Minsait Land Classification

Equipo MATRICS

UNIVERSITYHACK 2020®
DATATHON

GRUPO COOPERATIVO CAJAMAR

minsait by Indra

idealista/data

KABEL

VIEWNEXT
AN IBM SUBSIDIARY

CUNEF
COLEGIO UNIVERSITARIO DE ESTUDIOS FINANCIEROS

Cajamar UniversityHack 20 – Reto: Mindsait Land Classification

Juan Villasante Guerrero
Graduado en ADE

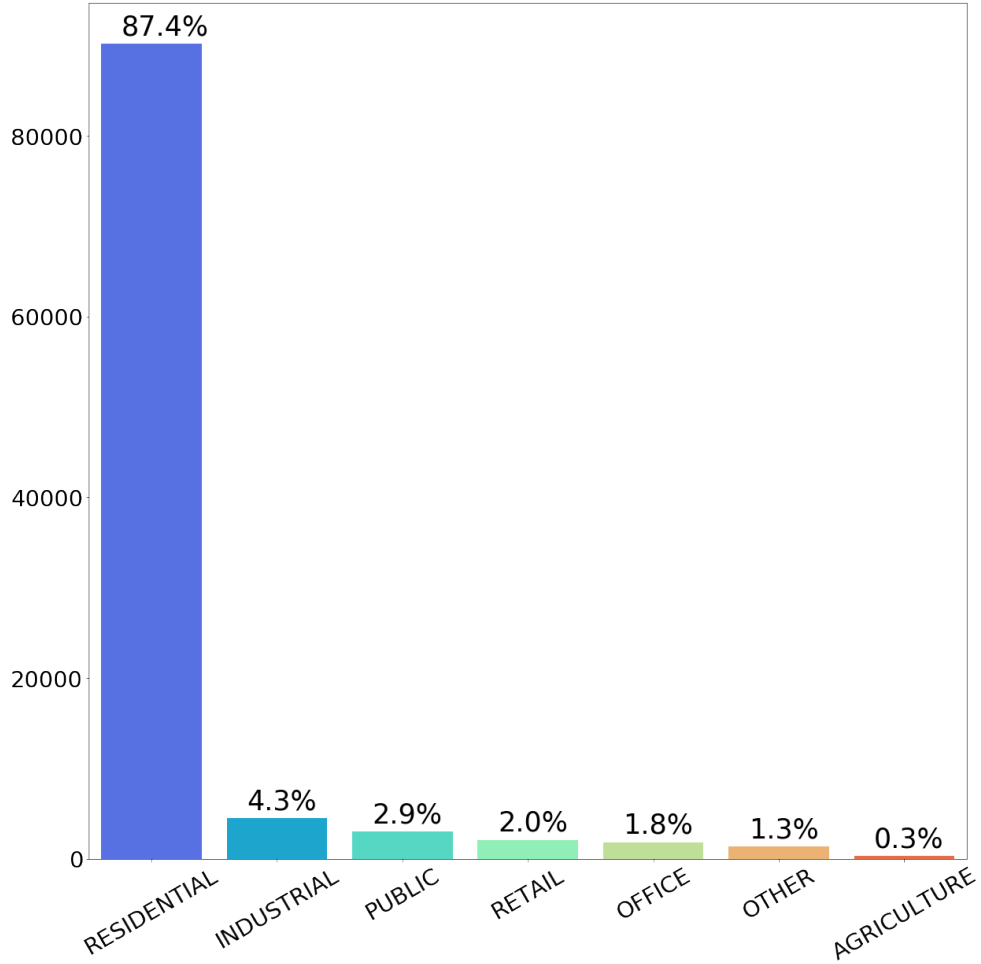
Miguel López Garralón
Graduado en Periodismo

Beltrán Aller López
Graduado en ADE

Índice

Objetivo

- Objetivo: clasificación de parcelas
- Problema **multiclase** con fuerte **desequilibrio**
- Existencia de valores extremos en la variable área
- Muy pocos valores ausentes
- Modelos basados en ***ensembles***



Algoritmos

ALGORITMOS PROBADOS

- Lightgbm
- Catboost
- Xgboost
- Random Forest
- Votingclassifier
- SVM
- Logistic regression
- Redes neuronales

