**COMPETICIÓN DETATHON 2020 – EQUIPO HAKUNA DATATA:**

**[INTRODUCCIÓN MINSAIT LAND CLASSIFICATION](#_Toc34115307)** [2](#_Toc34115307)

[**UTILIZA LA INFORMACIÓN DE LAS IMÁGENES DE SATÉLITE PARA CLASIFICAR EL SUELO** 2](#_Toc34115308)

[**EL OBJETIVO** 2](#_Toc34115309)

[**EL DATASET** 2](#_Toc34115310)

[**Ficheros** 2](#_Toc34115311)

[**Variables** 3](#_Toc34115312)

[**Ámbito geográfico y temporal** 3](#_Toc34115313)

[**Formato y estructura** 3](#_Toc34115314)

[**Dataset respuesta** 3](#_Toc34115315)

[**SE VALORARÁ** 3](#_Toc34115316)

[**AYUDAS AL DESARROLLO DEL RETO** 4](#_Toc34115317)

[**1º CORREO ACLARATORIO** 5](#_Toc34115318)

[**Sobre el reto** 5](#_Toc34115319)

[**Las variables** 5](#_Toc34115320)

[**La entrega intermedia** 5](#_Toc34115321)

[**Sobre las entregas** 5](#_Toc34115322)

# **INTRODUCCIÓN MINSAIT LAND CLASSIFICATION**

En la actualidad, un gran número de satélites orbitan la Tierra captando distintos tipos de imágenes, y no solo se toman en el espectro de luz visible. Estas imágenes se transforman en datos numéricos y tu objetivo será obtener un modelo que sea capaz de predecir qué tipo de terreno se trata.

Para este reto contamos con la colaboración de [Idealista](https://www.idealista.com/data/) como Data Partner.

# **UTILIZA LA INFORMACIÓN DE LAS IMÁGENES DE SATÉLITE PARA CLASIFICAR EL SUELO**

Actualmente, un gran número de satélites toman imágenes con distintos fines y usos. El gran número de imágenes y la gran cantidad de datos que se obtienen de las mismas hace necesario crear modelos predictivos para identificar el contenido de la imagen.

En este reto ya dispondrás de las variables extraídas de la imagen y georeferenciadas, así como variables categóricas asociadas al entorno para estimar un modelo.

# **EL OBJETIVO**

Te retamos a que encuentres el mejor modelo de clasificación automática de suelos en base a las imágenes proporcionadas por el satélite Sentinel II del servicio Copernicus de la Agencia Espacial Europea.

En este reto dispondrás de un conjunto de fincas catastrales asociados a una lista de atributos extraídos de la imagen.

Para ello puedes utilizar las distintas técnicas de Machine Learning disponibles para este tipo de problemas.

La métrica objetivo a maximizar es la “Exactitud”, ([en R](https://www.rdocumentation.org/packages/MLmetrics/versions/1.1.1/topics/Accuracy), [en Python](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html)) definida como el “Número de registros correctamente clasificados / Número total de registros proporcionados por la Organización”.



# **EL DATASET**

El dataset contiene un listado de superficies sobre las que se han recortado la imagen de satélite y se han extraído una serie de características de sus geometrías. Finalmente se ha etiquetado el conjunto de los datos según una clasificación de suelo.

Los datos provienen de distintas fuentes, por lo que son datos de tipo tabular.

## **Ficheros**

* **Dataset "Modelar\_UH2020.txt":** Con este fichero deberás construir un modelo predictivo que permita clasificar el uso del suelo en función de las variables proporcionadas. Constará de 103.230 registros.
* **Dataset "Estimar\_UH2020.txt":** El modelo generado lo aplicarás a los datos de este fichero, de modo que calcules, para cada referencia, la clasificación más probable. Constará de 5.618 registros.

## **Variables**

Los ficheros contienen un total 55 variables: las 3 primeras de ellas relativas a la identificación de los registros y las 8 últimas variables son distintas referencias geométricas y relativas al entorno (geometría del edificio, métricas geométricas generadas automáticamente -GEOM-, metros cuadrados, año construcción y nº de plantas de los edificios del entorno).

Las imágenes satelitales se han tratado y se ha extraído información de 4 canales (R, G, B y NIR), correspondientes a las bandas de color rojo, verde y azul, y el infrarrojo cercano. El valor mostrado corresponde a la intensidad por deciles en cada imagen. Estas variables empiezan con la letra “Q”.

El fichero "Modelar\_UH2020.txt" tiene, adicionalmente como última variable, la clase a predecir.

## **Ámbito geográfico y temporal**

El ámbito geográfico de las imágenes es una zona concreta del municipio Madrid. La referencia (ID) es distinta y representativa de un elemento diferenciado.



## **Formato y estructura**

Los datasets con formato txt tienen como estructura:

* **Nombres de campo**: Incluidos en la cabecera.
* **Separador**: "|".
* **Codificación**: UTF-8.

