



UNIVERSITYHACK 2019[®] DATA^ΔTHON



Deloitte.



Nosotros somos:



Pablo Manuel Garabal Vaquero

Grado Matemáticas



Emilia Musat

Grado en Ingeniería
Matemática



María del Mar Ruiz Martín

Doble Grado en Ingeniería Informática y
Matemáticas

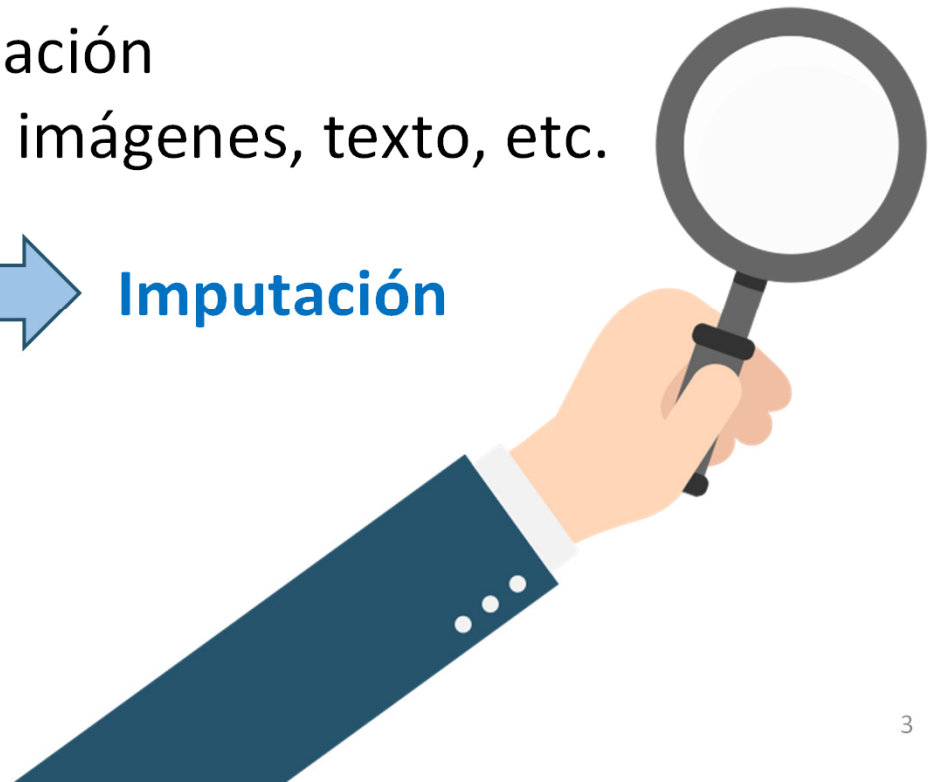
Análisis exploratorio

Diversidad de la naturaleza de los datos

- Diferentes fuentes de información
- Datos espaciales, numéricos, imágenes, texto, etc.

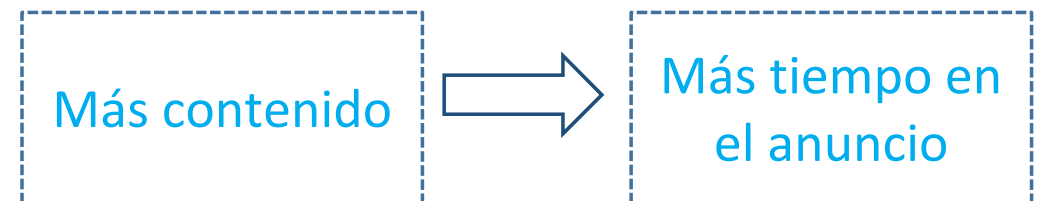
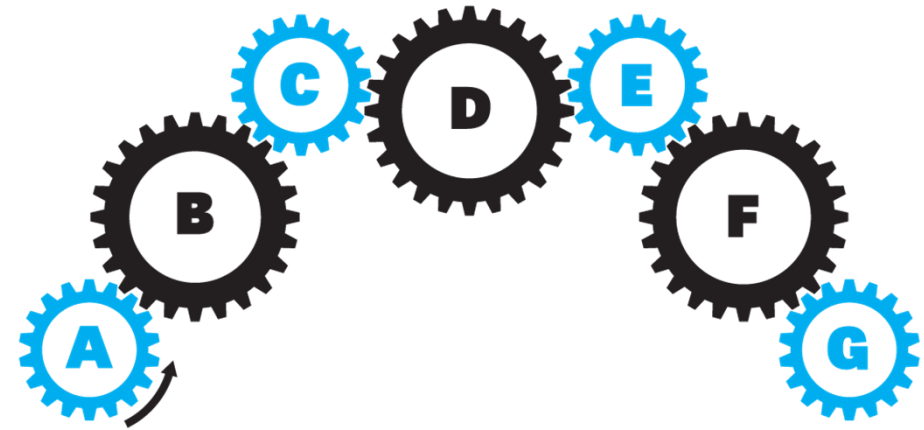
Hasta 83% de valores perdidos ➡ **Imputación**

