Fundamentos de Data Science

Unidad 3: Ciclo de Vida de un Proyecto de Ciencia de Datos

Semana 9 - Metodologías y características de un proyecto de ciencia de datos

CC216

Unidad 3

Ciclo de Vida de un Proyecto de Ciencia de Datos

Objetivos

- Conocer las principales diferencias entre un Proyecto de Ing. de Software y otro de Ciencia de Datos.
- Identificar cada una de las fases de un proyecto de DS.
- Conocer las principales metodologías utilizadas para la gestión de proyectos de TI.

Metodologías y características de un proyecto de ciencia de datos

Contenido

1. Proyecto de Ing. de Software vs Proyecto de Ciencia de Datos
2. Etapas de un Proyecto de Ciencia de Datos
3. Metodologías para la Gestión de Proyectos

.

Conclusiones y Preguntas

“La información es una fuente de aprendizaje. Pero a menos que esté organizada, procesada y disponible para las personas adecuadas, en un formato para la toma de decisiones, es una carga, no un beneficio ".

- William Pollard (1828-1893), clérigo inglés

1. Proyecto de Ing. de Software vs Proyecto de Ciencia de Datos

¿Qué entendemos por
"PROYECTO"?

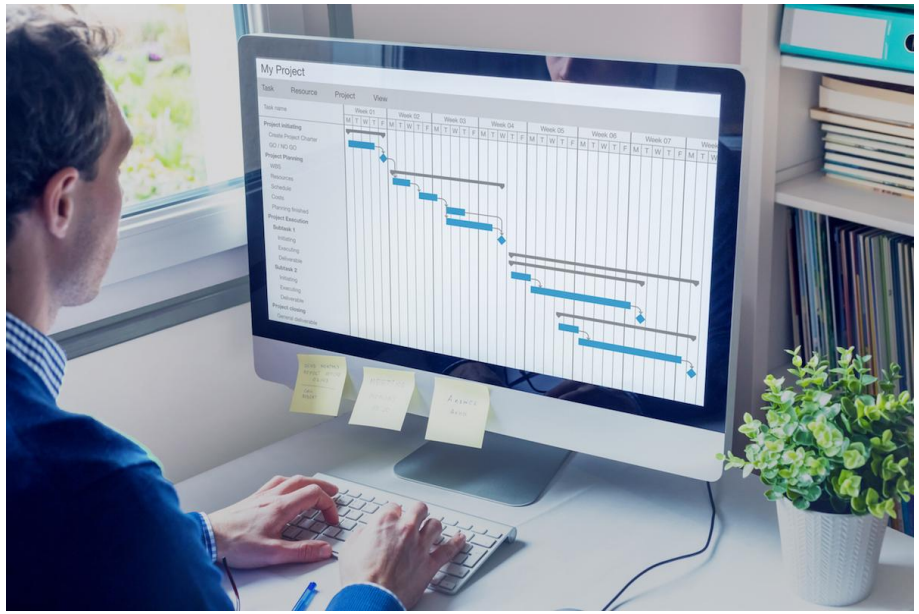
- ***Esfuerzo*** para crear o modificar un producto o servicio específico
- ***Trabajo temporal*** con un comienzo y un final claros

Un proyecto, involucra recursos:

- Personas
- Tecnologías
- Tiempo
- Dinero
- Nuevos productos o soluciones únicas

1. Proyecto de Ing. de Software vs Proyecto de Ciencia de Datos

Proyecto de Ing. de Software



Es un **proyecto de tecnología de la información (TI)** definido por una **fecha de inicio** y **fin**, por lo general, con **hitos** y **objetivos específicos** que deben cumplirse durante el ciclo de desarrollo.