MEJORES ALGORITMOS

- XGBoost
- Lightgbm
- Catboost

ALGORITMO ELEGIDO

- XGBoost

- Mejor equilibrio entre precisión y cómputo requerido

Métrica

¿Cuál es la métrica adecuada? → **Weighted - Recall**

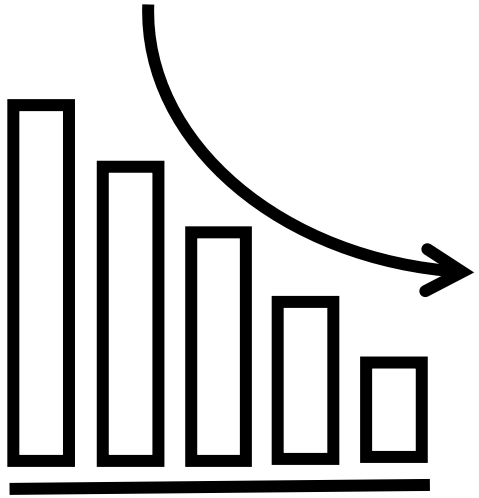
- Información a posteriori
- Distinta distribución del target en datos a modelar y a estimar
- Mayor peso de clases minoritarias
- **Mejorar *recall*** de las clases minoritarias para aumentar la precisión en test
- **Verificar resultados** con menor incertidumbre

Aplicamos un peso 4 veces mayor

Reducción de peso de la clase mayoritaria

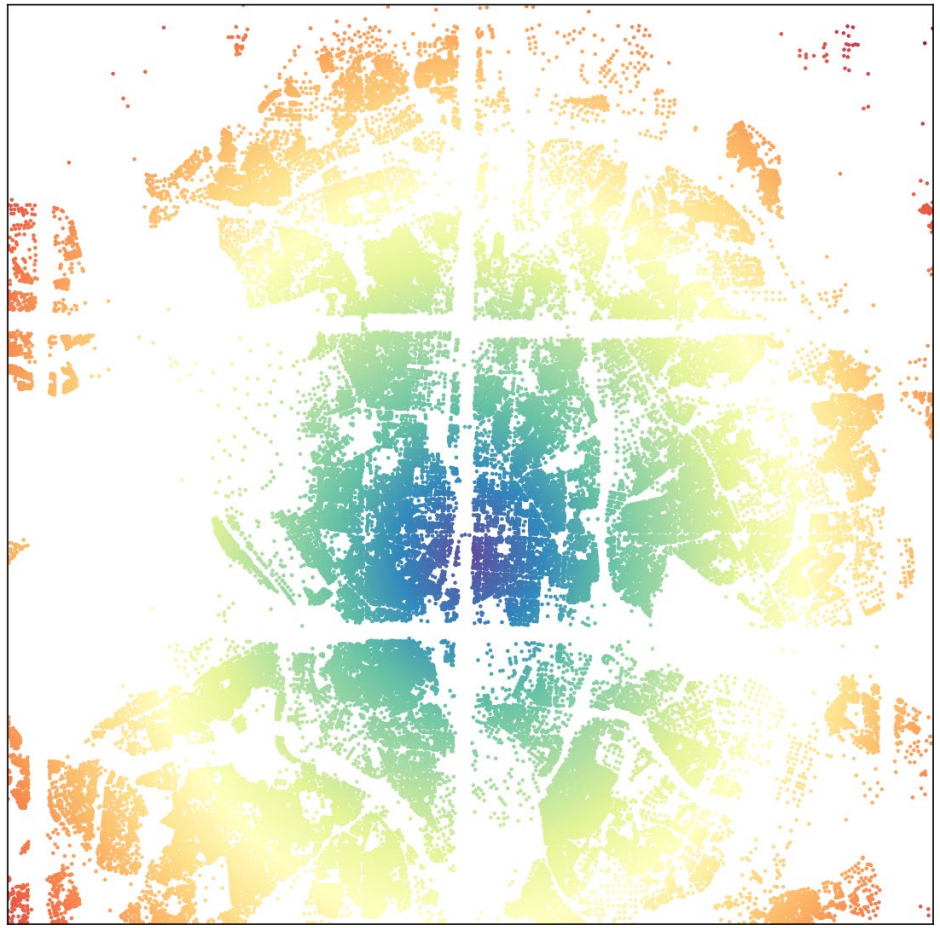
Reducción de la dimensión

- **Análisis PCA**
  - Variables de canales de color
  - **Kernel lineal**
  - De 44 variables a **5 componentes**
  - Mantenemos el **98,7% de la varianza** explicada
  - **Ganancia en rendimiento**, aunque con pérdida de explicabilidad



