Aclaración: Usaremos la estructura y los ejemplos del Capítulo 5 de "A Friendly Guide to Data Science" de Kelly P. Vincent para explicar en detalle qué es CRISP-DM, para qué sirve y cómo funciona cada uno de sus pasos.

---

# ¿Qué es CRISP-DM según Kelly P. Vincent?

Como bien señala el libro, aunque los proyectos de análisis de datos pueden variar, la mayoría sigue un proceso general o un **ciclo de vida**. **CRISP-DM** es el nombre de la metodología más popular y estandarizada para gestionar este ciclo de vida.

Es una **hoja de ruta** que guía tanto a los analistas de datos como a los científicos de datos a través de un proyecto, asegurando que se cubran todas las etapas importantes, desde entender el problema de negocio hasta presentar los resultados.

Lo más importante que destaca el autor, y que se ve claramente en el diagrama del libro, es que **CRISP-DM no es un proceso lineal, sino un ciclo altamente iterativo**. No es una lista de tareas que se tachan una vez y se olvidan. Constantemente estarás saltando de una fase a otra, hacia adelante y hacia atrás, a medida que aprendes más sobre el problema y los datos.

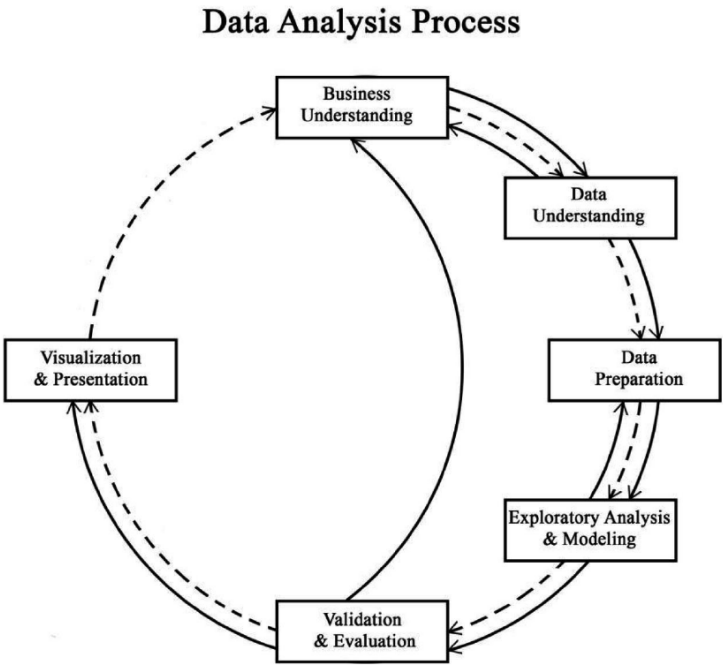


Figura 5-3. El proceso de análisis de datos CRISP-DM. El diagrama muestra un ciclo que conecta seis fases: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Análisis Exploratorio y Modelado, Validación y Evaluación, y Visualización y Presentación. Las flechas indican que el proceso es iterativo, con posibles saltos entre fases no consecutivas.

---

## Los 6 Pasos de CRISP-DM Explicados con el Ejemplo del Libro

El autor nos guía a través de las seis fases usando un ejemplo práctico: **un equipo de fútbol americano que quiere saber a qué jugadores defensivos debería traspasar.**

### 1. Comprensión del Negocio (Business Understanding)

- **¿De qué se trata?**: Es el punto de partida y el más crucial. Consiste en entender qué quiere lograr tu "cliente" (en este caso, la directiva del equipo) y traducir ese objetivo de negocio en preguntas de investigación claras.

- **Ejemplo del libro**:

- **Objetivo de Negocio**: "Decidir a qué jugadores defensivos traspasar".

- **Preguntas de Investigación**: Para lograr eso, necesitamos evaluar su rendimiento. Esto implica definir qué métricas son importantes. Hablas con los entrenadores y directivos y, juntos, definen que las métricas clave son: **placajes (*tackles*), capturas (*sacks*), intercepciones y balones sueltos (*fumbles*)**.

## 2. Comprensión de los Datos (Data Understanding)

- **¿De qué se trata?**: Una vez que sabes qué información necesitas, el siguiente paso es encontrarla. Esto implica buscar las fuentes de datos disponibles y hacer una primera evaluación para ver si contienen lo que necesitas.