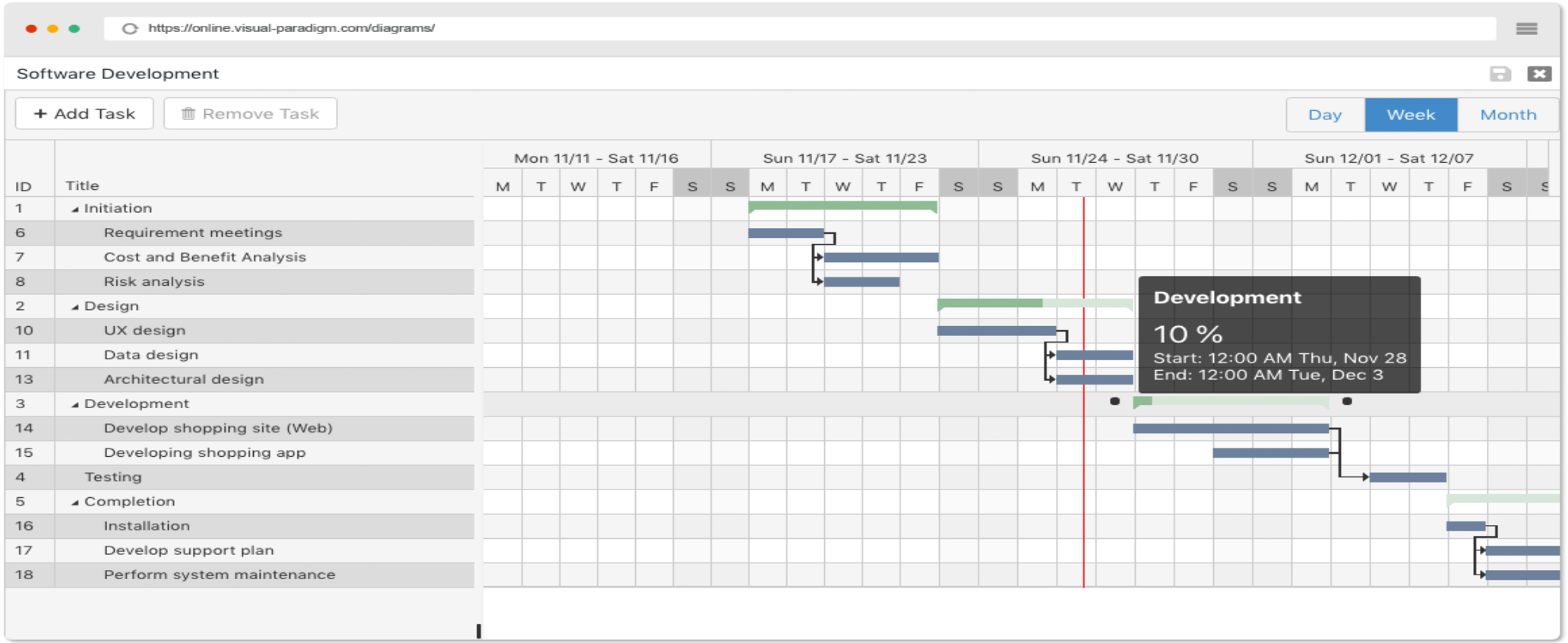
Para tener éxito en su ejecución, se debe mantener un equilibrio entre estos tres factores:

- Tiempo
- Coste
- Alcance

Los proyectos de ingeniería de software fluyen a través del ciclo de vida de desarrollo de software (SDLC), que a menudo se representa con cinco o seis fases (cada fase es distinta, por lo que cada una debe completarse antes de pasar a la fase siguiente).

1. Proyecto de Ing. de Software vs Proyecto de Ciencia de Datos

Diagrama de Gantt



1. Proyecto de Ing. de Software vs Proyecto de Ciencia de Datos

Gestión de Proyectos de Ing. de Software - Fases

1. Toma de requerimientos	2. Análisis	3. Diseño	4. Construcción	5. Pruebas	6. Puesta en Producción
Documentación al detalle de todas las solicitudes / requerimientos del cliente	Elaboración de diagramas / mapas que detallan el análisis funcional de los requerimientos	Elaboración de documentos, diagramas y especificaciones técnicas dejando por escrito y claramente, todo lo que los programadores deberán desarrollar.	Se traduce en código todo el diseño técnico de la etapa anterior. Los programadores deben seguir las especificaciones perfectamente detalladas.	El cliente valida el producto entregado luego que participa en pruebas manuales que simulan el uso que se hará de la aplicación. El equipo de pruebas por lo general no es el mismo que construye.	El cliente aprueba el producto entregado. Inicia el uso de la aplicación en “vivo” Antes de la puesta en producción, por lo general se realizan sesiones de capacitación dirigido a usuarios finales. Entrega de manuales de uso.

1. Proyecto de Ing. de Software vs Proyecto de Ciencia de Datos

Proyecto de Ciencia de Datos

Es también un **proyecto de tecnología de la información (TI)** definido por una **fecha de inicio** y **fin**, pero en el que **se desconoce el tiempo necesario** para dar muchos pasos, por tanto, no se pueden tener **hitos** y **objetivos específicos** que deben cumplirse durante el ciclo de desarrollo.

El seguimiento del progreso es mas ambiguo.

1. Proyecto de Ing. de Software vs Proyecto de Ciencia de Datos

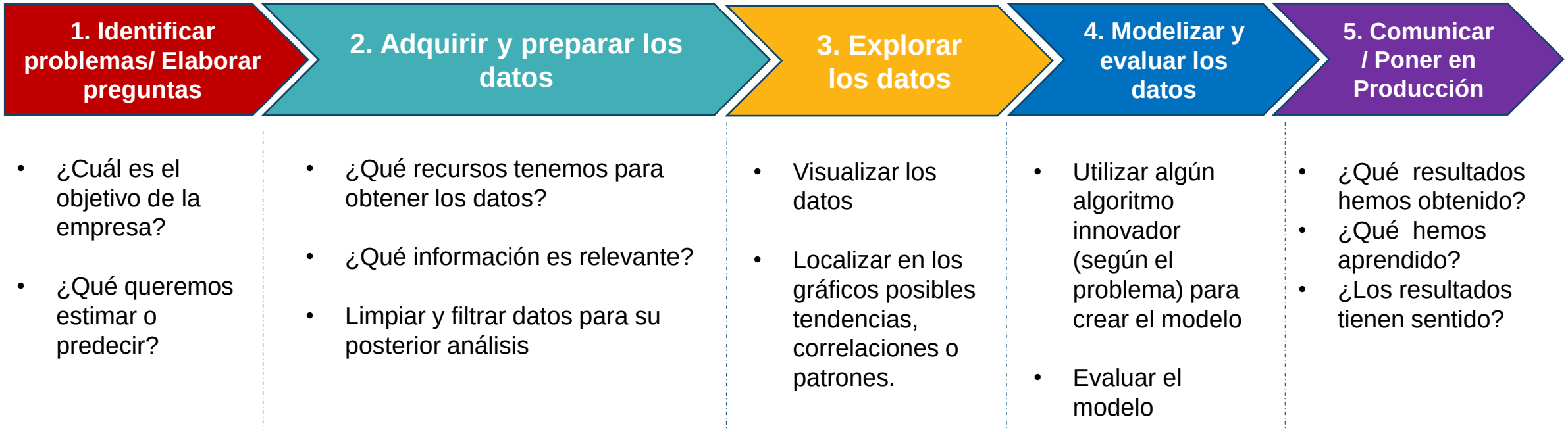
Diferencias entre proyectos de ingeniería de software y ciencia de datos

	Ingeniería de software	Ciencia de los datos
Viabilidad del proyecto	Generalmente se sabe de antemano si un proyecto es ejecutable	Es posible que no se sepa hasta las últimas fases del proyecto.
Enfocar	Entrega de sistemas de software en funcionamiento	Entrega de conocimientos prácticos
Fase más larga	Desarrollo (codificación)	Preparación de datos
Alcance	En gran parte definido por las partes interesadas y los gerentes de producto	Algo definible por las partes interesadas y los gerentes de producto, pero también debe descubrirse en función de lo que descubren los científicos de datos.
Estimación de tareas	El tiempo de finalización de la tarea generalmente se puede estimar	Se desconoce el tiempo necesario para dar muchos pasos.
Seguimiento de progreso	Algo definitivo a través de métricas como la *cantidad de características* o *puntos de historia completos*	Más ambiguo. Ejemplo: haber terminado al 50% con un modelo no significa nada.
Sabiendo que funciona	Mayormente binario. El software funciona según las especificaciones o no (por ejemplo, la interfaz de usuario se carga o no)	Muchos tonos de gris. Dado un modelo, una persona puede decir que está funcionando y otra podría decir que no. Ambos pueden tener razón dado su marco de referencia.

Fuente: https://www.datascience-pm.com/data-science-vs-software-engineering/

2. Etapas de un Proyecto de Ciencia de Datos

Gestión de Proyectos de Ciencia de Datos - Etapas



2. Etapas de un Proyecto de Ciencia de Datos

CIENCIA DE DATOS

Proyecto de Analítica

Identificar problemas/ Elaborar preguntas

Adquirir y preparar los datos

Explorar los datos

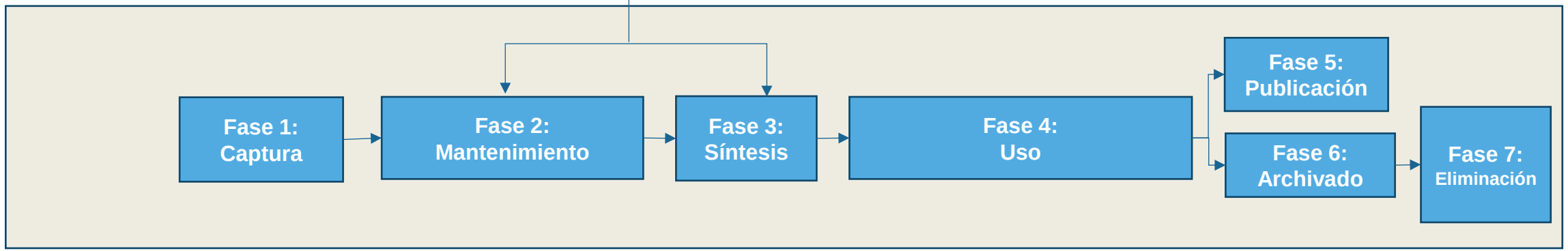
Modelizar y evaluar los datos

Comunicar / Poner en Producción

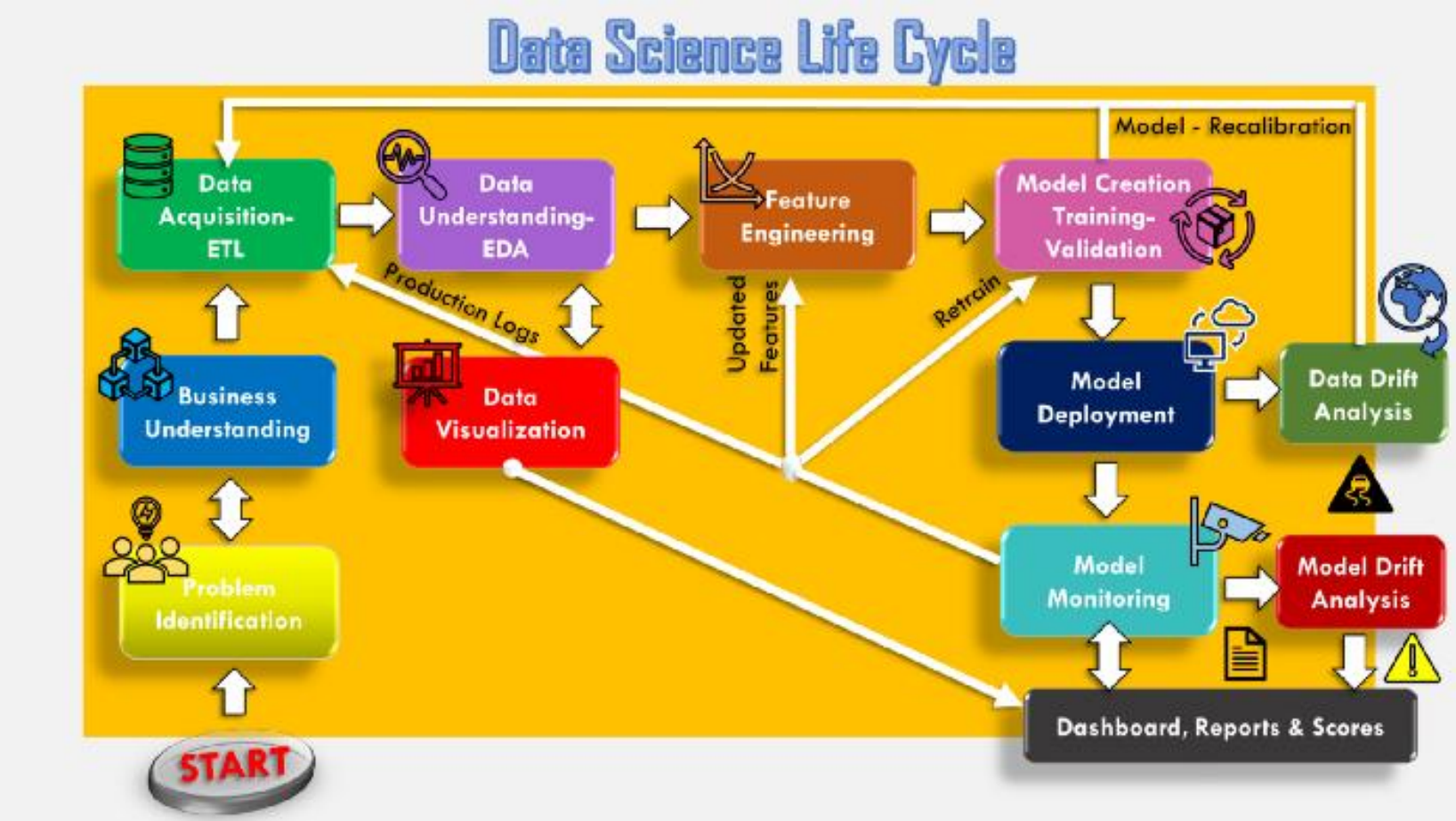
Proyecto de Minería de Datos

Pre-Procesamiento de datos

Ciclo de vida del dato



2. Etapas de un Proyecto de Ciencia de Datos



Fuente: https://towardsdatascience.com/stoend-to-end-data-science-life-cycle-6387523b5afc

2. Etapas de un Proyecto de Ciencia de Datos

EJEMPLO: Etapas del Proceso de Analítica de Datos en la Industria

Enlace: https://youtu.be/ptkPiRe8G30 (Duración: 10:59 min)

3. Metodologías para la gestión de Proyectos

¿Cuáles son las metodologías más utilizadas en la gestión de **Proyectos TI de Ing. de Software?**

❑ La metodología secuencial tradicional

Waterfall, Critical Path Method (CPM) y Critical Chain Project Management (CCPM).

❑ PMI/PMBOK

Establecida por el Project Management Institute. Ésta sigue las cinco fases de la gestión de proyectos descritas en la Guide to the Project Management Body of Knowledge (PMBOK), en español Guía del cuerpo de conocimiento de la gestión de proyectos.

❑ Agile

De ella surgieron otras metodologías: **Scrum**, **Kanban**, **Extreme Programming (XP)** y **Adaptive Project Framework** (APF).

❑ Gestión de cambio (change management)

Planificación de los riesgos y toman el control del cambio cuando se produce. Los métodos más conocidos son: **Event Chain Methodology** y **Extreme Project Management.**

3. Metodologías para la gestión de Proyectos

¿Cuáles son las metodologías más utilizadas en la gestión de **Proyectos TI de Ciencia de Datos?**

❑ CRISP-DM

Cross-Industry Standard Process for Data Mining, es un método probado para orientar los trabajos de minería de datos.
Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
Como modelo del proceso, CRISP-DM ofrece un resumen del ciclo de vida de los proyectos de minería de datos.

❑ SEMMA

Sample, Explore, Modify, Model, and Assess
Es una lista de pasos secuenciales desarrollada por SAS Institute, uno de los mayores productores de software de estadística e inteligencia empresarial. Guía la implementación de aplicaciones de minería de datos .

