1INF03 - Análisis de Datos

Lucio Cornejo

2022-03-28

Contents

About

Apuntes del curso **Análisis de Datos**, dictado en la *Pontificia Universidad Católica del Perú.*

1

Semana 03/21

1.1 Viernes

- Será necesario hacer un grupo con otros estudiantes del curso, con quienes se comparta afinidad de investigación, para el proyecto final del curso, el cual se irá desarrollando a lo largo del curso.
- Python y R son complementatios, no es que uno sea *mejor* que el otro.
- Potential project partner: Screenshot
- En el curso, usaremos Python en su mayoría, pero también se compartirá, después de clase, el código análogo ,en R, de lo que trabajemos.
- En la unidades 4 y 5, es donde más podremos contrastar el uso de Python y R. De esa manera, uno tendría más claro qué lenguaje escoger al momento de iniciar algún proyecto particular.
- Fechas de laboratorio
  - 9 abril
  - 23 abril
  - 7 mayo
  - 11 junio
  - 25 junio
- **Las dirigidas (perhaps a veces pcs) se me de IOP se me cruzan con todos los labs, excepto por el primero.**

1.1.1 Metodología KDD

1.1.1.1 ¿Qué es un dato?

- El dato es el valor de una característica/variable/atributo (edad, sexo, etc) de la población (población delimitada en espacio, tiempo, etc).
- Procesos paralelos

Variable ⇒ Variable aleatoria ⇒ Dato

Población ⇒ Muestra ⇒ Observación

- La **información** parte de la unión de los datos recopilados.
  - Es de utilidad para tomar decisiones.
  - Un solo dato, por su cuenta, no nos da información.
- El **conocimiento** es un conjunto de informaciones aplicadas, que permite preveer y planificar.
  - La información asociada a un **contexto** y una **experiencia** se convierte en conocimiento.

1.1.1.2 Descripción de la metodología KDD

- KDD: Knowledge Discovery in Databases
- Algunas definiciones:

  Knowledge Discovery in Databases is the non trivial process of identifying valid, novel, potentially useful, and utimately understandable patterns in data.

- *Nivel bajo de datos* se refiere a datos que no nos dice nada, pero que podría servir para generar conocimiento a partir de estos datos.

1.1.1.3 Etapas de la metodología KDD

- Esta metodología nos da pasos para cómo convertir **datos** en **conocimiento**.
- Estas etapas no son obligatorias … sirven de **guía**.
- En la etapa **selection**, se reduce la cantidad de data, quedándonos con la data que **nos va a servir** para lograr el objetivo de nuestro análisis.
  Implica filtrar filas y/o columnas/variables de la data (entendida como data frame).
  Requiere el entendimiento del objetivo del análisis.
- La parte de información surge en la etapa **Patterns** de la metodología KDD. Esa información requiere del bloque *interpretation/evaluation* (ver imagen) para convertirse en **knowledge** .
- El paso de **Transformed data** a patterns es vía “Descriptive methods”.

