

Minsait Land Classification

Equipo Talents

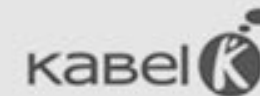


UNIVERSITYHACK 2020®
DATA Δ THON



minsait
by Indra

idealista/data



VIEWNEXT
AN IBM SUBSIDIARY

EOI Escuela de
organización
industrial



Pedro Álvarez Guirado

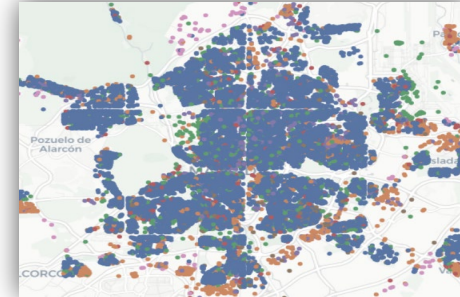
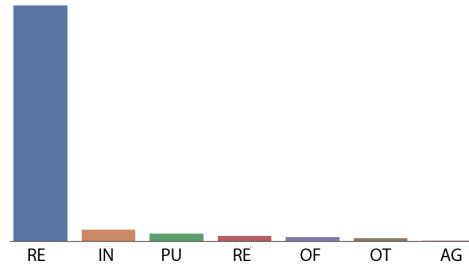


Alejandro Lanaspa Ferrer

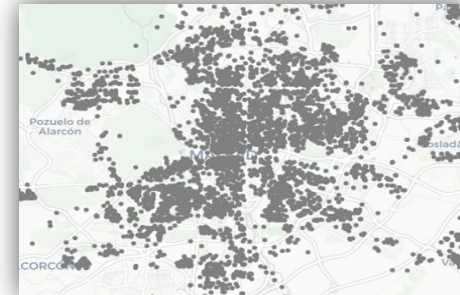
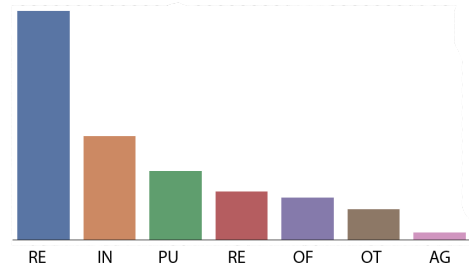