Inconsistencias en la información



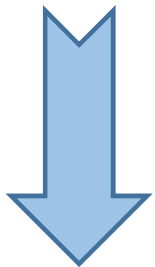
Feature engineering

- Transformación de variables
- Interacción de variables
- Creación de nuevas variables
 - A partir de imágenes
 - A partir de textos

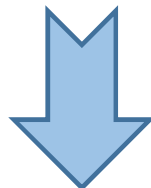


Variable Objetivo

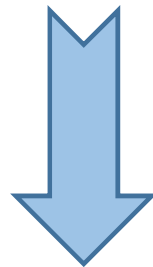
Distribución
asimétrica



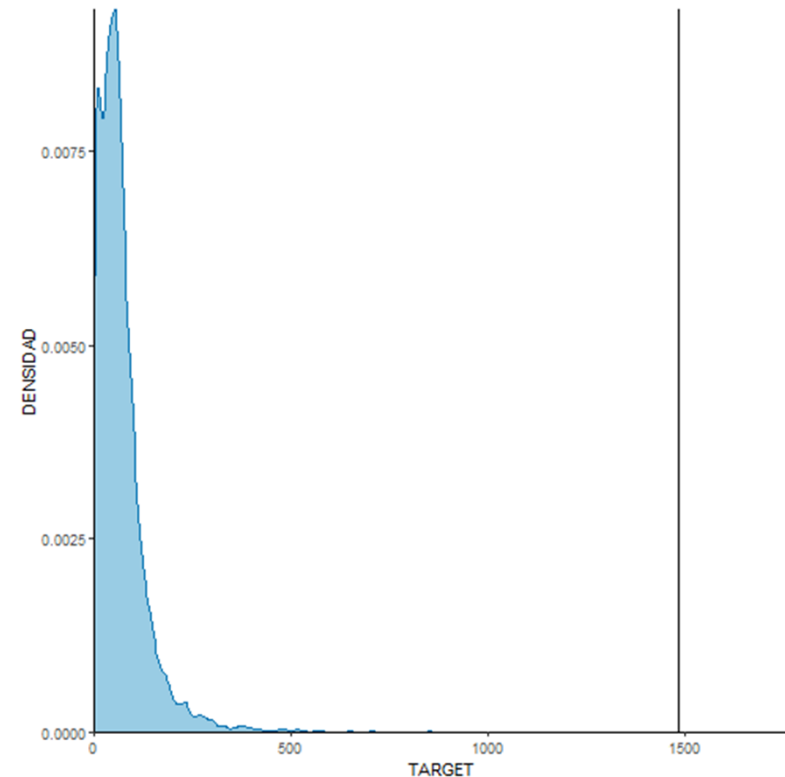
Outliers



Ausencia de
correlaciones



Eliminación de
High-leverage points



Construcción del modelo

Estrategias consideradas

1. Centrada en la calidad de los datos
2. Centrada en los hiperparámetros

Mejores modelos

- Gradient Boosting
- Random Forest
- XGB
- AdaBoost



Evaluación del modelo

Consideraciones

Medida evaluada: *Median Absolute Error (MAE)*

Validación Cruzada: K-fold = 5

Malla de hiperparámetros a optimizar para cada modelo

Diversidad

- ✓ Modelos
- ✓ Variables
- ✓ Conjuntos de entrenamiento

Modelo Final

Gradient Boosting XGB AdaBoost Random Forest

Averaging Ensemble

MAE = 15,95

Ideas descartadas

Clustering

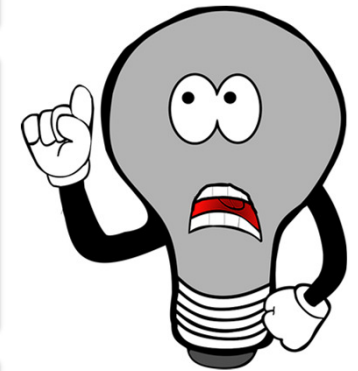
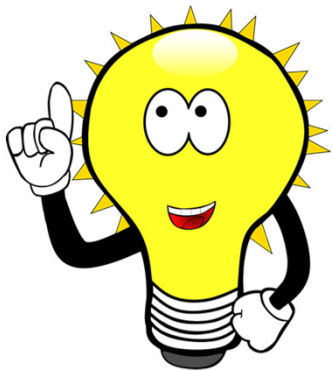
**Grupos
desbalanceados**

**Técnicas de
reducción de
dimensionalidad**

**Pérdida de
variabilidad**

Text mining

Hasta 66% NAs



Mejoras



Técnicas de *data augmentation*

Diferenciar por ubicación o tipo de anuncio

Stacked ensemble

Herramientas





La competición de analítica de datos
más grande de España.

Muchas gracias.