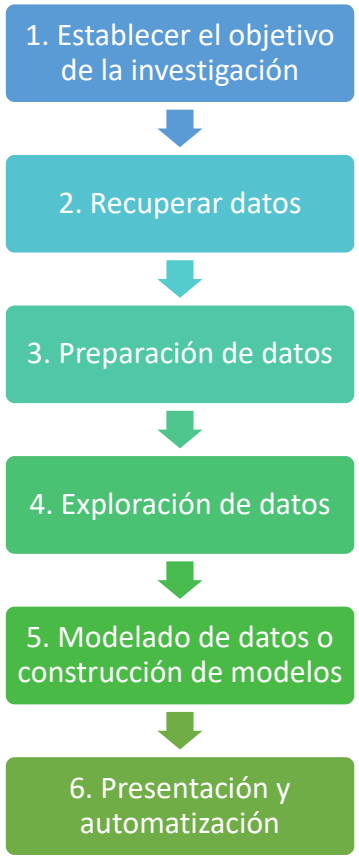
Proceso de Ciencia de Datos

Proceso de Ciencia de Datos

El proceso de ciencia de datos normalmente consiste de seis pasos.



1. Establecer el objetivo de la investigación

La ciencia de datos se aplica principalmente en el contexto de una organización. Cuando el negocio le pide que realice un proyecto de ciencia de datos, primero preparará una carta del proyecto. Esta carta contiene información como qué vas a investigar cómo la empresa se beneficia de eso, qué datos y recursos necesita, un cronograma y entregables.

1. Recuperar los datos

El segundo paso es recopilar datos. En la carta del proyecto ha indicado qué datos necesita y dónde puede encontrarlo. En este paso, se asegura de poder utilizar los datos en su programa, lo que significa verificar la existencia, calidad y acceso a los datos.

Los datos también pueden ser entregados por empresas de terceros y adoptan muchas formas que van desde hojas de cálculo de Excel hasta diferentes tipos de bases de datos.

3. Preparación de datos

La recopilación de datos es un proceso propenso a errores; en esta fase se mejora la calidad de datos y prepararlos para usarlos en los pasos siguientes. Esta fase consta de tres subfases:

- La limpieza de datos que elimina los valores falsos de una fuente de datos y las inconsistencias en todas las fuentes de datos.
- La integración de datos enriquece las fuentes de datos al combinar información de múltiples fuentes de datos.
- La transformación de datos asegura que los datos estén en un adecuado formato para usar en sus modelos.

4. Exploración de datos

La exploración de datos se ocupa de desarrollar una comprensión más profunda de sus datos. Intenta comprender cómo interactúan las variables entre sí, la distribución de datos y si existen valores atípicos. Para lograr esto, utiliza principalmente estadísticas descriptivas, técnicas visuales y modelado simple. Este paso a menudo se conoce con la abreviatura EDA, para análisis de datos exploratorios.

5. Modelado de datos o construcción de modelos

En esta fase, utiliza modelos, conocimiento del dominio e información sobre los datos que se encuentran en los pasos anteriores para responder a la pregunta de investigación. Se selecciona una técnica de los campos de la estadística, el aprendizaje automático, la investigación de operaciones, etc.

Edificar un modelo es un proceso iterativo que implica seleccionar las variables para el modelo, ejecución del modelo y diagnósticos del modelo.

6. Presentación y automatización

Finalmente, se presentan los resultados al negocio. Estos resultados pueden tomar muchas formas, desde presentaciones hasta informes de investigación. A veces, se necesitará automatizar la ejecución del proceso porque la empresa querrá utilizar los conocimientos que ha obtenido en otro proyecto o permitir que un proceso operativo utilice el resultado del modelo.

Proceso Iterativo en Ciencia de Datos

La descripción anterior del proceso de ciencia de datos le da la impresión de que recorre este proceso de forma lineal, pero en realidad, a menudo hay que dar un paso atrás y reelaborar ciertos hallazgos.

Por ejemplo, puede encontrar valores atípicos en la fase de exploración de datos que apuntan a errores de importación de datos. Como parte del proceso de ciencia de datos, obtiene incrementos conocimientos, que pueden dar lugar a nuevas preguntas. Para evitar retrabajos, asegúrese de que delimita la cuestión empresarial de forma clara y exhaustiva desde el principio.