Sumario

1. Resumen

El conjunto de datos generado resume en cifras los partidos de cada jornada de las últimas 20 temporadas de la liga de futbol de 1ª División Española. En él se almacenan desde los nombres de los equipos que participaron hasta la temperatura que hizo ese día entre otros datos relevantes del encuentro.

En cuanto a los orígenes de los datos utilizados en este proyecto:

- **www.bdfutbol.com:** ofrece, acceso a datos referentes a varias ligas de futbol nacionales y extranjeras, ya sea de manera gratuita en su web o pagando para usar su api. Entre sus datos podemos acceder a, entre otros, marcadores de los encuentros, protagonistas (jugadores, entrenadores, árbitros), estadísticas de los jugadores (nº de goles, nº y tipo de tarjetas recibidas, ...). Es por ello, que decidimos basarnos en esta web para nuestro proyecto.

- **Agencia Estatal de Meteorología (AEMet):** proporciona datos sobre temperaturas y precipitaciones, de sus estaciones meteorológicas. Todo ello de forma gratuita y mediante una API, pero que nos permite acceder a los datos que necesitamos de una forma rápida y sencilla. Por esto, decidimos obtener los datos climatológicos de AEMet.

- **Wikipedia:** Se han recuperado datos referentes a los estadios y a las ciudades de los equipos usando esta enciclopedia web.

- **Instituto nacional de estadística (Ine.es):** Proporciona multitud de juegos de datos distintos de manera gratuita, mediante descarga directa. En este caso se ha descargado un fichero con la relación de municipios y códigos por provincias.

- **GitHub:** Herramienta que proporciona repositorios de control de versiones para proyectos, tanto de manera publica como privada. Es en uno de estos repositorios públicos en donde se ha descargado un fichero con las provincias de España con sus códigos.

2. Definir un título para el dataset.

stats_and_clima_1div_es_2000_2020.csv

3. Descripción del dataset.

Tal y como se comentó anteriormente, cada registro del conjunto de datos generado resume en pocos atributos lo más importante de cada encuentro liguero, de cada jornada de las últimas 20 temporadas de la liga de futbol de 1ª División Española (2000-2020), así como los datos meteorológicos para la provincia[1] del encuentro en esa fecha.

4. Representación gráfica.

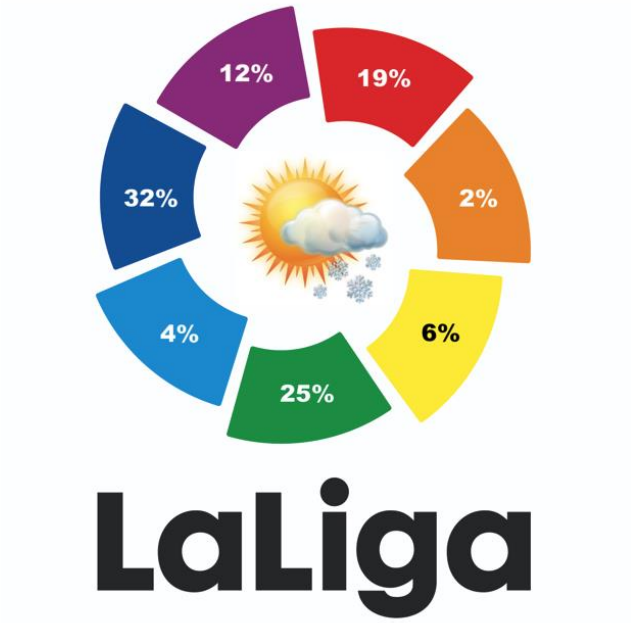

Imagen 1: "La Liga" (1ª división española de futbol), estadísticas y climatología unidas por este dataset.

5. Contenido.

Los datos recuperados para la creación de este dataset se corresponden con las siguientes columnas:

- EquipoAnf: Nombre del equipo anfitrión.
- EquipoVis: Nombre del equipo visitante.
- ResultadoAnf: Goles del equipo anfitrión.
- ResultadoVis: Goles del equipo visitante.
- Fecha: Fecha del encuentro.

1 Debido a la dificultad para poder relacionar los datos de los partidos con los climatológicos (se explica en el apartado contenido) es importante remarcar que tanto temperatura como precipitaciones son medias de la provincia, resultando en un proceso menos costoso de "matching".