1. Análisis Exploratorio (I)

Training

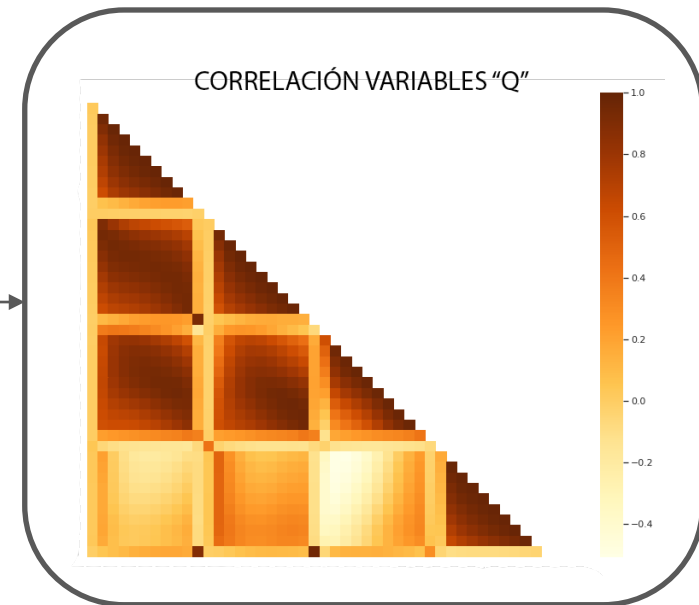
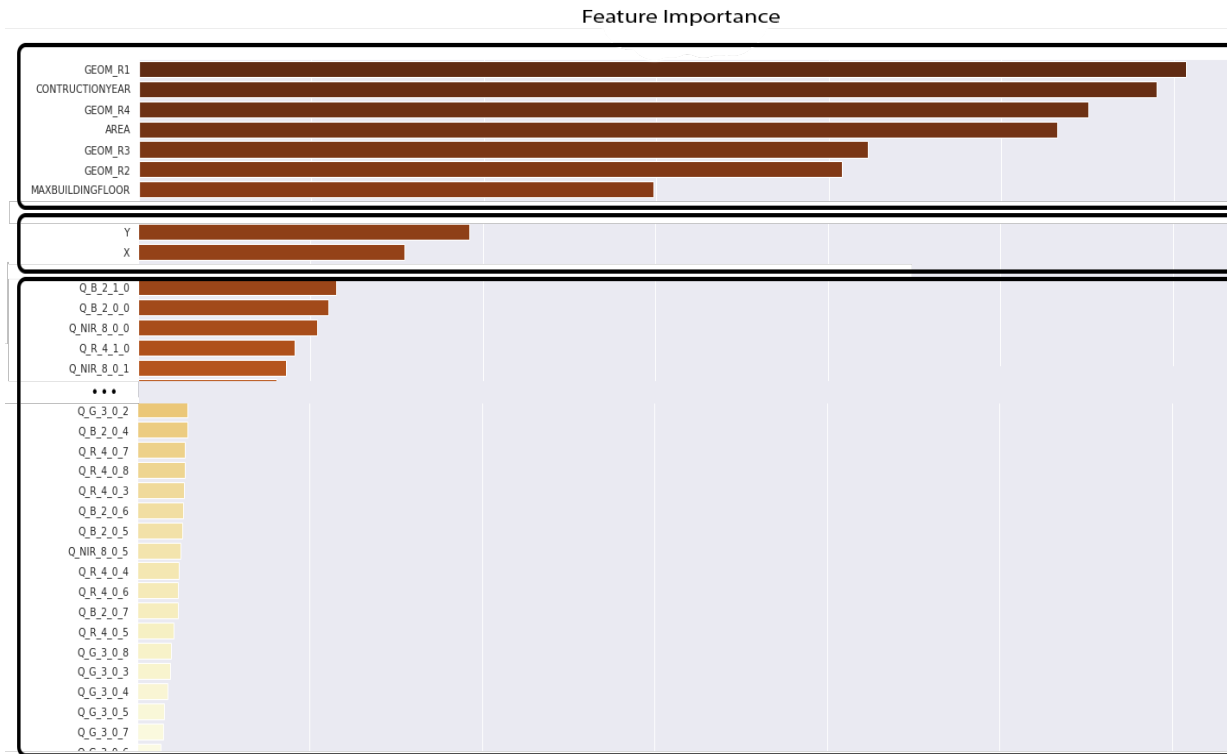


Test



El dataset presenta una alta correlación geoespacial y un alto desbalanceamiento

1. Análisis Exploratorio (II)



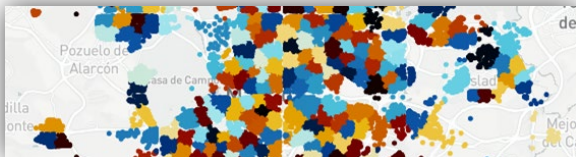
Existen 3 subgrupos en los datos, cada cual con unas propiedades completamente distintas

2. Feature Engineering

Coordenadas & Clases

X Y L

Modelo KNN



Variables "Q"

