GUÍA PRÁCTICA CRISP-DM

EQUIPO DE ANALÍTICA

2021

HISTORIAL DE REVISIONES

VERSIÓN	FECHA	CAMBIOS INTRODUCIDOS	AUTORES
0.1	18/04/2021	Creación inicial del documento	Equipo ACDT
0.2	08/06/2021	Complementación de información y diseño del documento	Equipo ACDT
0.3	22/06/2021	Actualización de tips explícitos para el usuario.	Equipo ACDT
0.4	10/08/2021	Adición de pautas generadas en la revisión inicial.	Equipo ACDT
1.0	03/11/2021	Ajustes etapas 4, 5, y 6	Equipo ACDT

¿SABIAS QUE?

- Una de las metodologías dominantes en el campo de la minería de datos es la de CRISP-DM y su objetivo primordial es ser neutral con respecto a herramientas a través de todo el proceso de minería de datos y elimina los procesos complejos y costosos de las tareas simples de minería.

- Las técnicas de Data Science o Data Analytics, surgieron en la década de los 90, cuando se establece el concepto de hallar conocimiento en los datos.

- 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.

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

  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.

  - 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.

METODOLOGÍA CRISP-DM

- El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases.
- La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.
- El modelo de CRISP-DM es flexible y se puede personalizar fácilmente.
- La secuencia de las fases no es rígida: se permite movimiento hacia adelante y hacia atrás entre diferentes fases. El resultado de cada fase determina qué fase, o qué tarea particular de una fase, hay 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.

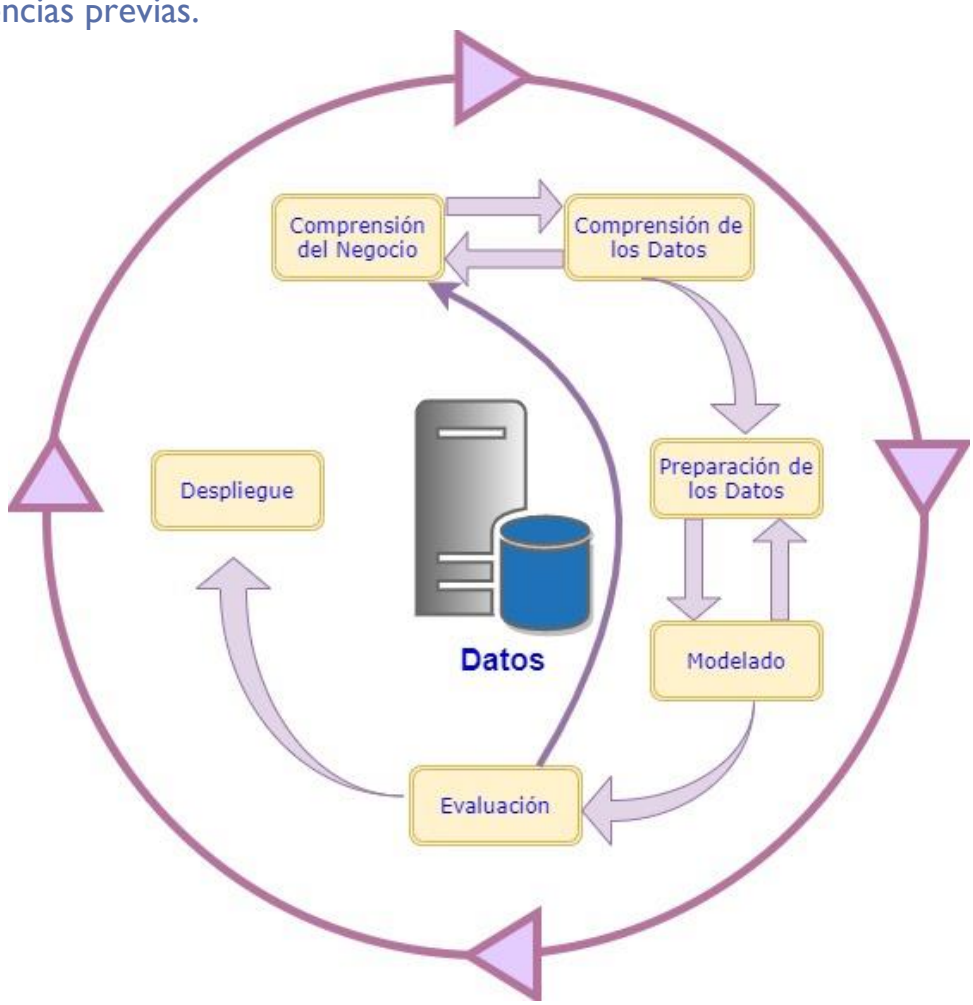

Ilustración 1.Modelo de referencia de las fases de CRISP-DM – Traducción propia

Mapeo de modelos genéricos a modelos especializados

Contexto de minería de datos

El contexto de minería de datos impulsa el mapeo entre el nivel genérico y especializado en CRISP-DM. Actualmente, distinguimos entre cuatro dimensiones diferentes de contextos de minería de datos:

- El dominio de aplicación es el área específica en la que se lleva a cabo el proyecto de minería de datos.
- El tipo de problema de minería de datos describe la(s) clase(s) específica(s) de objetivo(s) que trata el proyecto de minería de datos.
- El aspecto técnico cubre problemas específicos en la minería de datos que describen diferentes desafíos (técnicos) que generalmente ocurren durante la minería de datos.
- La dimensión de herramienta y técnica especifica qué herramientas y / o técnicas de minería de datos se aplican durante el proyecto de minería de datos.