- **Ejemplo del libro**:

  - Buscas en las bases de datos del equipo. Encuentras tablas con datos sobre placajes, intercepciones y balones sueltos para cada jugador.

  - **¡Problema!** No encuentras ninguna tabla con una columna llamada "sacks". Sin embargo, encuentras otra tabla con todos los placajes registrados con fecha y hora, y la posición del jugador placado. Como sabes que un "sack" es un placaje a un *quarterback*, te das cuenta de que **quizás** podrías cruzar ambas tablas para derivar esa métrica.

  - **Iteración**: Vuelves a hablar con los interesados (Paso 1) y les preguntas: "¿Qué tan importante es la métrica de 'sacks'? ¿Vale la pena el esfuerzo de intentar derivarla?".

## 3. Preparación de los Datos (Data Preparation)

- **¿De qué se trata?**: Esta es la fase de "arremangarse". Es el proceso de tomar los datos brutos y desordenados y convertirlos en un conjunto de datos limpio y estructurado, listo para el análisis. Como dice el libro, esta fase puede llevar una gran proporción del tiempo del proyecto.

- **Ejemplo del libro**:

  - **Limpieza**: Reemplazas los valores nulos en las estadísticas de los jugadores por ceros, después de confirmar con el negocio que esa es la lógica correcta.

  - **Unión de Tablas**: Creas una nueva tabla que une toda la información de

rendimiento para cada jugador.

  - **Ingeniería de Características**: Si decidiste que la métrica "sacks" era importante, aquí es donde harías el trabajo complejo de cruzar las dos tablas por fecha y hora para crear esa nueva columna.

## 4. Análisis Exploratorio y Modelado (Exploratory Analysis and Modeling)

- **¿De qué se trata?**: ¡La fase de descubrimiento! Aquí finalmente te "sumerges" en los datos limpios. Usas **estadísticas descriptivas** y, sobre todo, **visualizaciones** (gráficos de barras, histogramas, etc.) para entender la distribución de tus datos, encontrar patrones, detectar valores atípicos (*outliers*) y empezar a responder las preguntas de negocio. A esto se le llama **EDA (Exploratory Data Analysis)**.

- **Ejemplo del libro**:

  - Creas histogramas para ver la distribución de placajes entre todos los jugadores.

  - Identificas a los jugadores que son *outliers* (los que tienen un rendimiento excepcionalmente alto o bajo).

  - Si los interesados te han dado una calificación (de 1 a 10) para algunos jugadores, aquí podrías entrenar un modelo simple (como una regresión lineal) para predecir la calificación de los jugadores restantes basándote en sus estadísticas.

## 5. Validación y Evaluación (Validation and Evaluation)

- **¿De qué se trata?**: Es la fase de "revisar tu trabajo". No puedes confiar ciegamente en tus resultados. Debes validar que tus cálculos son correctos, que tus suposiciones son lógicas y que el modelo, si lo has creado, es preciso.

- **Ejemplo del libro**:

  - **Validación**: Verificas que la transformación que hiciste para calcular los "sacks" no tiene errores lógicos.

  - **Evaluación**: Evalúas si la regresión lineal que creaste predice bien las calificaciones de los jugadores, o si es inútil.

  - **Revisión por Pares (*Peer Review*)**: Le pides a un compañero que revise tu código y tu enfoque para detectar posibles errores o suposiciones incorrectas

que tú no habías visto.

## 6. Visualización y Presentación (Visualization and Presentation)

- **¿De qué se trata?**: La fase final. Tu análisis no sirve de nada si no puedes comunicarlo de forma efectiva a las personas que toman las decisiones. Aquí es donde compilas tus hallazgos en un formato comprensible.

- **Ejemplo del libro**:

  - El resultado final para el equipo de fútbol podría ser una **tabla simple** que clasifica a los jugadores defensivos de mejor a peor según una puntuación combinada de sus métricas.

  - También podrías incluir un par de **gráficos clave** que resalten las tendencias más importantes que descubriste durante tu exploración.

  - Si el análisis va a ser recurrente, el resultado final podría ser un ***dashboard*** en una herramienta como Tableau que se actualice automáticamente cada semana.

Este ciclo, como ves, es una guía estructurada que asegura que un proyecto de datos esté siempre conectado al negocio, basado en datos de calidad y que sus resultados sean confiables y accionables.