# DataFlow





# **Dataset Inicial**

Información relacionada al objetivo, presentada en diveros formatos.





## **DATA CLEANING**

Completitud Precision Consistencia Normalizacion

StopWords - Hapaxes - Lemmatización





# **EDA**

Univariado - No gráfico Univariado - Gráfico Bivariado - No gráfico Bivariado - Gráfico

# **Definir Objetivo**

1- Supervisado



2- No Supervisado







### Selección de carácteristicas

- Relación de valor pérdido
- Filtro de baja varianza



### **OUTLIERS**

Univariado: Z- Score, IQR,

Percentiles

Multivariado: LOF, Isolation **Forest** 



Creación de variables en las cuales no sea necesario utilizar toda la data. Además de utilizar las variables que no tienen missings.

- Suma de variables
- Multiplicación de variables



# variables

Creación de variables utilizando la información que tenemos completa de forma inicial







# Train/Test

**Train - 80%** Test - 20%



# Missings

Dato a imputar - Train Imputar dato - Train - Test





Selección de carácteristicas Train- Filtro de alta correlación

Train- Filtro de baja correlación con el objetivo

Train- Multicolinealidad Train - PCA



**Supervisado** 



# Missings

Dato a imputar - Todo el conjunto de datos Imputar dato - Todo el conjunto de datos



### Selección de carácteristicas

- Filtro de alta correlación - Filtro de baja correlación
- con el objetivo
- Multicolinealidad - PCA