FASES

Fase I.	Comprensión del negocio (Business Understanding)
Fase II.	Comprensión de los Datos (Data Understanding)
Fase III.	Análisis de los datos y selección de características (Data Preparation.)
Fase IV.	Modelado (Modeling)
Fase V.	Evaluación (Evaluation)
Fase VI.	Despliegue (Deployment)

FASE I.

COMPRENSIÓN DEL RETO

Definición de necesidades del usuario

EN ESTA FASE:

- Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto.
- La idea es convertir 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.

Tips:

- *Comprender a fondo, lo que el usuario realmente quiere lograr.*
- *Enumerar los riesgos o eventos que podrían retrasar el proyecto o hacer que falle.*
- *Compilar un glosario de terminología relevante para el proyecto.*
- *Definir los criterios para un resultado exitoso del proyecto.*

FASE II.

COMPRENSIÓN DE LOS DATOS

Fase 2

Estudio y comprensión de los datos

EN ESTA FASE:

- La fase comienza con la colección de datos inicial.
- Identificar los problemas de calidad.
- Tener el contexto preliminar sobre los que significan datos.
- Descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

- *Incluir la carga de datos y la integración de fuentes.*
- *Registrar los problemas encontrados y las resoluciones logradas.*
- *Describir los datos que se han conseguido, incluido el formato de los datos, la cantidad de datos, las identidades de los campos, etc.*
- *Verificar y Examinar la calidad de los datos.*

FASE III.

ANÁLISIS DE LOS DATOS

Análisis de los datos y selección de características.

EN ESTA FASE:

- Esta fase cubre todas las actividades necesarias para construir el conjunto final de datos.
- Se seleccionan 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.
- También la transformación y la limpieza de datos para las herramientas que modelan.

Tips:

- *Decidir los datos que se utilizarán para el análisis. Los criterios incluyen la relevancia para los objetivos de minería de datos, la calidad y las limitaciones técnicas, como los límites en el volumen de datos o los tipos de datos.*
- *Considerar las transformaciones y las agregaciones de los datos con fines de limpieza y el posible impacto en los resultados del análisis.*
- *Realizar el ajuste en el formato de datos o transformaciones de formato de algunas columnas del set de datos.*

FASE IV. MODELADO.

Modelado

EN ESTA FASE:

- Se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor).
- Se calibran los parámetros a valores óptimos.
- Algunas técnicas tienen requerimientos específicos sobre la forma de los datos.
- Casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

- *Documentar la técnica de modelado real que se va a utilizar, por ejemplo, la construcción de un árbol de decisiones con 5.0 o la generación de redes neuronales con retropropagación.*
- *Describir el plan previsto para entrenar, probar y evaluar los modelos.*
- *Ejecutar la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos.*

FASE V. EVALUACIÓN.

Obtención de resultados

EN ESTA FASE:

- En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la 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.
- Es necesario comparar el modelo obtenido con los objetivos de negocio.
- También se analiza 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 V

Tips:

- *Evaluar el grado en que el modelo cumple con los objetivos comerciales y busca determinar si existe alguna razón comercial por la cual este modelo es deficiente.*
- *Determinar dependiendo de los resultados de la evaluación y la revisión del proceso, el equipo decide si finaliza este proyecto y pasa a la implementación, inicia más iteraciones o configurar nuevos proyectos de minería de datos.*

FASE VI. DESPLIEGUE.

Puesta en producción

EN ESTA FASE:

- Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el usuario 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 entidad.

Tips:

- *Tomar los resultados de la evaluación y determinar una estrategia para la implementación.*
- *Establecer una estrategia de monitoreo y mantenimiento, incluyendo los pasos necesarios y cómo llevarlos a cabo.*
- *Revisar el proyecto. Evaluar qué salió bien y qué salió mal para mejorar.*
- *Realizar una presentación final y completa de los resultados de la minería de datos.*

FUENTES DE REFERENCÍA

- CRISP-DM 1.0 Step-by-step data mining guide
- Guía de CRISP_DM de IBM SPSS Modeler - https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview
- CRISP-DM: La metodología para poner orden en los proyectos- https://www.sngular.com/es/data-science-crisp-dm-metodologia/
- Metodología Analítica de Datos - https://ideca.gov.co/sites/default/files/MetodologiaAnaliticaDatos.pdf
- Metodologías aplicadas al proceso de Minería de Datos- https://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf