Datos crudos y datos limpios

Curso de Ingeniería de Características

https://mcd-unison.github.io/ing-caract/

Definición de dato

Datos son valores de variables cualitativas o cuantitativas, que pertenecen a un conjunto de elementos

Definición de dato

Datos son valores de variables cualitativas o cuantitativas, que pertenecen a un conjunto de elementos

- *conjunto de situaciones u objetos*: Situaciones, objetos, población, es el conjunto de datos en los que estamos interesados

Definición de dato

Datos son valores de variables cualitativas o cuantitativas, que pertenecen a un conjunto de elementos

- *variables*: Una medición de un atributo de un elemento

Definición de dato

Datos son valores de variables cualitativas o cuantitativas, que pertenecen a un conjunto de elementos

- *cualitativas*: País, sexo, tratamiento, toman su valor de un conjunto finito
- *cuantitativas*: Talla, temperatura, voltaje, toman su valor de un subconjunto de $\mathbb{R}$

Datos crudos (*raw data*)

- Fuente original de datos
- Dificil de usar para análisis de datos
- Pueden incluir preprocesamiento, muchas veces desconocido
- Puede no necesitar procesamiento

Ejemplos de datos crudos

- El extraño archivo binario que escupe un sensor o una máquina
- El archivo en excel con 10 hojas cada una con formatos diferentes que envía el cliente
- Un JSON complicado que viene de una API
- Un archivo texto con datos anotados después de un análisis visual de un operador

Un dato es crudo si...

- No se a ejecutado ningún software en el
- No se ha manipulado
- No se ha eliminado nada del conjunto de datos
- No se ha realizado ninguna agregación de la información

Datos procesados (*tidy data*)

- Datos listos para análisis o modelado
- El procesamiento incluye mezclado, transformación, selección, etc...
- Puede haber estandares de procesamiento
- Todos los pasos de procesamiento se deben de registrar

Datos procesados

- Cada variable (o característica) se encuentra en una columna diferente
- Cada observación se encuentra en un renglón diferente
- Hay una tabla por cada *tipo* de variable
- Si hay varias tablas, hay características comunes entre tablas que permitan que sean relacionadas

Cosas deseables

- Incluye una fila inicial con el nombre de las características
- Las características tienen nombre descriptivos y comprensibles
- Cada tabla se guarda en un solo archivo

Diccionario de datos (*code book*)

- Información sobre las variables (características)
- Incluye unidades de medida
- Información sobre el proceso de transformación seleccionado
- Información sobre el origen de los datos

Lista de instrucciones (*ETL, Data pipeline*)

- Libretas o scripts con **todos** los pasos para pasar de los datos crudos a los procesados
- La entrada principal son los datos crudos
- No hay parámetros en los scripts
- El orden de aplicación de los scripts es claro

Componentes de datos procesados

- Datos crudos (o donde se obtienen)
- Datos procesados
- Diccionario(s) de datos
- Lista de instrucciones

La primera parte del curso se centra en este proceso.