posible sobrecarga si el proyecto es pequeño.

Ejercicio final
Aplicación comparativa de metodologías

Caso práctico: una empresa de servicios de salud quiere reducir las ausencias injustificadas de pacientes a sus citas médicas. Sospechan que hay patrones en los historiales de pacientes y el tipo de cita que podrían predecir si un paciente no asistirá. Te piden desarrollar un modelo para anticipar estas ausencias y tomar medidas preventivas (como enviar recordatorios o cambiar el tipo de cita).

Parte 1: Aplicar la metodología CRISP-DM

- **Comprensión del negocio.** ¿Cuál es el objetivo empresarial? ¿Qué impacto tendría reducir las ausencias?
- **Comprensión de los datos.** ¿Qué datos podrías necesitar? Ej.: edad del paciente, tipo de cita, historial de asistencia, etc.
- **Preparación de los datos.** ¿Cómo prepararías los datos? ¿Qué variables crearías?
- **Modelado.** ¿Qué técnica usarías para predecir si un paciente faltará?
- **Evaluación.** ¿Tu modelo cumple los objetivos? ¿Detecta correctamente a los pacientes propensos a faltar?
- **Despliegue.** ¿Cómo implementarías el modelo en la práctica? ¿Quién usaría sus resultados?

Parte 2: Aplicar la metodología TDSP

- **Comprensión del problema empresarial.** Replantea el objetivo como una pregunta de ciencia de datos. Ej.: ¿Será clasificación o detección de anomalías?
- **Adquisición y comprensión de los datos.** ¿Qué arquitectura y canalización usarías? ¿Cómo asegurarías la calidad?
- **Modelado.** ¿Entrenarías varios modelos? ¿Cómo elegirías el mejor?
- **Despliegue.** ¿Harías una API? ¿Dónde se integraría este sistema?
- **Aceptación del cliente.** ¿Cómo presentarías resultados al cliente? ¿Qué plan de validación harías?

Parte 3: Aplicar la metodología ASUM-DM

- **Análisis.** ¿Qué necesidades del usuario identificas? ¿Qué entregables se esperan?
- **Diseño.** ¿Qué recursos técnicos y humanos necesitarás? ¿Qué entorno usarás (cloud, local, etc.)?
- **Configuración y construcción.** ¿Cómo organizarías ciclos iterativos para desarrollar y probar la solución?
- **Despliegue.** ¿Qué plan tendrías para una transición sin impacto en operaciones actuales?
- **Operación y optimización.** ¿Cómo garantizarías el rendimiento del sistema a lo largo del tiempo?
- **Gestión del proyecto.** ¿Qué tareas y roles definirías para coordinar todo el trabajo?