Portafolio de evidencias

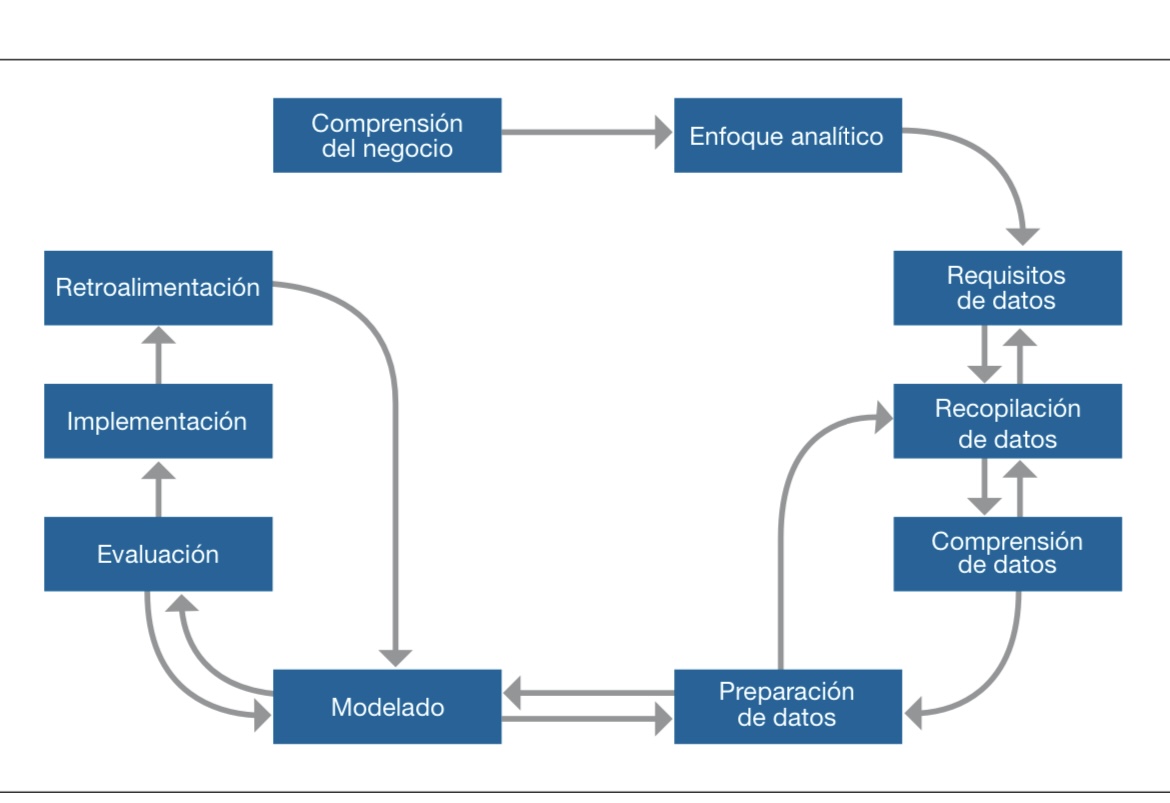
Emely Berrocal

Metodología Fundamental para la Ciencia de Datos

¿Qué es una metodología?

Una metodología es una estrategia general que sirve de guía para los procesos y actividades que están dentro de un dominio determinado. La metodología no depende de tecnologías ni herramientas específicas, ni es un conjunto de técnicas o recetas. Más bien, la metodología proporciona al científico de datos un marco sobre cómo proceder con los métodos, procesos y argumentos que se utilizarán para obtener respuestas o resultados.

**Fases:**



Etapa 1: Comprensión del negocio

Todos los proyectos comienzan con la comprensión del negocio. Los promotores de negocios que necesitan la solución analítica desempeñan el papel más importante en esta etapa, al definir el problema, los objetivos del proyecto y los requisitos de la solución desde una perspectiva empresarial. Esta primera etapa sienta las bases para que el problema empresarial sea resuelto con éxito.

Etapa 2: Enfoque analítico

Cuando el problema empresarial se haya establecido claramente, el científico de datos podrá definir el enfoque analítico para resolver el problema. Esta etapa implica expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático, para que la organización pueda identificar las más adecuadas para el resultado deseado.

Etapa 3: Requisitos de datos

El enfoque analítico elegido determina los requisitos de datos. Más concretamente, los métodos analíticos a utilizar requieren de determinados contenidos de datos, formatos y representaciones, orientados por el conocimiento en el dominio.