Q1 Q2 Q3 Q4 ... Q41 Q42 Q43 Q44

PCA

W1 W2 W3 W4 W5 W6 W7

Referencias Geométricas

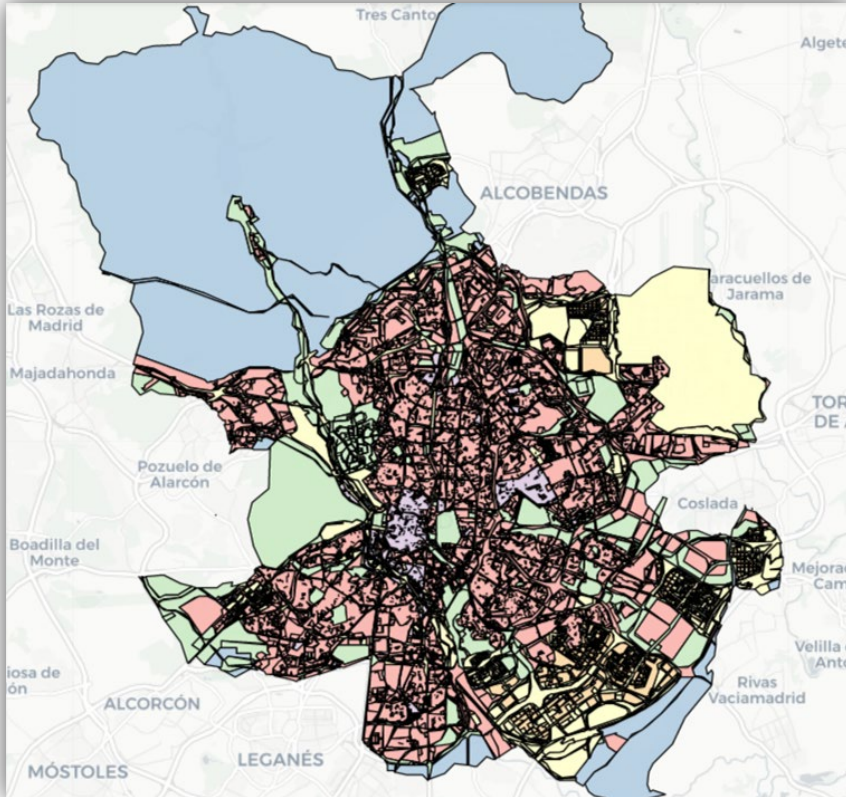
G1 G2 G3 G4 M A C T

Agregaciones

Ag1 Ag2 Ag3 Ag4 ... M A C T

Feature Engineering particularizado para cada subgrupo

3. Datos Externos (I)



Clasificación y **Calificación urbanística** de la CCAA de Madrid (2014)



Atlas Urbano del servicio Copernicus de la EU (2012)



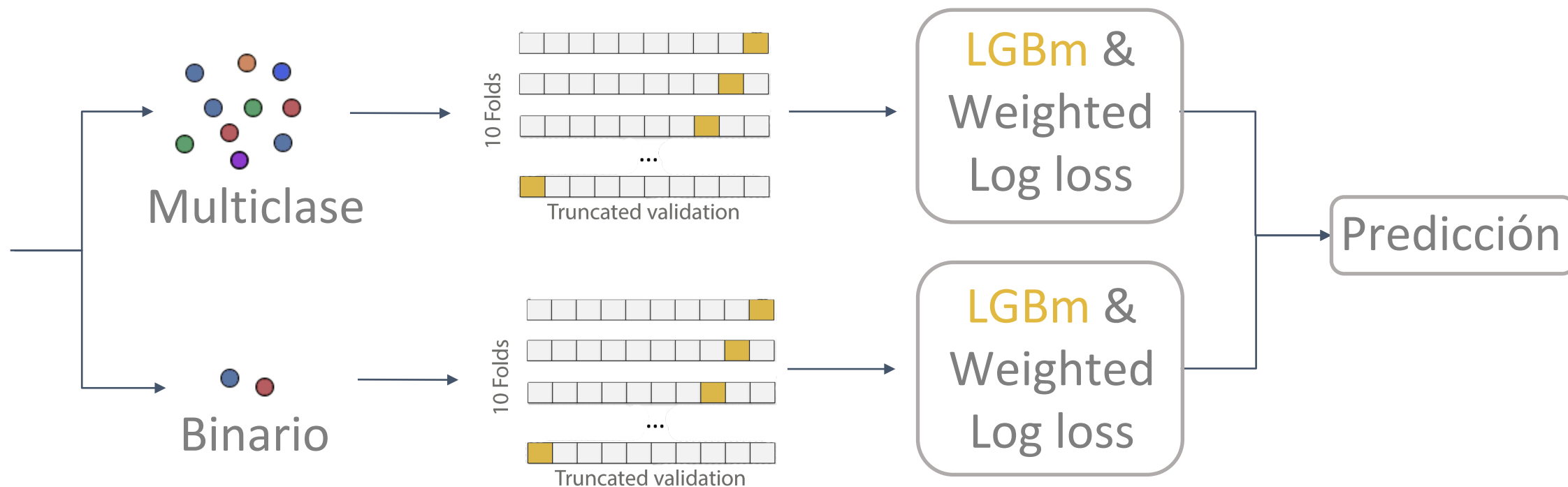
Censo Demográfico del INE (2011)



Datos de **OpenStreetMap** (2020)

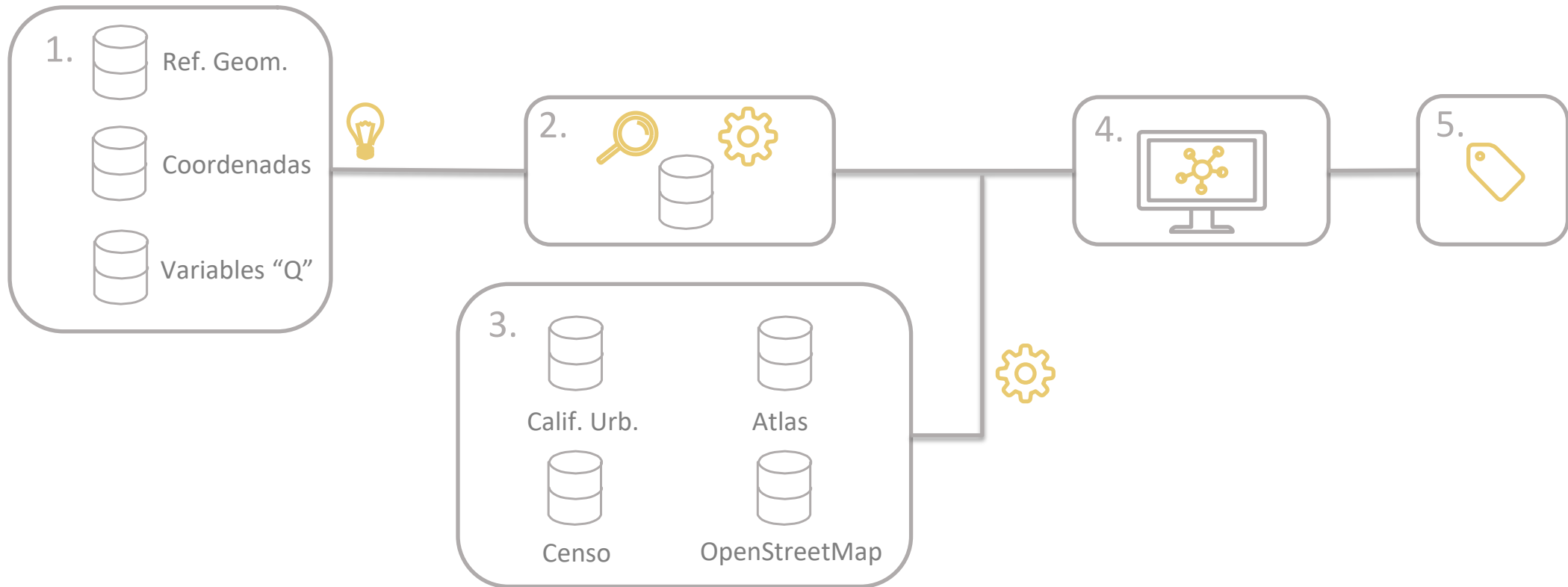
Diversidad y complementariedad en los datos externos

4. Modelado



Combinación de modelo multiclase y binario.