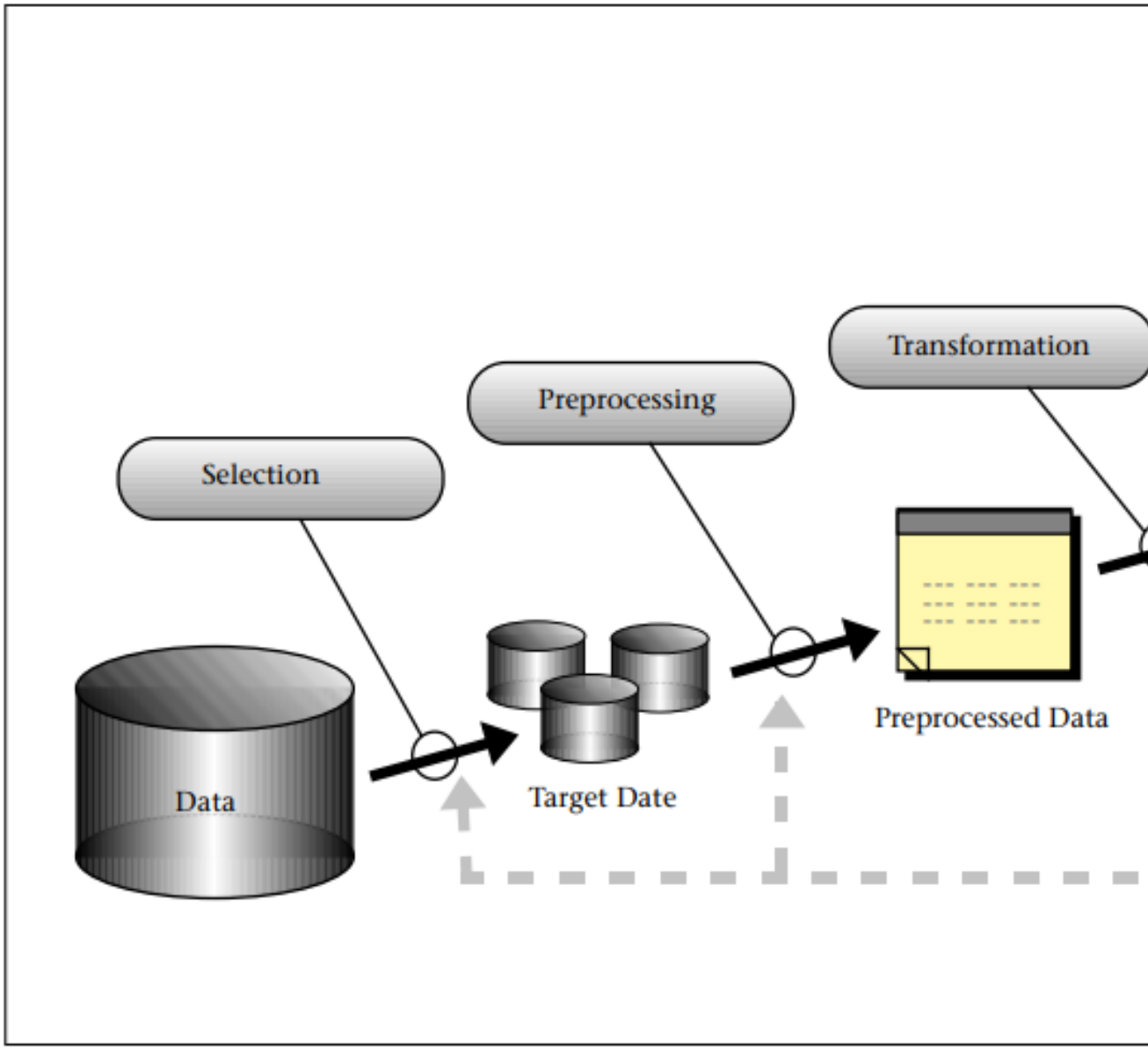

Figure 1. An Overview of the Steps T

Figure 1.1: Etapas de la metodología KDD

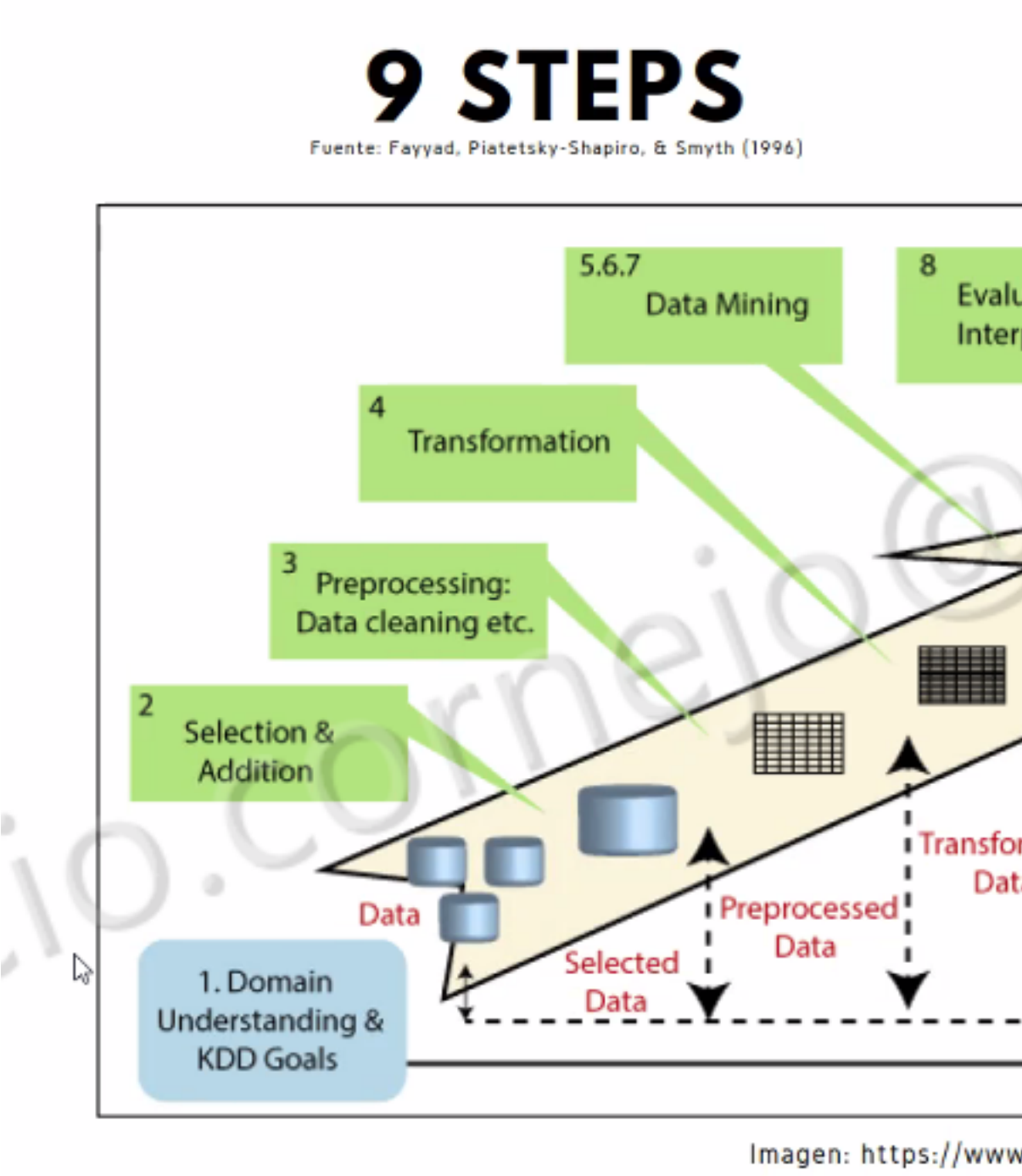

Figure 1.2: Etapas (más a detalle) de la metodología KDD

- Las flechas verticales indican que, a medida que avanzamos en las etapas, podemos volver al inicio para poder obtener nueva data que haya surgido la necesidad de requerir para el análisis.
- El bloque **Active DM (Data Mining)** se refiere a que el proceso *Data mining* forma parte de TODO el proceso de 9 pasos (es otro enfoque). **Regresar a cualquier paso es válido**.

1. Paso 1
   - **Es el paso principal**.
   - Reunirse con los expertos del tema en que se va a trabajar. Se discuten cosas como
     - ¿Cómo sucede el fenómeno?
     - ¿Qué agentes intervienen con el fenómeno?
     - ¿Qué datos se recolectan (variables disponibles) o se pueden recolectar para el fenómeno?
     - ¿Para qué población se va a construir el proyecto?
   - Se habla en lenguaje entendible para todos los expertos, no usando, por ejemplo, palabras particulares de Estadística.
   - Se busca **entender el negocio/problema**.
   - Se busca identificar la **meta** del proceso KDD desde la perspectiva del **customer**.
   - Es más que nada un proceso *cualitativo* que servirá para formalizar el análisis futuro.
   - Es recomendable crear una ficha resumen sobre este paso, donde se anota la información recopilada en la reunicón (o reuniones) con el customer.
     - Asignar un experto del negocio como encargado del proyecto. Esta persona debe ir validando el avance del proyecto, en cada uno de los 9 pasos.
     - Anotas una meta principal y las secundarias.
     - Una vez completa esta ficha resumen es que podemos pasar al siguiente paso; debe redactarse, quedar como evidencia.