Etapa 4: Recopilación de datos

En la etapa inicial de recopilación de datos, los científicos de datos identifican y reúnen los recursos de datos disponibles (estructurados, no estructurados y semiestructurados) y relevantes para el dominio del problema. Por lo general, deben elegir si realizan inversiones adicionales para obtener elementos informativos menos accesibles. Lo mejor puede ser aplazar la decisión de inversión hasta que se sepa más sobre los datos y el modelo.

Etapa 5: Comprensión de datos

Después de la recopilación de datos inicial, los científicos de datos suelen utilizar estadísticas descriptivas y técnicas de visualización para comprender el contenido de los datos, evaluar su calidad y descubrir insights iniciales sobre ellos. Para llenar los huecos es posible que sea necesario volver a recopilar datos.

Etapa 6: Preparación de datos

Esta etapa abarca todas las actividades para construir el conjunto de datos que se utilizará en la subsiguiente etapa de modelado. Entre las actividades de preparación de datos están la limpieza de datos (tratar con valores no válidos o que faltan, eliminar duplicados y dar un formato adecuado), combinar datos de múltiples fuentes (archivos, tablas y plataformas) y transformar los datos en variables más útiles.

Etapa 7: Modelado

La etapa de modelado utiliza la primera versión del conjunto de datos preparado y se enfoca en desarrollar modelos predictivos o descriptivos según el enfoque analítico previamente definido. En los modelos predictivos, los científicos de datos utilizan un conjunto de capacitación (datos históricos en los que se conoce el resultado de interés) para construir el modelo.

Etapa 8: Evaluación

Durante el desarrollo del modelo y antes de su implementación, el científico de datos evalúa el modelo para comprender su calidad y garantizar que aborda el problema empresarial de manera adecuada y completa. La evaluación del modelo implica el cálculo de varias medidas de diagnóstico y de otros resultados, como tablas y gráficos, lo que permite al científico de datos interpretar la calidad y la eficacia del modelo en la resolución del problema. Para los modelos predictivos, los científicos de datos usan un conjunto de pruebas, que es independiente del conjunto de capacitación, pero sigue la misma distribución de probabilidad y tiene un resultado conocido. El conjunto de pruebas se utiliza para evaluar el modelo para ajustarlo según las necesidades. A veces, el modelo final también se aplica a un conjunto de validación para realizar una evaluación final.

Etapa 9: Implementación

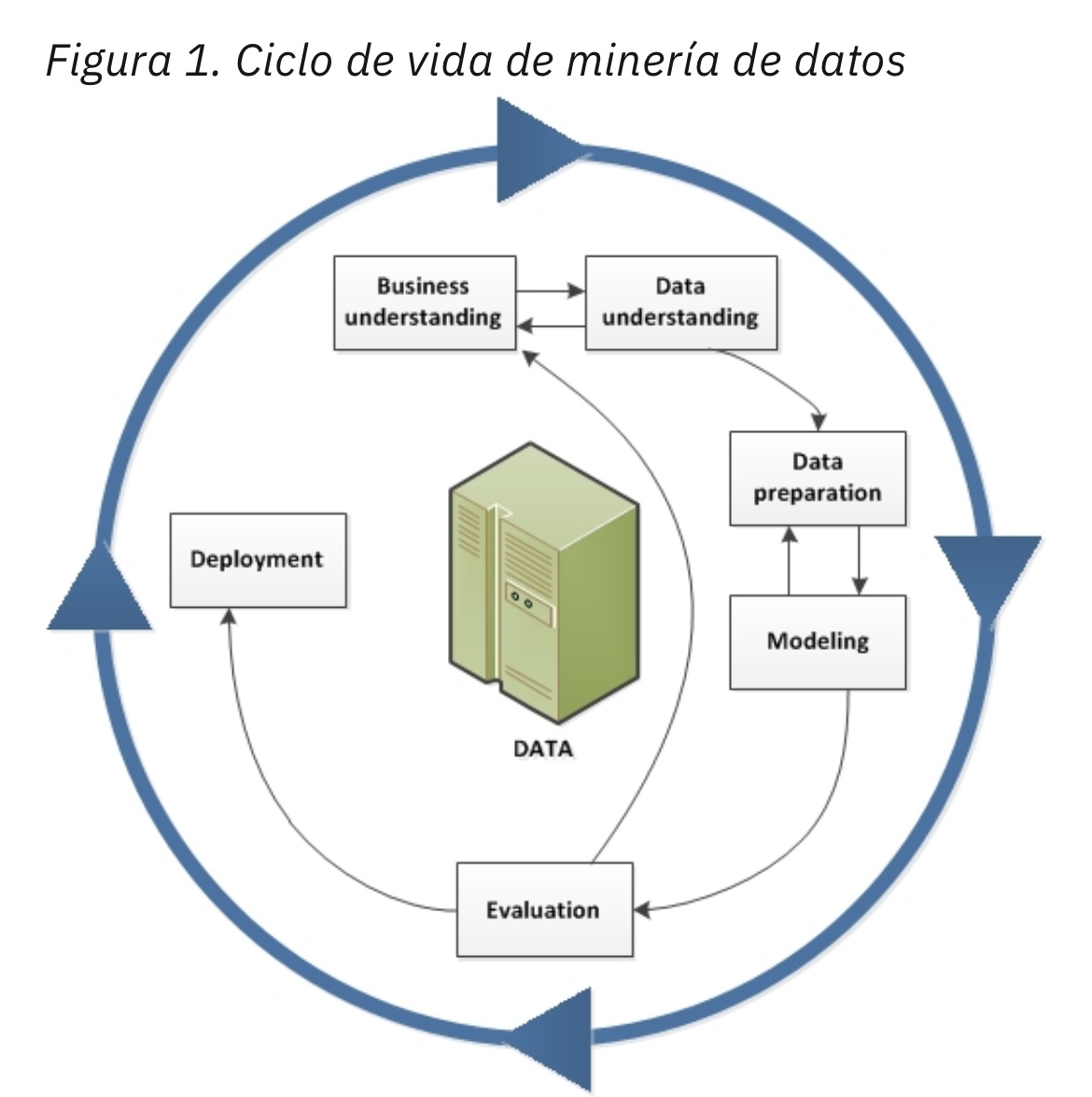
Cuando el modelo satisfactorio ha sido desarrollado y aprobado por los promotores del negocio, se implementa en el entorno de producción o en un entorno de pruebas comparable. Por lo general, se implementa de forma limitada hasta que su rendimiento se haya evaluado completamente.