Diagrama de flujo



Proceso de mejora continuo

Resultados (I)

Matriz de confusión

Precisión CV

73.5%

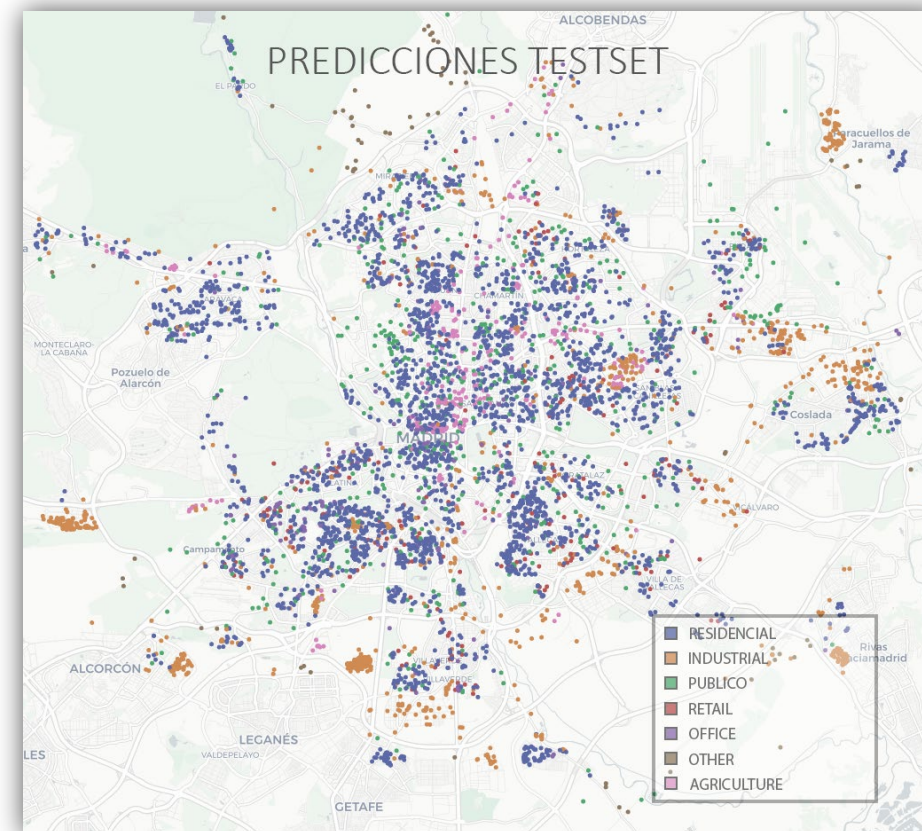
RE	0.97	0.01	0.01	0	0.01	0.01	0
IN	0.16	0.74	0.02	0.04	0	0.02	0
PU	0.22	0.06	0.63	0.03	0.03	0.02	0.01
RE	0.32	0.24	0.06	0.35	0.01	0.02	0.01
OF	0.22	0.08	0.07	0	0.61	0.02	0
OT	0.51	0.12	0.05	0.04	0.02	0.27	0
AG	0.03	0.12	0.03	0	0.03	0	0.79
	RE	IN	PU	RE	OF	OT	AG

- Posible **mejora** en clasificación de **Retail** y **Others**
- **Extracción** de mayor **información** de las variables de imagen/infrarrojos así como de variables descriptivas del inmueble.

Mejora respecto del baseline de un 13.5%

Resultados (II)

- Variables geométricas y área tienen una gran importancia de decisión.
- La clasificación geoespacial mantiene una lógica conforme a los distintos distritos.
- Modelo con alta capacidad de generalización gracias al Ensembling.
- Modelo explicable al estar basado en arboles de decisión.





UNIVERSITYHACK 2020®
DATATHON

La competición de analítica de datos
más grande de España.

Muchas gracias.