Construcción de nuevas variables

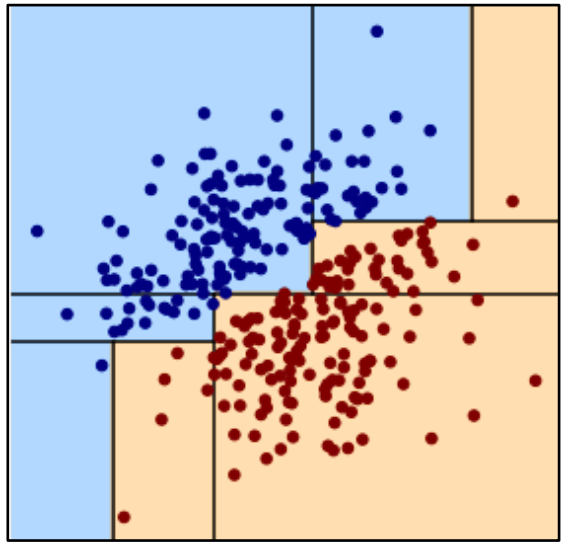
- **Distancia al centro**
  - Estandarización de las coordenadas
  - Diferente distribución de clases entre centro y periferia



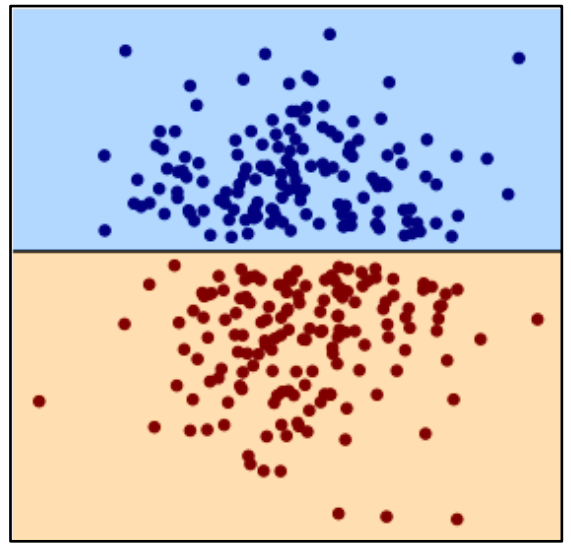
Construcción de nuevas variables

- **Rotación de coordenadas**
  - Algoritmos basados en árboles de decisión (XGBoost)
  - Mejor división de las variables espaciales

