

# proyectoTID

*Alejandro Casado Quijada y Gustavo Rivas Gervillas*

## Introducción

## Descripción del dataset

Este dataset contiene datos recogidos de la aplicación *PokemonGo*, esta aplicación es un juego de realidad aumentada que emplea el GPS del móvil para principalmente localizar y capturar pokemon en el mundo real. El dataset contiene 296021 muestras cada una de las cuales dispone de los siguientes campos:

- **pokemonId**: el identificador del pokemon, denota su clase.
- **latitude**: latitud de la posición donde se ha localizado el pokemon.
- **longitude**: longitud de la posición donde se ha localizado el pokemon.
- **appearedLocalTime**: momento exacto en el que se encontró el pokemon, con el formato yyyy-mm-ddThh-mm-ss.ms.
- **cellId 90-5850m**: la localización geográfica del pokemon proyectada en una celda S2.
- **appearedTimeOfDay**: momento del día en el que apareció el pokemon (night, evening, afternoon, morning).
- **appearedHour**: hora local de una observación del pokemon.
- **appearedMinute**: minuto local de una observación del pokemon.
- **appearedDayOfWeek**: día de la semana en la que se produjo el avistamiento (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday).
- **appearedDay**: día del avistamiento.
- **appearedMonth**: mes del avistamiento.
- **appearedYear**: año del avistamiento.
- **terrainType**: tipo del terreno donde se avistó el pokemon. Este dato viene dado por un valor número según una tabla de tipos de terreno.
- **closeToWater**: si está el pokemon a 100m del agua o no.
- **city**: ciudad donde se ha visto el pokemon.
- **continent**: continente donde se ha avistado el pokemon.
- **weather**: un string indicando el tiempo que hacía en el momento del avistamiento.
- **temperature**: temperatura en grados Celsius en el momento del avistamiento.
- **windSpeed**: velocidad del viento en el momento del avistamiento km/h.
- **windBearing**: dirección del viento entre 0 y 360 grados.
- **pressure**: presión en el momento del avistamiento en bares.
- **weatherIcon**: el tiempo atmosférico en el momento del avistamiento clasificado según un sistema de categorías más simple que el empleado en *weather* (fog, clear-night, partly-cloudy-night, partly-cloudy-day, cloudy, clear-day, rain, wind).
- **sunriseMinutesMidnight**: tiempo de la aparición relativo al amanecer.
- **sunsetMinutesBefore**: tiempo de la aparición relativo a la puesta de sol.
- **population density**: densidad de población por  $km^2$  en un avistamiento.
- **urbal-rural**: cómo de urbana es la localización donde apareció el pokemon relativa a la *population density* (<200 rural, >= 200 && < 400 midUrban, >= 400 && < 800 subUrban, >800 urban).
- **gymDistanceKm**: distancia al gimnasio más cercano al punto de aparición del pokemon.
- **pokestopDistanceKm**: distancia a la pokestop más cercana al punto de aparición del pokemon.
- **gymIn100m - pokestopIn5000m**: son atributos booleanos que indican si hay un gimnasio o una pokestop a 100m/200m/.../5000m de la localización donde se avistó el pokemon.
- **cooc1 - cooc151**: booleano que indica si el avistamiento de un pokemon coincidió con el de otro (de una clase entre 1 y 151) en un radio de 100m y en un rango de tiempo de 24 horas.
- **class** dice qué pokemon se trata, y en la página del dataset indica que es el atributo a predecir.

## Preprocesamiento

En primer lugar vamos a ver cuántas muestras y atributos tiene nuestro dataset. Además veremos si las clases están balanceadas, para echo emplearemos el comando `xtab`: