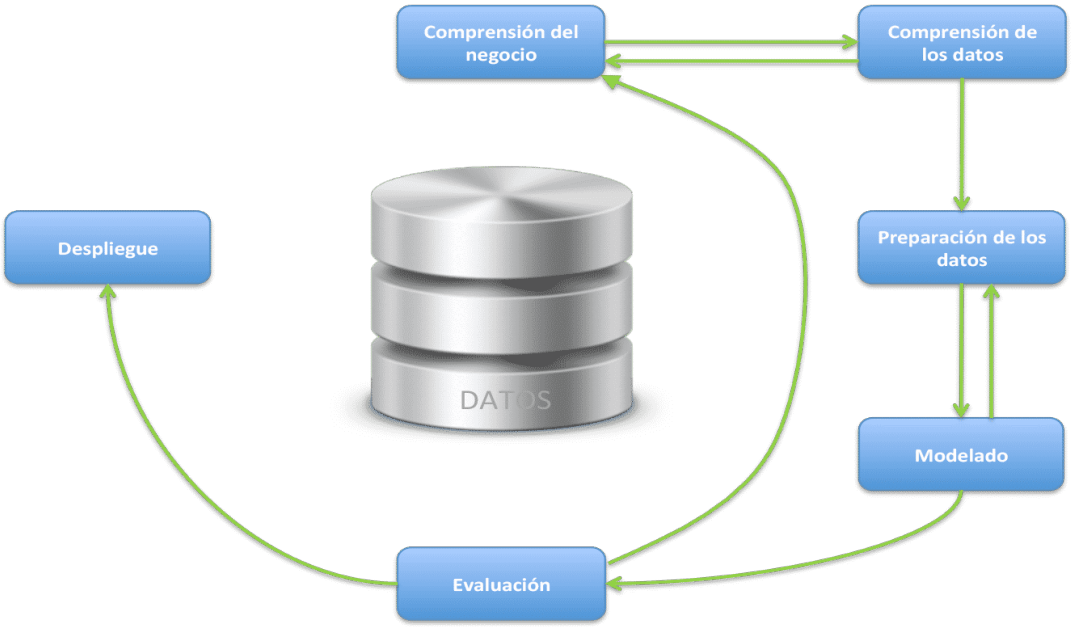
Metodología CRISP-DM

Por **Daniel Álvarez Gil** - 14 enero, 2021



Seguir una buena metodología de trabajo nos asegura obtener unos buenos resultados. Como en cualquier área profesional, la minería de datos tiene su propia guía para afrontar proyectos reales.

En este artículo se explica cómo sería el ciclo de vida de un proyecto de minería de datos llevando a cabo una metodología CRISP-DM.

Tabla de contenidos

1. Introducción

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) integra todas las tareas necesarias en los proyectos de minería de datos, desde la fase de comprensión del problema hasta la puesta en producción de sistemas automatizados analíticos, predictivos y/o prospectivos. Esta metodología se puede aplicar a una gran variedad de problemas tales como encontrar perfiles de clientes fraudulentos, estimar la probabilidad de que los clientes de una empresa se vayan a la competencia o también determinar patrones de compra para luego, recomendar productos de interés.

CRISP-DM está compuesta por seis fases, las cuales dependen entre sí tanto en forma secuencial como cíclica, pudiendo encontrarse interacciones que permitan mejorar la aproximación obtenida en otras fases anteriores. Las fases son las siguientes:

- *Comprensión del negocio.*
- *Comprensión de los datos.*
- *Preparación de los datos.*
- *Modelado.*
- *Evaluación.*
- *Despliegue.*

2. CRISP-DM: Comprensión del negocio

Esta es la primera fase por donde debe comenzar todo proyecto de minería de datos, comprendiendo en profundidad el problema que se quiere resolver y estableciendo los requisitos y objetivos del proyecto desde una perspectiva empresarial para luego trasladarlos a objetivos técnicos y a un plan de proyecto.

En primer lugar, se establece cuáles serán los criterios para medir el éxito en el proyecto, ya sean de tipo cualitativo o cuantitativo. Después, se realiza una evaluación de la situación actual determinando los antecedentes y requisitos del problema, tanto en términos de negocio como de minería de datos. Y por último, se realiza un plan de proyecto

donde se tiene en cuenta qué pasos se deben seguir y qué procedimientos se emplearán en cada uno de ellos.

