CRISP-DM: La metodología para poner orden en los proyectos

Las técnicas de Data Science o Data Analytics, que tanto interés despiertan hoy en día, en realidad surgieron en la década de los 90, cuando se usaba el término KDD (Knowledge Discovery in Databases) para referirse al (amplio) concepto de hallar conocimiento en los datos. En un intento de normalización de este proceso de descubrimiento de conocimiento, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM (Cross Industry Standard Process for Data Mining) y **SEMMA** (Sample, Explore, Modify, Model, and Assess). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase.

la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente (en encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, 4 veces más que SEMMA), y es la que nosotros empleamos en nuestros proyectos. En este post vamos a realizar una introducción a la metodología CRISP-DM, sus objetivos, fases de las que consta y tareas contenidas en cada fase, resumido de la guía del consorcio de empresas que propuso la metodología (Chapman et al., 2000).

> **Azevedo, Ana; Zantos, Manuel Filipe. KDD, SEMMA and CRISP-DM: a parallel overview. 2008.**

> **Chapman, Pete (NCR); Clinton, Julian (SPSS); Kerber, Randy (NCR); Khabaza, Thomas (SPSS); Reinartz, Thomas (DaimlerChrysler); Shearer, Colin (SPSS); Wirth, Rüdiger (DaimlerChrysler). Step-by-step data mining guide. 2000.**

Introducción

CRISP-DM (Cross Industry Standard Process for Data Mining) proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software. El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción no es posible identificar todas las relaciones; las relaciones podrían existir entre cualquier tarea según los objetivos, el contexto, y el interés del usuario sobre los datos.

La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

El ciclo de vida del proyecto de minería de datos consiste en seis fases mostradas en la figura siguiente.

que hacer después. Las flechas indican las dependencias más importantes y frecuentes.

El círculo externo en la figura simboliza la naturaleza cíclica de los proyectos de análisis de datos. El proyecto no se termina una vez que la solución se despliega. La información descubierta durante el proceso y la solución desplegada pueden producir nuevas iteraciones del modelo. Los procesos de análisis subsecuentes se beneficiarán de las experiencias previas.

A continuación vamos a describir brevemente cada una de las fases.

Fase I. Business Understanding. Definición de necesidades del cliente (comprensión del negocio)

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Fase II. Data Understanding. Estudio y comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Fase III. Data Preparation. Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Fase IV. Modeling. Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos

Fase V. Evaluation. Evaluación (obtención de resultados)

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.
Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Fase VI. Deployment. Despliegue (puesta en producción)

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

Conclusión

A modo de conclusión-resumen, la siguiente figura presenta una guía visual de todas las fases, listando las tareas a realizar en cada fase, así como las conexiones entre ellas y las iteraciones que pueden llevarse a cabo. Esta figura ha sido tomada de "**A visual guide to CRISP-DM methodology**".

El consorcio que planteó CRISP-DM se disolvió hace unos años. Pese a ello, CRISP-DM es la metodología que se utiliza de facto, de una forma u otra, en los proyectos de análisis de datos que se pretendan abordar con seriedad y asegurando la calidad de los resultados.

Recientemente ha resurgido una nueva iniciativa, **crisp-dm.eu**, sin mucho impacto hasta el momento.

En 2015, IBM Corporation, uno de los impulsores tradicionales de CRISP-DM, planteó una nueva metodología methodology llamada **Analytics Solutions Unified Method for Data Mining/Predictive Analytics** (ASUM-DM) que extiende CRISP-DM, y es parte de la metodología general ASUM (Analytics

Nuestro equipo de profesionales puede abordar proyectos de Data Analytics en cualquier escenario complejo con las máximas garantías de éxito, aplicando la metodología CRISP-DM de forma seria y consistente, aunque siendo pragmáticos y en combinación con metodologías ágiles. Si tiene cualquier pregunta o necesidad en estas áreas, por favor, no dude en contactar con nosotros, que estaremos encantados de ayudarle.

Country

España

DKCs

Data Science

Tags

Data Science

Volver a la lista Siguiente artículo >

Data & AI
Design
Scalable Platforms
Studios

Talent
Marketing Tranformation
Media
Ventures

Energía
Servicios financieros
Salud
Industria

Museos
Administraciones Públicas
Retail
Telecomunicaciones

Noticias
Uniq
Futurizable

Newsletter

It can be done

Follow us