Coordenadas sin rotar

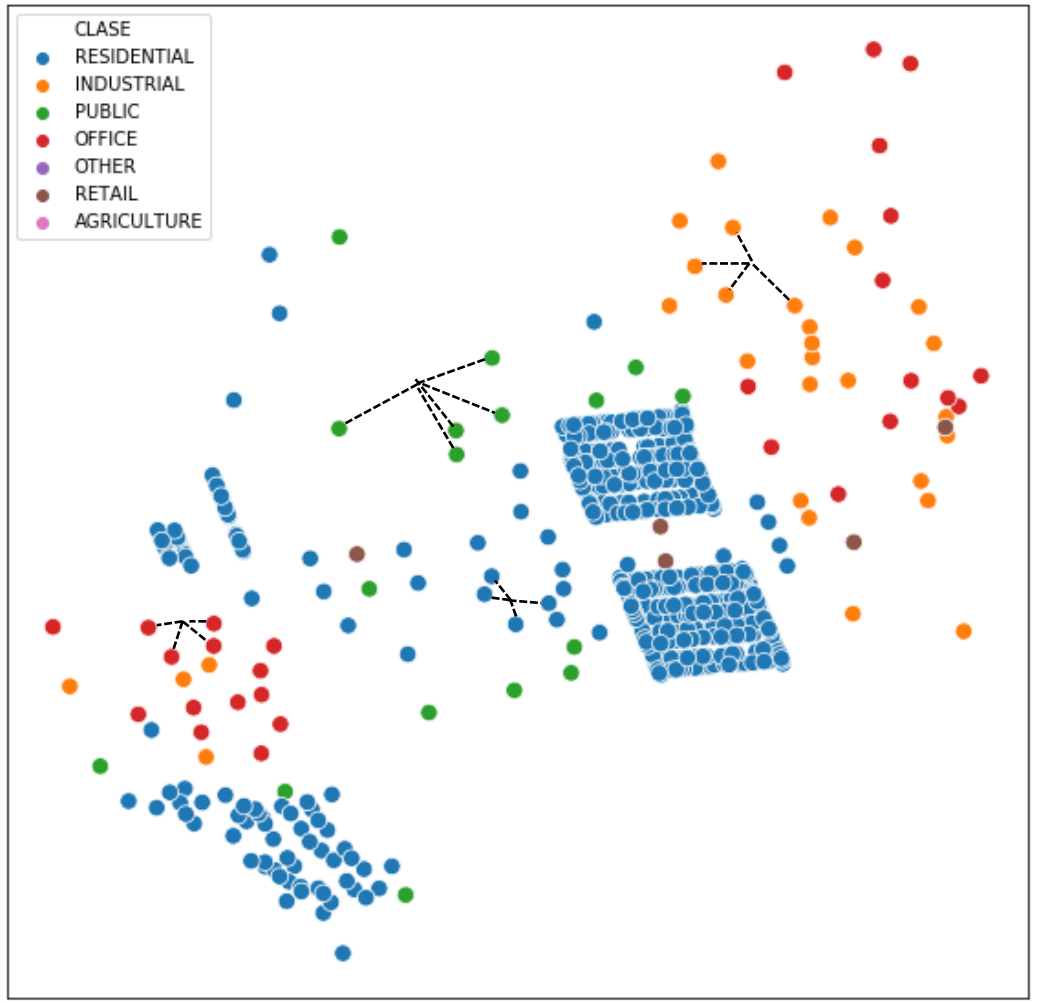


Coordenadas rotadas

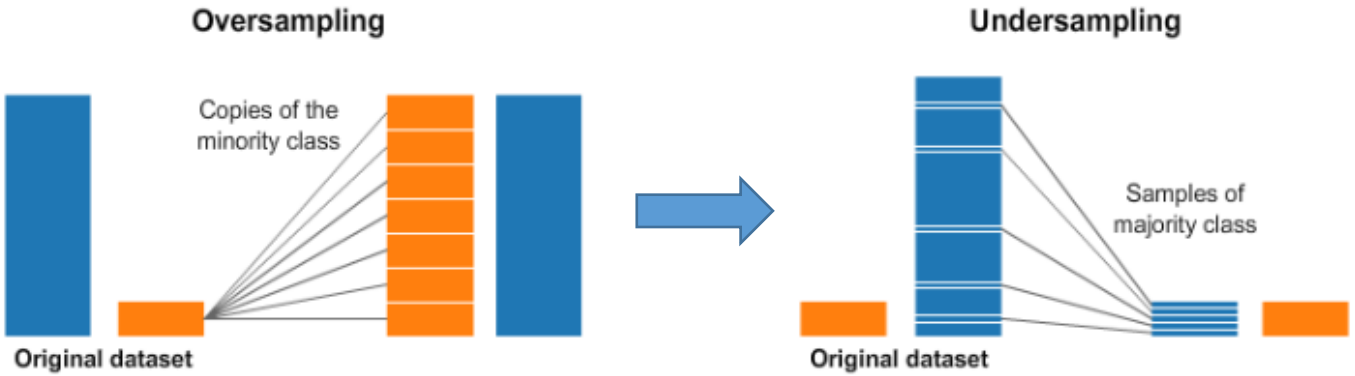


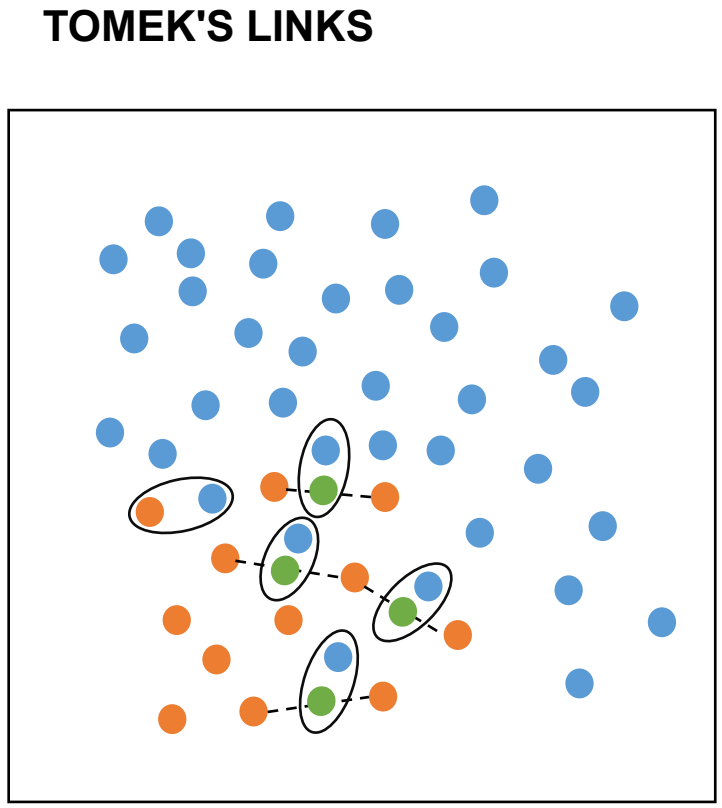
Construcción de nuevas variables – Posición relativa