❑ KDD

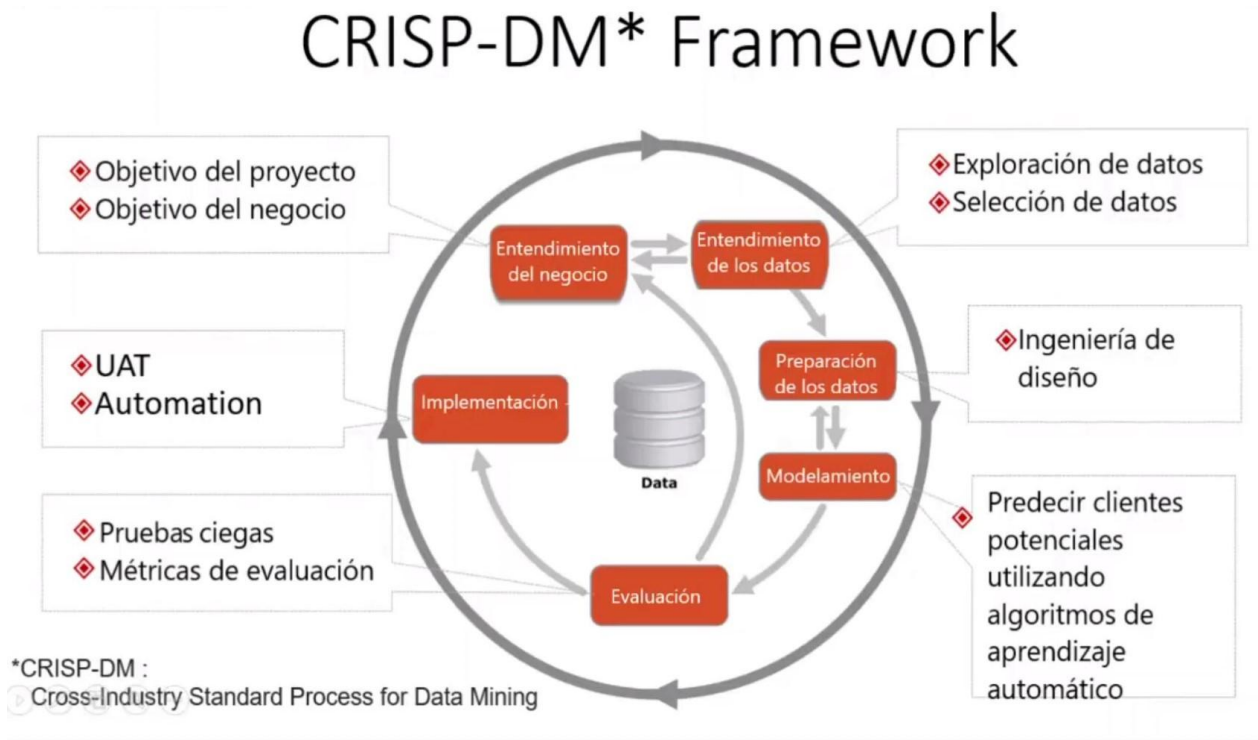
Knowledge discovery in databases (KDD), también llamado "descubrimiento de conocimiento en bases de datos".
Tiene por finalidad interpretar patrones, modelos y un profundo análisis de la información.

❑ Six Sigma

Esta metodología (desarrollada inicialmente por Motorola) promueve mejorar la confiabilidad, reduciendo la variación o los defectos de productos, servicios y procesos. Orientada a la calidad.

NOTA: Todas las metodologías, en general, son comparables y comparten fases comunes.

3. Metodologías para la gestión de Proyectos de Ciencia de Datos

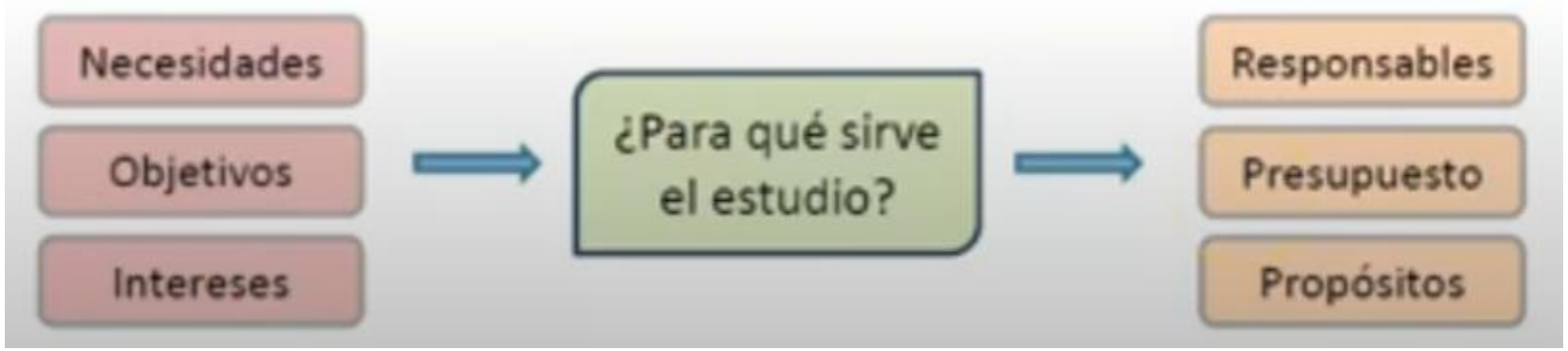


Enlace: https://youtu.be/xtkl_RgWakw (Duración: 7:54 min)

