proyectoTID

Alejandro Casado Quijada y Gustavo Rivas Gervillas

Introducción

Descripción del dataset

Este dataset contiene datos recogidos de la aplicacion *PokemonGo*, esta aplicación es un juego de realidad aumeta 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 goegráfica 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, Saturdy, 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 aparicion 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`:

```
ds <- read.csv("/mnt/Datos/Master/TID/proyectoTID/300k.csv")
```

```
xtabs(~ class, ds)
```

```
## class
##     1     2     3     4     5     6     7     8     9    10    11    12
##  1368   100    12   738    23     8  1490    99    18  9854   596    95
##    13    14    15    16    17    18    19    20    21    22    23    24
## 27367  1807   199 52114  3290   446 39637  1233 12337   393  4147   146
##    25    26    27    28    29    30    31    32    33    34    35    36
##   516    10  2025    56  3797   205    16  4090   272    17  3565    64
##    37    38    39    40    41    42    43    44    45    46    47    48
##   568     8  1120    25 10143  1035  3603   229    16  7209   227  8325
##    49    50    51    52    53    54    55    56    57    58    59    60
##   251   721    25  1757    29  3950   131  1912    60  1750    32  3897
##    61    62    63    64    65    66    67    69    70    71    72    73
##   253    12  1360   107     7   564    38  3468   204    19  1153   135
##    74    75    76    77    78    79    80    81    82    83    84    85
##  2470   158    20  1210    41  1824    59  1090    40     1   551   302
##    86    87    88    89    90    91    92    93    94    95    96    97
##   692    28   138     5  1344    29  2419   145    12   205 10505   318
##    98    99   100   101   102   103   104   105   106   107   108   109
##  4638   153  1073    34  1786    29   930    29    56    40    46   604
##   110   111   112   113   114   115   116   117   118   119   120   121
##    16  1183    43    22   337    34  2297    84  3842   152  3555    58
##   122   123   124   125   126   127   128   129   130   131   133   134
##   345   307   755   130   202  1404   670  7938    10    20 11740    26
##   135   136   137   138   139   140   141   142   143   147   148   149
##    10    18    14   247     7   262     5    34    73   593    47    24
```