3. CRISP-DM: Comprensión de los datos

Es esta fase se lleva a cabo la recolección y exploración inicial de los datos, con el objetivo de establecer un primer contacto con el problema. Esta fase suele ser crítica en el proyecto dado que un mal entendimiento de los datos tiene como consecuencia un aumento en el tiempo global del proyecto y también reduce las garantías de éxito.

Para llevar a cabo esta fase hay que desarrollar una serie de tareas. La primera es recolectar datos iniciales y adaptarlos a las necesidades del proyecto para su posterior procesamiento. Luego se deben describir formalmente los datos obtenidos: número de instancias (filas) y atributos (columnas), el significado de los atributos y una descripción rigurosa del formato de los datos. Después se exploran los datos aplicando técnicas básicas de estadística descriptiva que revelan propiedades de estos. Por último, se lleva a cabo una verificación de los datos para determinar su consistencia, la cantidad y distribución de los valores nulos o valores fuera de rango que puedan provocar ruido en el modelado posterior.

4. CRISP-DM: Preparación de los datos

La fase de preparación trata de seleccionar, limpiar y generar conjuntos de datos correctos, organizados y preparados para la fase de modelado. Esta es una fase sumamente crítica en un proyecto de minería de datos. Los errores en los datos que se pasan por alto y que no son resueltos en esta fase se trasladan hasta la fase de modelado, lo que genera una reducción en la exactitud de los modelos o incluso, es posible entregar al cliente resultados basados en datos que aún contienen errores no detectados.

Por esta razón, esta fase es crucial y generalmente demanda siempre el mayor esfuerzo y tiempo del proyecto, aproximadamente un 75% del tiempo total.

5. CRISP-DM: Modelado

En esta fase se lleva a cabo la creación de modelos de conocimiento a partir de los datos suministrados desde la fase anterior. Estos modelos de conocimiento pueden ser de distintos tipos, por ejemplo, se pueden crear modelos de clasificación o regresión con el objetivo de estimar o inferir el valor de una determinada variable.

Para afrontar esta fase hay que seguir una serie de pautas que nos ayudarán a obtener mejores resultados. Aunque parezca obvio, el primer paso es **seleccionar los algoritmos de modelado** más apropiados al problema. Posteriormente se **genera un plan de prueba**, donde configuramos los valores de los parámetros que se usarán para los algoritmos de aprendizaje automático (machine learning), ya que muchos de estos pueden ser configurados para determinar las características del modelo que se generará. También, se determinan las métricas de evaluación que se calcularán para evaluar la bondad de los modelos. Luego **construimos los modelos**, ejecutando los algoritmos seleccionados sobre los datos preparados con el fin de generar uno o más modelos y calculando las métricas. Y acabamos **evaluando los resultados**, donde se analizan las métricas de evaluación obtenidas con el fin de conocer la bondad de los modelos generados y garantizar que cumplan con los criterios de éxito definidos al inicio del proyecto.

6. CRISP-DM: Evaluación

Si los modelos obtenidos cumplen con las expectativas de negocio, se procede a la explotación del modelo. Si no, se evalúa en esta fase si se procede a iterar nuevamente sobre los pasos anteriores con el objetivo de encontrar nuevos resultados.

Para ello, se realiza una evaluación formal de los resultados obtenidos en las fases anteriores del proyecto, teniendo en cuenta los criterios de éxito de negocio y explicando las causas que provocaron los grados de éxitos alcanzados y revisando todo el proceso llevado a cabo con el fin de

encontrar errores que puedan afectar al éxito.

7. CRISP-DM: Despliegue

¡Es hora de poner el modelo a trabajar! En esta última fase es donde se definen las estrategias para su implementación, monitorización y mantenimiento de los modelos. Esto nos ayudará a observar cualquier comportamiento irregular del sistema y corregirlo de forma que no provoque deficiencias en el servicio del cliente.

Por último se lleva a cabo una revisión final de todo el proceso llevado a cabo en el proyecto, evaluando las acciones realizadas de forma correcta e incorrecta, con el fin de enumerar las lecciones aprendidas en el proyecto.

Enlaces y referencias

- Guía oficial de la metodología CRISP-DM

Daniel Álvarez Gil

Desarrollador Software. Graduado en Ingeniería de Computadores y haciendo un máster en Inteligencia Artificial. Puedes encontrarme en Autentia: Ofrecemos servicios de soporte a desarrollo, factoría y formación. Somos expertos en Java/Java EE