- **Densidad**
  - Distancia media a los 5 vecinos
  - Diferentes concentraciones según la clase
- **Moda de la altura**
  - Altura más frecuente de los 15 vecinos
  - Las clases tienden a agruparse
  - Difiere mucho entre clases



Equilibrio



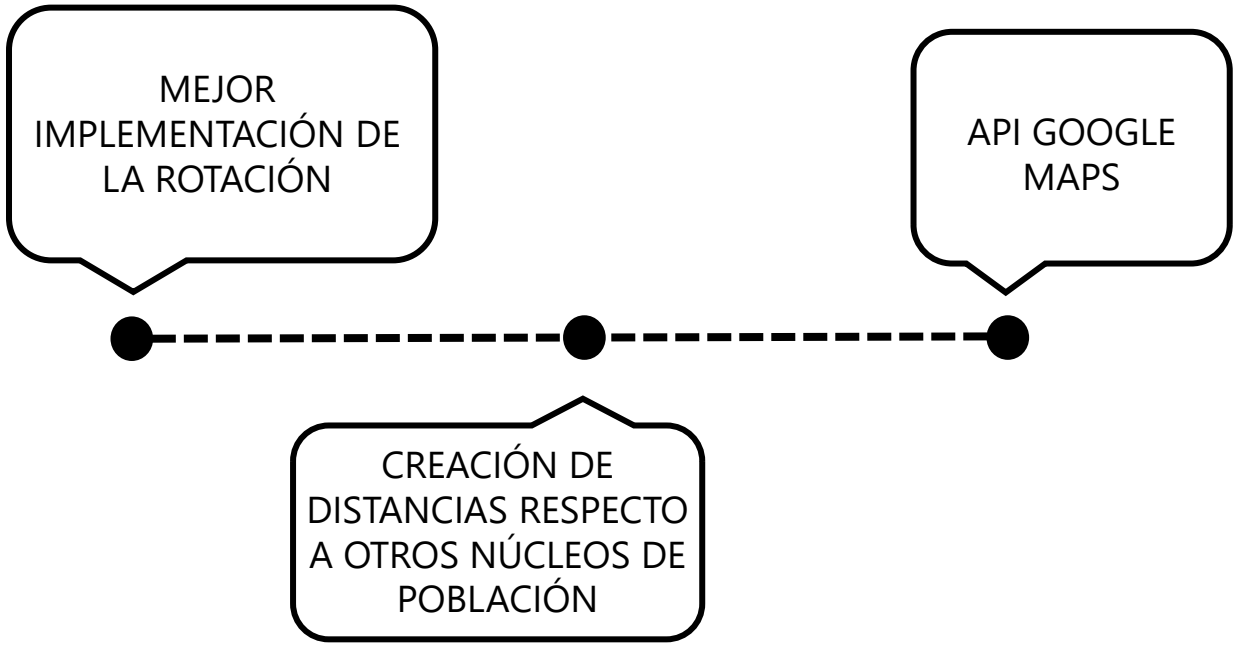


- Combinación de ***oversampling*** **(SMOTE)** y ***undersampling*** **(Tomek's links)**
- Mayor **distanciamiento** entre clases
- Mejor identificación de clases minoritarias

Conclusiones

- **Modelo XGBoost**
- Generación de **nueva métrica**
- Creación de **nuevas variables**:
  - Distancia al centro
  - Rotación de las coordenadas
  - Moda de la altura
  - Densidad
- Equilibrio de clases: *SmoteTomek*
- Optimización bayesiana
- Validación cruzada: 4 *Folds*
- **Métrica final**: 73%

Posibles Mejoras



UNIVERSITYHACK 2020®
DATATHON

La competición de analítica de datos más grande de España.

Muchas gracias.