2. Paso 2
   - **Creating a target data set**.
   - Filtramos la data para obtener un subconjunto, tanto en variables (columnas) y data samples (filas), al cual se le analizará durante pasos siguientes.
   - No se trata de la selección de variable que se realiza con código, por ejemplo la que busca explicar un fenómeno con las variables *independientes*.
   - Esta selección **no** tiene que ver con la **calidad de datos.** Esa selección ocurrirá más adelante.
   - Formalmente, estos filtros se realizan en base a **criterios de inclusión/exclusión**.
3. Paso 3
   - **Data cleaning and preprocessing**.

- Se le dice también *remover el ruido*. Donde, el *ruido* hace referencia a los **datos atípicos**.
- Se ve la forma de trabajar los *datos perdidos*.
  - Para construir un modelo, necesitamos lidiar primero con los datos perdidos.
  - Dependiendo del contexto, y requiriendo fundamento, se pueden imputar/reemplazar los datos vacíos por **cero**, **la mediana de esa variable**, etc.
  - Desde el punto de vista de la profesora, máximo se debería imputar el 30% de los valores vacíos de una misma variable (que tiene varios valores vacíos). Pues, sino, se estaría trabajando con una variable *ficticia*, y podría así generar ruido en los resultados obtenidos.
    Pero eso **no es una regla**. La decisión de imputación dependerá del contexto/fenómeno, y debe estar fundamentada **numéricamente**, además de tener sentido respecto al negocio.
    (Por ejemplo, si imputar una variable por cero tiene sentido en cierto contexto particular).
- Debido, en parte, a estas razones, es importante la comunicación constante con un experto del negocio.

4. Paso 4
   - **Data reduction and projection**.
   - **La transformación de la data debe suceder después de la limpieza de esta**.