## **Dataset respuesta**

Se denominará “Equipo\_UH2020.txt” donde Equipo será el nombre del equipo con el que te has inscrito y constará de dos columnas que corresponden a la variable “ID” (identificación de la referencia en “Estimar\_UH2020.txt”) y la variable “CLASE”, que la predicción del tipo de terreno estimado por tu modelo.

# **SE VALORARÁ**

* **La calidad y la técnica utilizada para generar un modelo**  
  Se analizará la técnica analítica utilizada y se compararán objetivamente los valores reales frente a los valores predichos por el modelo. La métrica para maximizar es la “precisión”.
* **Comunicación**  
  Que la documentación interna aportada (códigos, comentarios) esté correctamente expresada y estructurada, y sea reproducible. En la Fase de Presentación de mejores trabajos, el Jurado tendrá en cuenta que el modelado y los resultados obtenidos se transmitan de forma clara y concisa.

# **AYUDAS AL DESARROLLO DEL RETO**

De cara a abordar el reto, y centrándonos en los dos principales lenguajes de programación que nos permiten realizar analítica avanzada, Python y R, os mostramos algunas librerías que os pueden servir de ayuda:

# **1º CORREO ACLARATORIO**

**Sobre el reto**

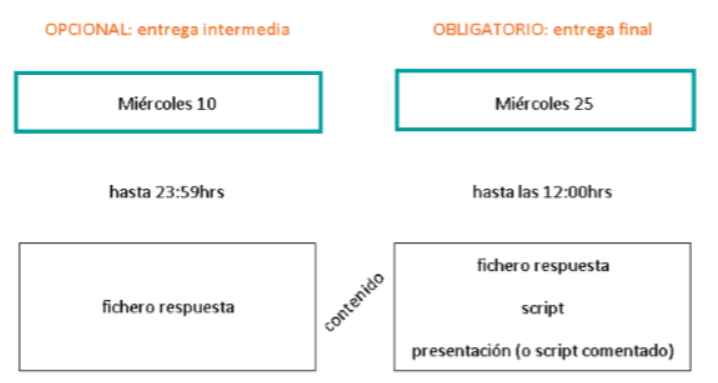
Los datos representan las características de un grupo de referencias catastrales de la zona de Madrid.    
El objetivo es predecir el uso de cada registro propuesto. Es un problema de clasificación multietiqueta: hay 6 categorías no balanceadas.   
Se presentan 2 conjuntos de datos: “Modelar…” contiene 103.230 registros con 56 variables incluyendo la variable objetivo y “Estimar…” contiene 5.618 registros con 55 variables.   
Se permite el enriquecimiento de datos, siempre y cuando se cumplan que los datos sean públicos, accesibles y legales.

**Las variables**

La referencia catastral está ofuscada por lo que no es consultable, solo es válida como identificador de registro.    
La información de longitud-latitud ha sido escalada y desplazada aleatoriamente (manteniendo la relación de posición con el resto de puntos).   
El principal grupo de variables es información sobre los canales de color R-G-B e infrarrojo de las imágenes tomadas por satélite. Hay 11 variables extraídas de cada uno de estos canales, que representan el valor o densidad en cada decil.    
Así, la variable Q\_R\_4\_0\_0 representa el valor para el canal Red en el decil 0, la variable Q\_G\_3\_0\_8 representa el valor para el canal Green del decil 8 etc. La identificación “Q\_NIR…” representa el canal en infrarrojo.   
A continuación, hay 1 variable de área y 4 variables que condensan información de tipo geométrico sobre la parcela/registro.    
El MAXBUILDINGFLOOR se refiere a la altura máxima de los registros colindantes, no al propio registro. El año de construcción también se refiere al entorno.   
Por último, se presenta una variable categórica representativa de la calidad según los datos de catastro (menor valor, menor calidad).

**La entrega intermedia**

Se habilita un día para presentar una “entrega intermedia”, en el que los equipos podrán enviar 2 “ficheros\_respuestas” distintos (es decir, la predicción para los registros de “Estimar…”). No hay que enviar script ni presentación.  La Organización, al día siguiente, comunicará la métrica y la posición relativa del equipo en el total de la competición. Este mecanismo es optativo y no será considerado como una entrega oficial evaluable por el jurado local, tan solo será un medio para retroalimentar y mejorar vuestro trabajo.



**Sobre las entregas**

El trabajo debe ser reproducible, por lo que se debe realizar con herramientas ‘open source’. Se prefieren lenguajes como R o Python. No se permite la importación de modelos pre-entrenados. Todas las etiquetas deben generarse mediante modelos endógenos, caso contrario, el trabajo no califica. Pedimos el fichero\_respuesta + presentación + script\_con\_código.    
Como presentación sirve: pdf, html, rmarkdown, jupiter o notebook con el código muy bien comentado. Si se opta por pdf o similares, máximo 5 slides.   
El script con el código se ejecutará por la Organización para generar el fichero respuesta. Indicar lenguaje y versión.