Clase 2: Metodologías para proyectos de Ciencia de Datos.

Diplomado de Ciencia de Datos
Notas por José Daniel Conejeros - jdconejeros@uc.cl

Junio del 2020

Profesora: Francia Berna - f.berna@uc.cl

I. Iniciar un proyecto de Ciencia de Datos

¿Por dónde partir?

Antes de considerar los aspectos técnicos se le debe dar un **enfoque** a un proyecto de ciencia de datos: (1) definir objetivos y (2) diseñar estrategias que sean innovadoras.

Para el planteamiento y definición de un problema hay que hacerse todas las preguntas posibles. Si queremos que los datos trabajen para nosotros, debemos ser capaces de hacer las preguntas adecuadas. No hay patrones únicos a la hora de diseñar un proyecto ya que dependerá de las características de cada desafío, por lo que cada proyecto de ciencia de datos será diferente (no es solo aplicar códigos). Además, se debe tener las siguientes consideraciones:

- Requiere de un trabajo multidisplinarios para robustecer la visión integral del proyecto.
- Mantener una comunicación y gestión constantes con las empresas para obtener sus datos.

Actividades que realiza un cientista de datos

A partir de un estudio realizado por Crowd Flower (2017) a 179 científicos de ciencia de datos en todo el mundo, se identifico que las principales actividades que se realizan son las siguientes:

- Colectar, etiquetar, limpiar y organizar los datos (51 % del tiempo)
- Construir y modelar los datos (19 % del tiempo)
- El modelado de datos para patrones (10 % del tiempo)
- Refinar algoritmos (9 % del tiempo)
- Otras actividades (8 % del tiempo)

A su vez, un reporte de Kaggle (2017) presentó los principales obstaculos a la hora de desarrollar un proyecto de ciencia de datos:

- Datos sin limpieza adecuada (49 %)
- Falta de talento (42 %)
- Falta de apoyo administrativo y/o Financiero (37 %)
- Falta de claridad en el reto a resolver (30 %)
- Datos no disponibles o de difícil acceso (30 %)
- Otros

En consecuencia, es necesario trabajar con una metodología de proyectos de ciencia de datos capaz de responder a estos obstaculos de manera sistemático y así, alcanzar los objetivos fijados en el proyecto de ciencia de datos.

II. Metodología para proyectos de Ciencia de Datos

Existen diferentes metodologías para desarrollar proyectos de ciencia de datos:

- Knowledge Discovery in Databases (KDD)
- Sample, Explore, Modify, Model and Access (SEMMA)
- Cross-Industry Standard Process for Data Mining (CRIPS-DM)

A pesar de que no existe un consenso sobre la forma más adecuada de trabajar un proyecto de ciencia de datos, se recomienda un proceso de 6 etapas que es independiente de la metodología:

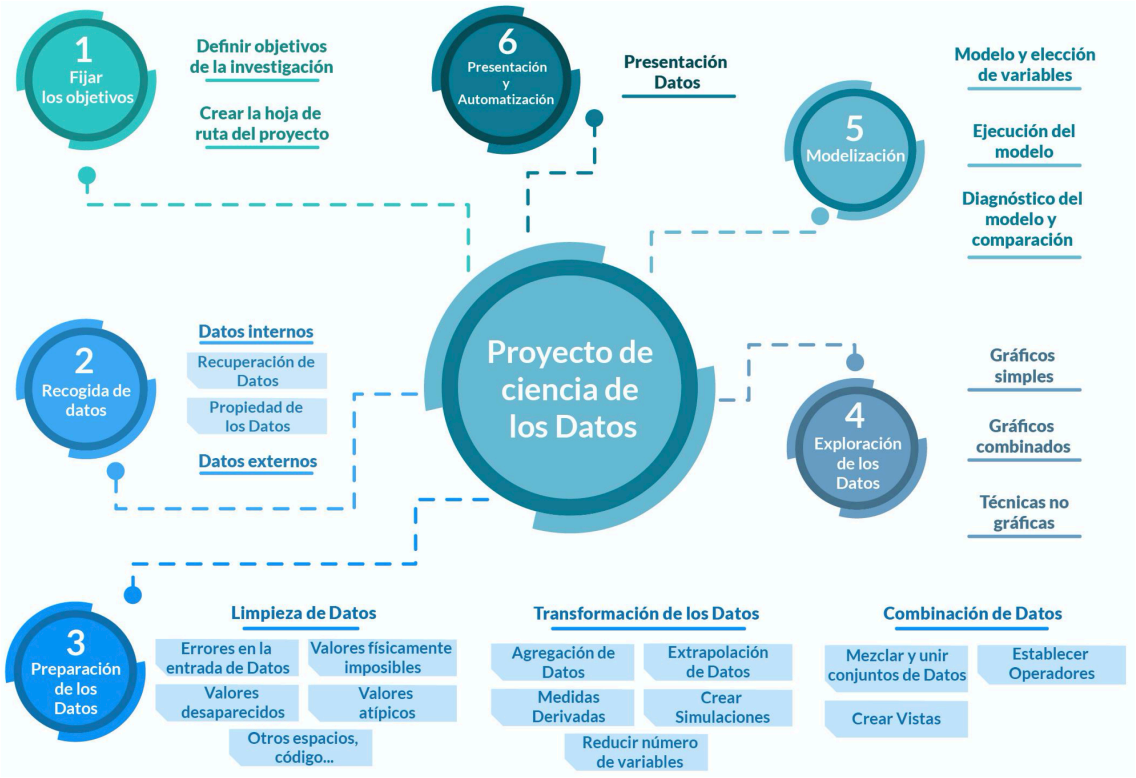

Figura 1: Metodología de Proyectos de Ciencia de Datos

1. **Fijar objetivos:** definir y crear la hoja de ruta del proyecto. Es necesario definir el problema de ciencia de datos y dedicar tiempo en construir una idea de proyecto que captura el qué, el cómo y el por qué de dicho proyecto. Todos los/as participantes deben comprender estos objetivos.

2. **Recogida de datos:** Diseñar un proceso de recolección de datos (internos/externos). El resultado son los datos en crudo o primarios que siempre es necesario procesar antes de utilizar en nuestros análisis.

3. **Procesar los datos:** Tranaformación de los datos primarios en datos utilizables para el análisis. La preparación de los datos considera su **limpieza, transformación y combinación**. Se corrigen distintos tipos de errores en los datos y se fusionarán de diferentes fuentes, normalizando, agrupando, etc. No solo se puede eliminar observaciones, se puede buscar diferentes fuentes de información para acceder a los datos.

4. **Exploración de los datos:** obtener una comprensión de los datos con técnicas visuales y descriptivas para buscar patrones, correlaciones y desviaciones. Estos conocimientos nos permiten comenzar a modelar.
5. **Construcción del modelo o modelado de datos:** Modelo y elección de variables, ejecución del modelo, diagnóstico del modelo y comparación. En este espacio se obtienen los conocimietos relevantes ("insights").
6. **Presentar resultados y automatizar:** presentar resultados ojalá de una manera atractiva y utilizando la automatización (ayuda a ahorrar tiempo). Además es necesario utilizar la información para convencer a los interesados de que los resultados realmente cambiarán el proceso del negocio.

El avance no es lineal, de hecho, lo habitual es trabajar de **forma iterativa** en un proceso recursivo. La división de un proyecto en etapas más pequeñas también permite trabajar de manera colaborativa con especialistas en cada uno de los procesos.

Documento elaborado con las herramientas de Rmarkdown