UNIVERSITYHACK 2019®
DATATHON

GRUPO COOPERATIVO CAJAMAR

EY Building a better working world

minsait by Indra

Deloitte.

KABEL

KPMG

VIEWNEXT AN IBM SUBSIDIARY

Go Deeper Or Go Home

UNIVERSITYHACK 2019 DATATHON

Santiago
García Gisbert

Master Data Science para Finanzas CUNEF 2018-2019

Carlos
Vecina Tebar

Master Data Science para Finanzas CUNEF 2017-2018

Data Scientist en Gnarum

/CarlosVecina

/in/carlos-vecina

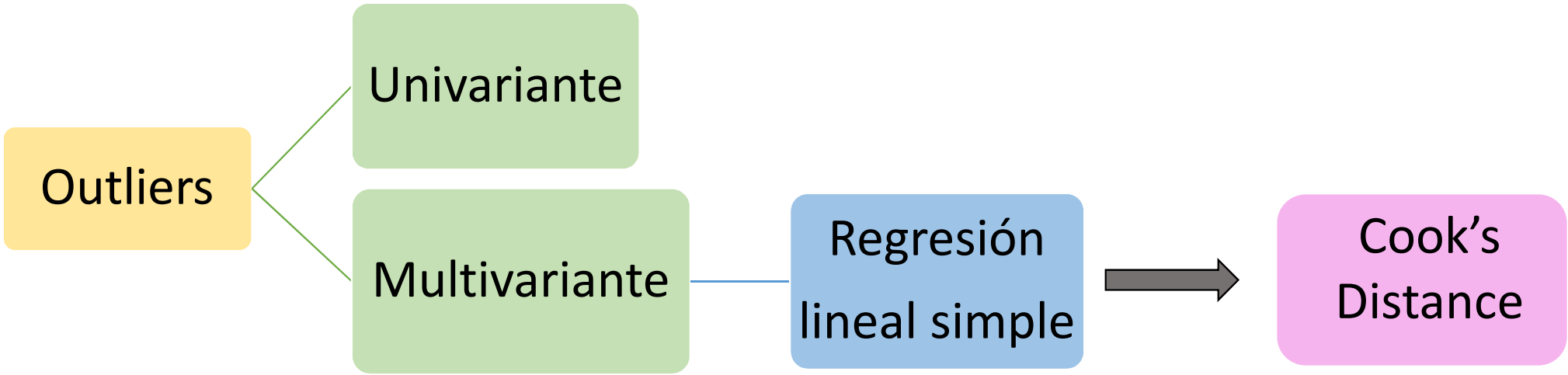
Proyecto:

Minsait Real State Modelling

1. Análisis exploratorio.
2. Estructura del algoritmo y modelado.
3. Procesamiento de imágenes.

❖ *1a. Exploratorio. Imputación y outliers.*

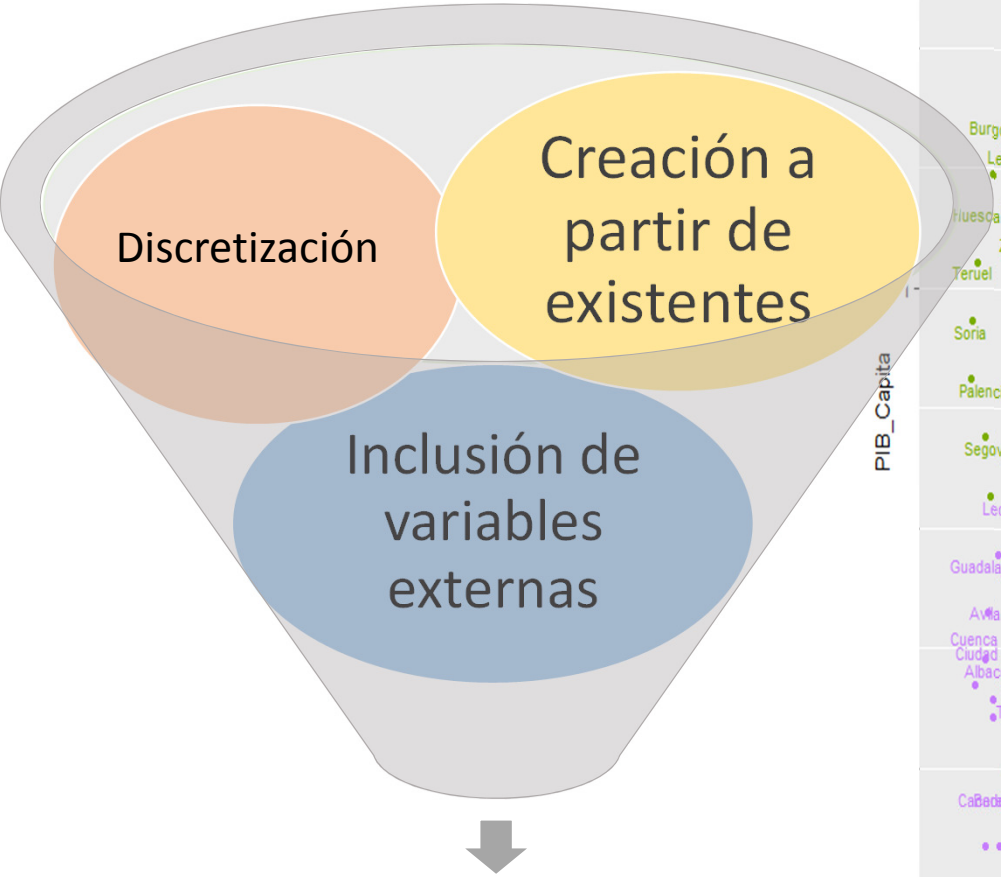
NA → Imputación por agrupación de medianas en base a terceras variables

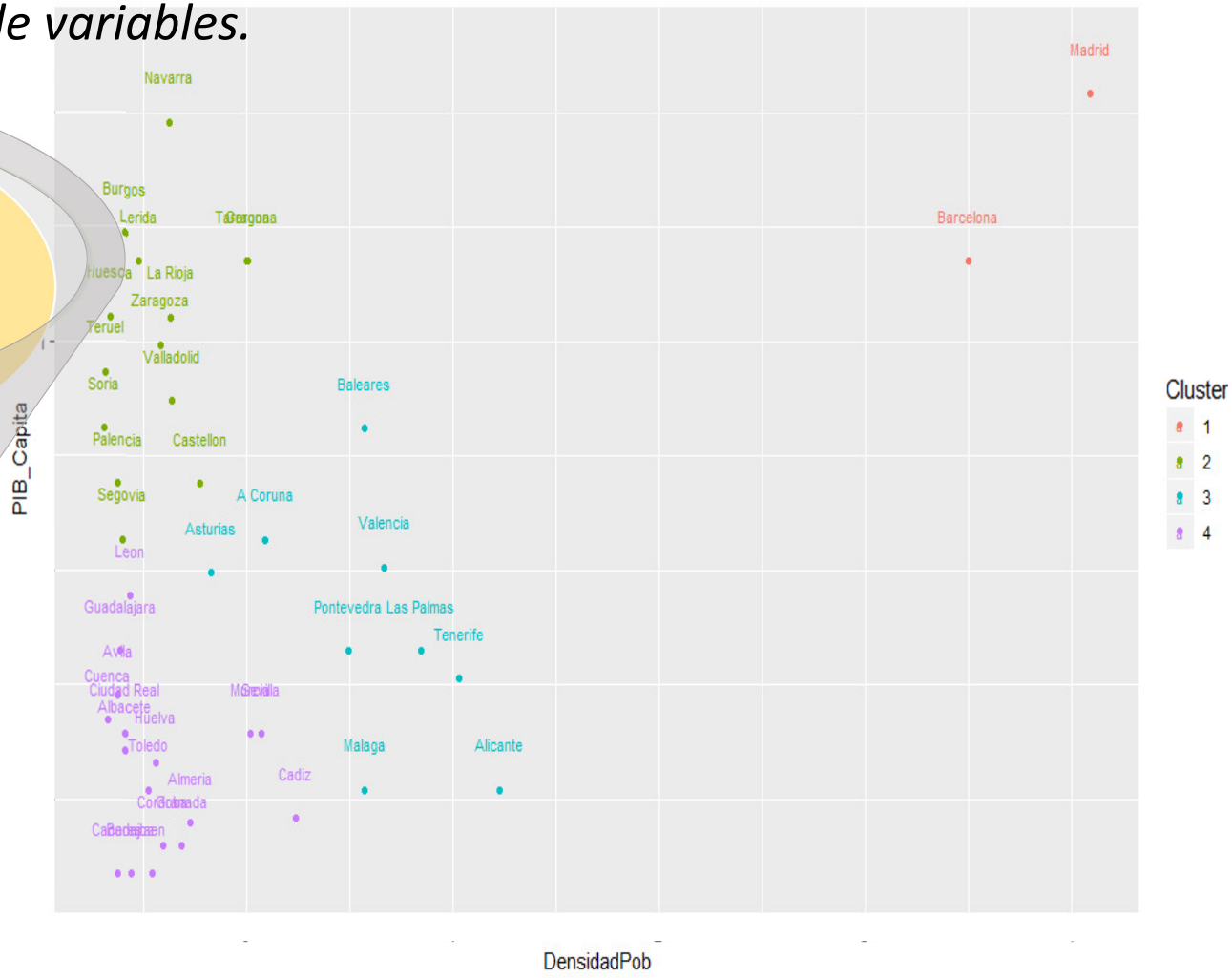


GO DEEPER OR GO HOME

UNIVERSITYHACK 2019 DATATHON

❖ *1b. Exploratorio. Creación de variables.*





Craft de variables

❖ *2a. Estructura del algoritmo (I)*

Características:

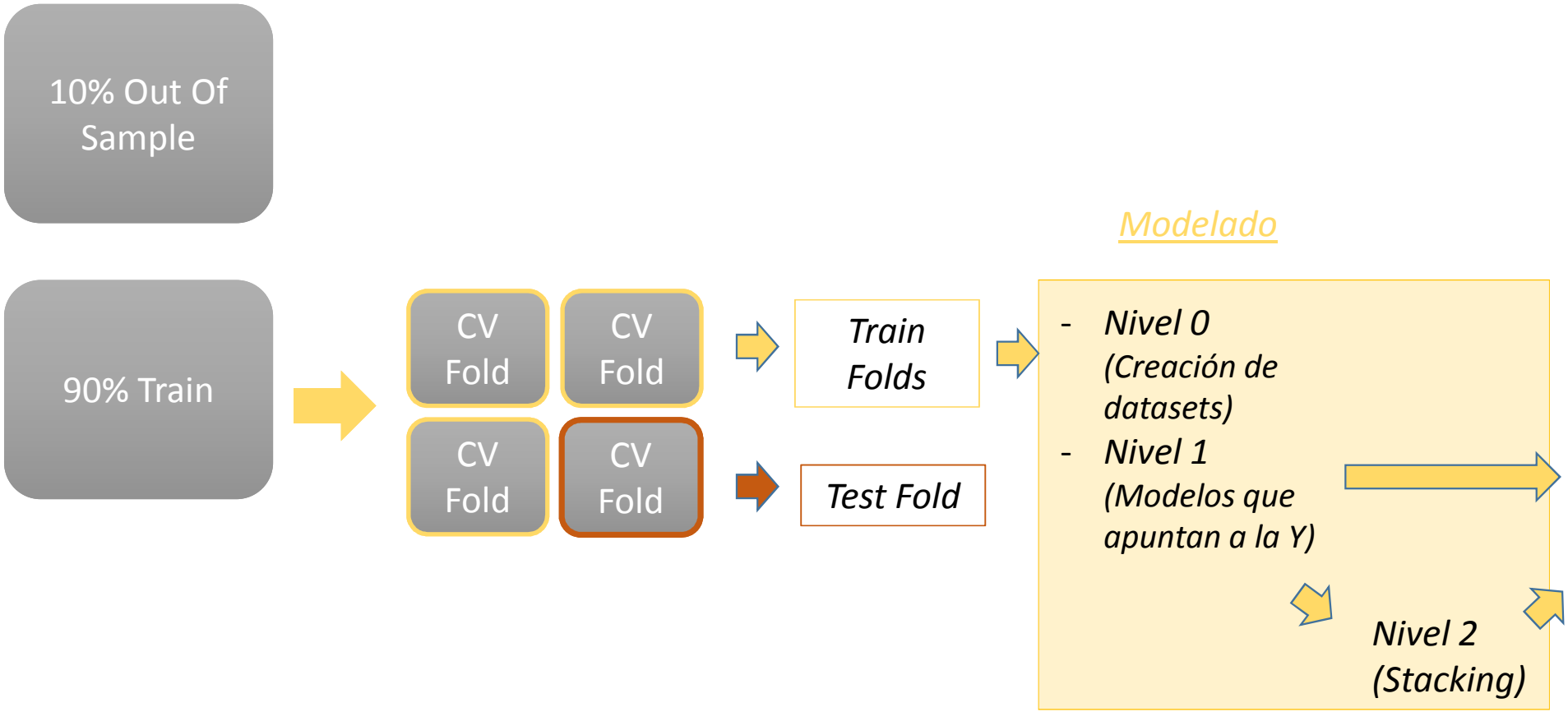
- **Flexible** para combinar (*ensemble y stacking)* modelos de diferente naturaleza y que vean diferentes datos (variables y/o observaciones);
- ***Cross Validation* personalizada** para calcular errores reales en los *folds* y sacar predicciones del *train* «legítimas»

Problemas con las librerías:

- Difícil tratamiento de *outliers* e imputación en los diferentes folds ⇨ Métrica de error distorsionada

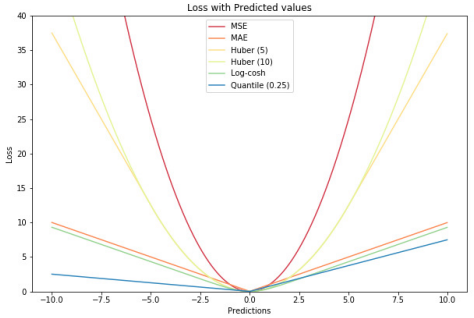
- Predicciones en el *train* obtenidas tras calculo de hiperparámetros óptimos sobre ese mismo conjunto ⇨ Predicciones con un error infra estimado que contaminarán el *stacking*

❖ *2a. Estructura del algoritmo (II)*



❖ *2b. Modelado*

Nivel 0	Creación de datasets	- Filtrado de observaciones. - Modelo clasificación decil de la Y.
Nivel 1	Modelos que apuntan a la Y	- XGB parámetros optimizados *y Log-cosh loss* - SVM - Catboost - Aproximate Bayessian Computation Random Forest
Nivel 2	Stacking de predicciones de los niveles anteriores	- Media ponderada - Regresión penalizada Lasso - XGB con depth muy reducida



❖ *3. Procesamiento de imágenes (I)*

Dos enfoques

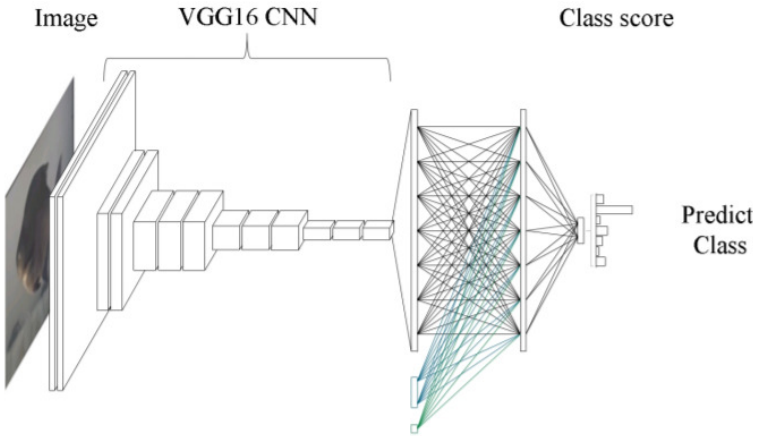
Usadas como variables en los modelos anteriores