2. Metodología CRISP-DM

01 Entender el Negocio

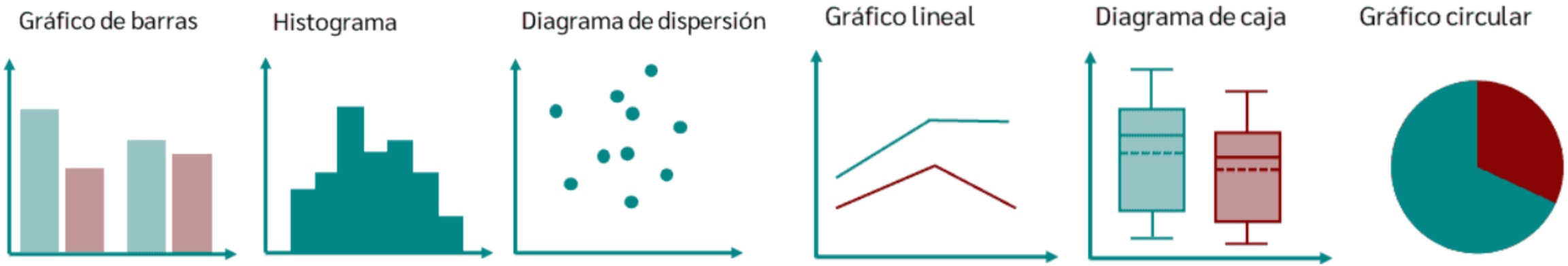
- **Definir los objetivos de negocio** y comprender a fondo lo que se desea lograr.
- **Evaluar la situación** actual o problemática y realizar un análisis de costo beneficio.
- **Establecer los objetivos técnicos** de data mining y cómo se medirá el éxito del proyecto.
- **Desarrollar un plan de proyecto**, seleccionando las tecnologías y herramientas a utilizar.



2. Metodología CRISP-DM

02 Entender los Datos

- **Recopilar los datos iniciales** y cargarlos en herramientas para su análisis.
- **Descripción de los datos** como propiedades, formato, número de registros y campos clave.
- **Exploración de los datos** mediante visualizaciones e identificando relaciones entre ellos.
- **Verificar la calidad de los datos** evaluando datos faltantes, irrelevantes, atípicos.



2. Metodología CRISP-DM

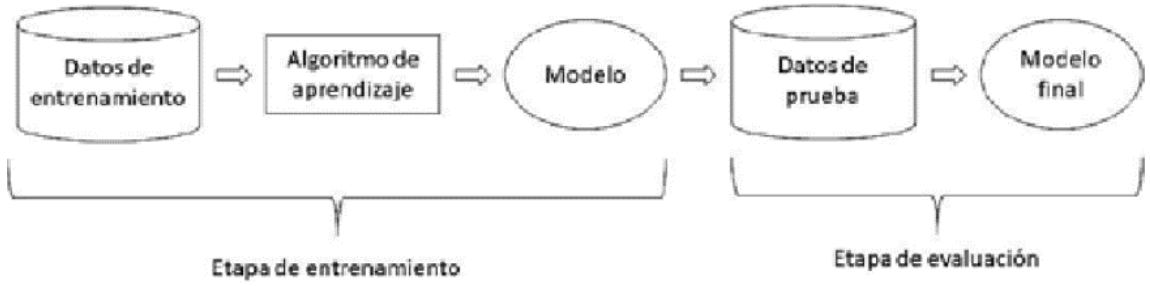
03 Preparar los Datos

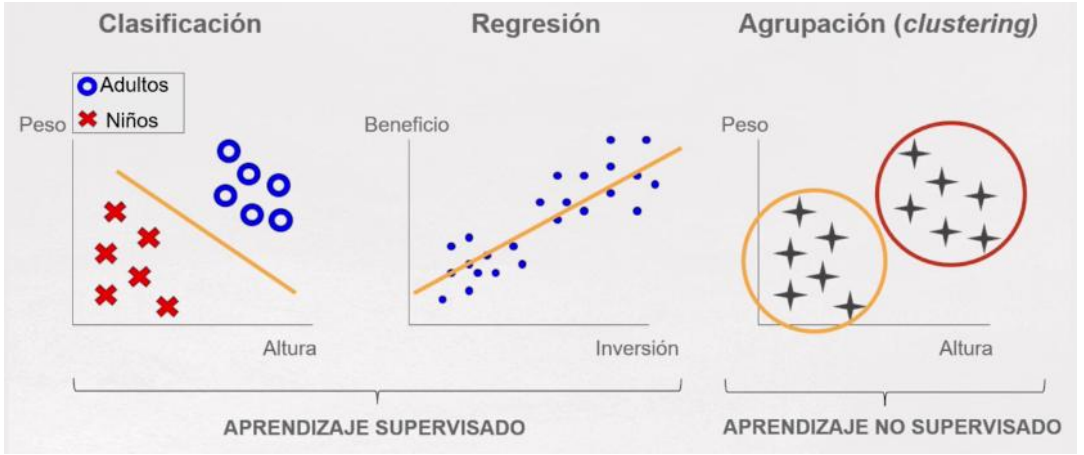
- **Seleccionar los datos** que se utilizarán y documentar los motivos de la inclusión o exclusión.
- **Limpiar los datos** como corregir, imputar o eliminar valores erróneos.
- **Estructurar los datos** mediante visualizaciones e identificando relaciones entre ellos.
- **Integrar los datos** de varias fuentes para crear nuevos conjuntos de datos.
- **Formato de los datos** según sea necesario.