Etapa 10: Retroalimentación

Al recopilar los resultados del modelo implementado, la organización obtiene retroalimentación sobre el rendimiento del modelo y su impacto en el entorno en el que se implementó. Por ejemplo, la retroalimentación puede ser en forma de porcentajes de respuesta a una campaña promocional dirigida a un grupo de clientes que ha sido identificado por el modelo como respondedores de alto potencial. Los científicos de datos pueden analizar esta retroalimentación para ajustar el modelo para mejorar su precisión y utilidad.

Conceptos básicos de ayuda de CRISP-DM

CRISP-DM, que son las siglas de 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.

**BADIR:**

**Step 1: Business Question**

IStart with the traditional What, Who, Where, When, Why and How questions to help identify the problem in its context. Understanding the context, the impacted segment and potential reasons as understood by the business may provide a quicker path to resolving the problem.

**Step 2: Analysis Plan**

The Analysis Plan step includes 5 parts; analysis goals, hypothesis, methodology, data specification and project plan as shown below in the diagram from BEGD.

**Analysis Goals**

Business problems and questions can be large in scope so narrowing the objectives or breaking it into parts, where each has its own set of analysis goals is important.  These goals should follow the S.M.A.R.T methodology – Specific, Measurable, Attainable, Relevant and Time-bound.

In most types of business analytics we start with what we have experienced and know, then brainstorm some hypothesis (educated guesses to cause and effect related to the Business Question and Analysis Goals) that we test and confirm or reject based on observations or data.

**Methodology**

Some common analytics methodologies are used to help solve business problems.

**Data Specification**

The data specification describes the data that will be needed to prove or disprove the hypothesis and support the chosen methodology.  It’s important to only collect relevant data. The keys are to determine the granularity of the data, weekly, monthly, geography, etc. A data field should serve as a unique ID.

**Project Plan**

The project plan will tie together all of the information gained and serve as the road map for the analysis execution and steps that follow.

**Step 2: Data Collection**

* Data pull which is the collection of the data following the data specification from the previous step.
* Data cleansing and validation to avoid the GIGO (garbage in, garbage out) syndrome the data needs to cleaned for usability and validated for accuracy.

**Step 4: Derive Insights**

* Review patterns: It helps validate patterns in the data if there is a real problem and if there are unusual patterns in key variables.
* Prove or disprove hypotheses: Look at each hypothesis, one at a time, and examine the relevant data needed to prove or disprove each one. This will help eliminate some hypotheses and identify the ones on which you should focus your energies.
* Present findings: Finally, present your findings in terms of quantified impact to guide prioritization of the hypotheses for analysis.

**Step 5: Recommendations**

1. Engage the audience by presenting a short, concise, insightful set of recommendations without getting bogged down in detail.
2. Be perceived as an effective business partner by presenting credible recommendations.
3. Drive the audience towards actions that create impact by solving the business problem