- Tiene / No tiene foto
- Número de fotos
- Número de canales de la foto
- Dimensiones y calidad mediana
- Histograma de luminosidad

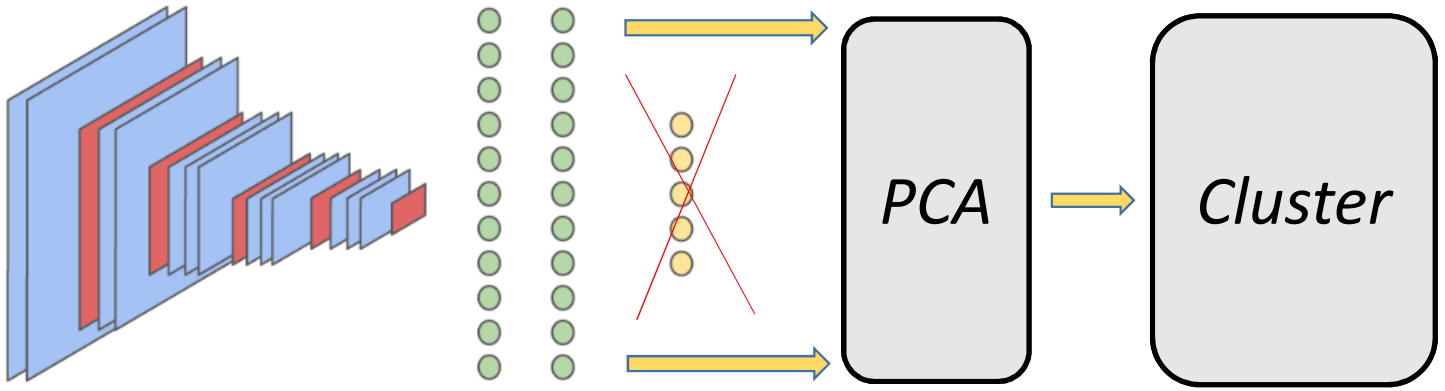
Modelos de *Deep Learning…*

❖ *3. Procesamiento de imágenes (II)*

Transfer Learning



Kratzert, Frederik & Mader, Helmut. (2018)

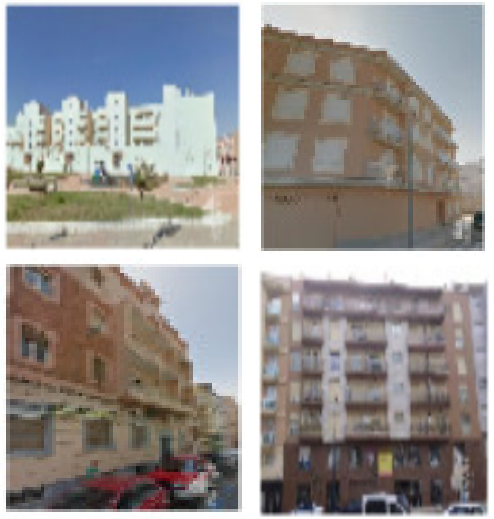


❖ *3. Procesamiento de imágenes (III)*

Cluster PCA 1

Cluster PCA 2

Cluster PCA 3



❖ *4. Conclusiones alejando el foco. Negocio.*

- Variables más relevantes respecto al TARGET

- Interpretación del modelo / predicciones

Local
Interpretable
Model-agnostic
Explanations

UNIVERSITYHACK 2019®
DATATHON

La competición de analítica de datos más grande de España.

Muchas gracias.