2. Metodología CRISP-DM

04 Crear el modelo

- **Seleccionar las técnicas de modelado** determinando qué algoritmos se aplicarán.
- **Generar el plan de prueba** y los conjuntos de datos para entrenamiento, prueba y validación.
- **Construir el modelo** usando las técnicas de data mining apropiado al objetivo planteado.
- **Evaluar el modelo** probando la calidad y validez de este.

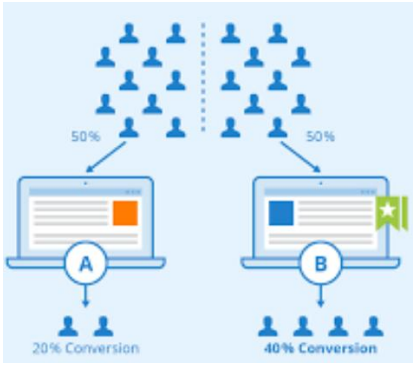




2. Metodología CRISP-DM

05 Evaluar el modelo

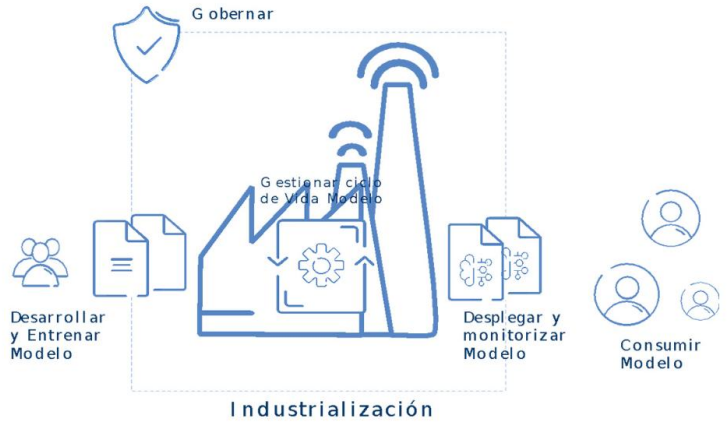
- **Evaluar los resultados** preguntándose si los modelos cumplen con los criterios de éxito.
- **Revisar el trabajo** analizando si se pasó algo por alto y/o se ejecutaron todos los pasos.
- **Determinar los siguientes pasos,** revisar si los resultados son fiables o es aconsejable probar otros modelos. Lista de acciones o planes a desarrollar.



2. Metodología CRISP-DM

06 Desplegar el modelo

- **Desarrollar y documentar** un plan para implementar el modelo.
- **Plan de monitoreo,** ajuste y optimización para evitar problemas durante la fase operativa.
- **Documentación y resumen** del proyecto en un informe final de los resultados del proyecto.
- **Revisión del proyecto** y análisis de retrospectiva sobre lo que salió bien, lo que podría haber sido mejor y cómo mejorar en el futuro.



Conclusiones

1. Un proyecto es un ***Esfuerzo*** para crear o modificar un producto o servicio específico. Es de un ***Trabajo temporal*** con un comienzo y un final claros e involucra varios recursos:
   - Personas
   - Tecnologías
   - Tiempo
   - Dinero
   - Nuevos productos o soluciones únicas

2. Un **Proyecto de Ing. de Software, e**s un **proyecto de tecnología de la información (TI)** definido por una **fecha de inicio** y **fin**, por lo general, con **hitos** y **objetivos específicos** que deben cumplirse durante el ciclo de desarrollo. Para tener éxito en su ejecución, se debe mantener un equilibrio entre estos tres factores:
   - Tiempo
   - Coste
   - Alcance

   Los proyectos de ingeniería de software fluyen a través del ciclo de vida de desarrollo de software (SDLC), que a menudo se representa con cinco o seis fases (cada fase es distinta, por lo que cada una debe completarse antes de pasar a la fase siguiente).
   El **Diagrama de Gantt** nos ayuda a gestionar un proyecto desde el inicio hasta el fin.

3. Las fases para **la Gestión de Proyectos de Ing. de Software** son 6:
   1) **Toma de requerimientos:** Documentación al detalle de todas las solicitudes / requerimientos del cliente
   2) **Análisis:** Elaboración de diagramas / mapas que detallan el análisis funcional de los requerimientos
   3) **Diseño:** Elaboración de documentos, diagramas y especificaciones técnicas dejando por escrito y claramente, todo lo que los programadores deberán desarrollar.
   4) **Construcción:** Se traduce en código todo el diseño técnico de la etapa anterior. Los programadores deben seguir las especificaciones perfectamente detalladas.

Conclusiones

5) **Pruebas:** El cliente valida el producto entregado luego que participa en pruebas manuales que simulan el uso que se hará de la aplicación. El equipo de pruebas por lo general no es el mismo que construye.
6) **Puesta en Producción:** El cliente aprueba el producto entregado. Inicia el uso de la aplicación en “vivo”. Antes de la puesta en producción, por lo general se realizan sesiones de capacitación dirigido a usuarios finales. Entrega de manuales de uso.

4. Un **Proyecto de Ciencia de Datos, e**s también un **proyecto de tecnología de la información (TI)** definido por una **fecha de inicio** y **fin**, pero en el que **se desconoce el tiempo necesario** para dar muchos pasos, por tanto, no se pueden tener **hitos** y **objetivos específicos** que deben cumplirse durante el ciclo de desarrollo. El seguimiento del progreso es mas ambiguo.
5. Existen diferencias marcadas entre un proyecto de Ingeniería de Software y de Ciencia de Datos:
6. Las fases para **la Gestión de Proyectos Ciencia de Datos** son 5:
   1) **Identificar problemas/ Elaborar preguntas**
      - ¿Cuál es el objetivo de la empresa?
      - ¿Qué queremos estimar o predecir?
   2) **Adquirir y preparar los datos:**
      - ¿Qué recursos tenemos para obtener los datos?
      - ¿Qué información es relevante?
      - Limpiar y filtrar datos para su posterior análisis
   3) **Explorar los datos:**
      - Visualizar los datos
      - Localizar en los gráficos posibles tendencias, correlaciones o patrones.
   4) **Modelizar y evaluar los datos:**
      - Utilizar algún algoritmo innovador (según el problema) para crear el modelo
      - Evaluar el modelo

Conclusiones

5) Comunicar / Poner en Producción:

- ¿Qué resultados hemos obtenido?
- ¿Qué hemos aprendido?
- ¿Los resultados tienen sentido?

7. Las metodologías más utilizadas en la gestión de Proyectos TI de Ing. de Software son:

- **La metodología secuencial tradicional:** Waterfall, Critical Path Method (CPM) y Critical Chain Project Management (CCPM).
- **PMI/PMBOK:** Establecida por el Project Management Institute. Ésta sigue las cinco fases de la gestión de proyectos descritas en la Guide to the Project Management Body of Knowledge (PMBOK), en español Guía del cuerpo de conocimiento de la gestión de proyectos.
- **AGILE:** De ella surgieron otras metodologías: **Scrum**, **Kanban**, **Extreme Programming (XP)** y **Adaptive Project Framework** (APF).
- **Gestión de cambio (change management):** Planificación de los riesgos y toman el control del cambio cuando se produce. Los métodos más conocidos son: **Event Chain Methodology** y **Extreme Project Management.**

8. Las metodologías más utilizadas en la gestión de Proyectos TI de Ciencia de Datos son:

- **CRISP-DM: Cross-Industry Standard Process for Data Mining**, es un método probado para orientar sus trabajos de minería de datos.Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

- **SEMMA: Sample, Explore, Modify, Model, and Assess** Es una lista de pasos secuenciales desarrollada por SAS Institute, uno de los mayores productores de software de estadística e inteligencia empresarial. Guía la implementación de aplicaciones de minería de datos .

Conclusiones

- **KDD: Knowledge discovery in databases** (KDD), también llamado "descubrimiento de conocimiento en bases de datos". Tiene por finalidad interpretar patrones, modelos y un profundo análisis de la información.
- **Six Sigma:** Esta metodología (desarrollada inicialmente por Motorola) es mejorar la confiabilidad, reduciendo la variación o los defectos de productos, servicios y procesos. Orientada a la calidad.

9. Factores que contribuyen al fracaso de los Proyectos de Ciencia de Datos
    - **¿Por qué el 87% de los proyectos de ciencia de datos fracasan?**
        1. CREER EN UN FALSO ÉXITO ASEGURADO
        2. FALTA DE ACCESO A LOS DATOS
        3. FALTA DE COLABORACION

10. Factores que contribuyen al éxito de los Proyectos de Ciencia de Datos
    - **¿Por qué sólo un 13% de los proyectos de Ciencia de Datos realmente llegan a producción?**

      Porque sus claves de éxito son:
        1. EDUCAR A LOS LÍDERES EMPRESARIALES EN TODA LA ORGANIZACIÓN
        2. MANTENER SIMPLE EL PROCESO
        3. **TENER OBJETIVOS Y RECURSOS CLAROS**
            - Eligen un pequeño proyecto para empezar (no intentar hervir el océano), en el cual se pueda demostrar el progreso.
            - Trabajan con el equipo adecuado.
            - Se apoyarse en proveedores de soluciones analíticas con experiencia.

PREGUNTAS

Dudas y opiniones