- Estadio: Nombre del estadio.
- Ciudad: Nombre de la ciudad del encuentro.
- Arbitro: Nombre del arbitro del encuentro.
- Temp[2]: Temperatura media del día (ºC).
- Prep[3]: Media de las precipitaciones del día (ml)

El periodo de tiempo al que hacen referencia los datos comprende desde el año 2000 al 2020.

Para obtener los datos futbolísticos, se aplicaron técnicas de "Web Scraping" sobre la página *bdfutbol.com*. Para tomar esta decisión, se tuvo en cuenta, después de analizar la página, que no hay disponible una Api libre (aunque si de pago) y que los datos se pueden ver, de forma gratuita, en la web. Como observaciones, cabe mencionar que se comprobó que el archivo "*robot.txt*" no limitaba los accesos a las rutas de donde se obtuvieron los datos y que para evitar saturar el servidor de con las peticiones, estas se espaciaron en el tiempo aplicando un retraso entre petición y petición (delay).

Por otro lado, los datos de los estadios (equipo, estadio, ciudad) se han recuperado de la *Wikipedia*[4], usando también técnicas de "Web Scraping", creando un diccionario y añadiendo los datos, usando el nombre del equipo anfitrión como nexo. Estos datos también existen en "*bdfutbol*", sin embargo, se decidió recuperar estos datos de otra fuente, con el objetivo de no sobrecargar de peticiones a la página.

En cuanto a los datos referentes a la climatología, se obtuvieron mediante el uso de la API Rest de la Agencia Estatal de Meteorología (AEMet OpenData[5]). AEMet proporciona, mediante su API, parte de la información de la cual es propietaria[6].

Para obtener las mediciones climatológicas, hemos creado un script con Python que nos permite obtener los datos mediante llamadas al "endpoint" Valores Climatológicos Diarios de Todas las Estaciones (JSON). Los datos se obtienen con una limitación de 31 días.

Es importante tener en cuenta que el uso de la API obliga a solicitar una "*API Key*"[7], asociada a una cuenta de correo, que debe ser utilizada en cada llamada.

Finalmente, para la integración de los dos dataset nos hemos encontrado el siguiente problema, el JSON resultante tiene información recogida de cada una de las estaciones meteorológicas de España, sin embargo, solo se dispone de la información discretizada por provincia y nombre (nombre de la zona, pueblo, villa, ciudad, etc. donde está localizada la estación), que no siempre

2 Ver nota 1
3 Ver nota 1
4 https://es.wikipedia.org/wiki/Anexo:Estadios_de_f%C3%BAtbol_de_Espa%C3%B1a
5 https://opendata.aemet.es/centrodedescargas/inicio
6 ©AEMET: "Información elaborada por la Agencia Estatal de Meteorología"
7 https://opendata.aemet.es/centrodedescargas/altaUsuario?

coincide con el nombre de una ciudad (como se necesita). Se optó por la conexión mediante provincia por su menor coste en tiempo de ejecución y de desarrollo. Sin embargo, los datos cruzados de la wikipedia y bdfutbol no disponian de la provincia, unicamente se tenia la ciudad, ha sido necesario, cruzar estos datos con 2 dataset obtenidos del INE y de Github, uno con los datos de los pueblos y ciudades de España con sus códigos de provincia y el otro con las provincias con sus códigos, respectivamente.

Como posible mejora[8], se podría solucionar esto y hacer un "matching" más preciso, mediante el uso de la librería *Python* "GeoPy", ya que podríamos obtener la ciudad asociada a una medición, simplemente pasando como parámetros el nombre de localización de la estación, provincia y país. Aunque el alto coste en tiempo de ejecución, así como limitaciones de la API utilizada por la librería, nos obligarían a dedicar mayor tiempo y recursos para su consecución.

6. Agradecimientos.

Agradecemos a www.bdfutbol.com por publicar los datos recogidos en esta practica.

Agradecemos a la Agencia Estatal de Meteorología que pongan a disposición pública dichos datos.

7. Inspiración.

Los datos recogidos en este dataset podrían ser utilizados tanto con fines lúdicos como lucrativos. Por ejemplo, personas aficionadas al futbol o simplemente a la estadística, podría hallar interesante este conjunto de datos con el cual, no solo consultar los datos más relevantes de un partido, sino tratar de dar respuesta mediante la minería de datos a la pregunta: ¿Podría encontrar los factores pudieron influir en las derrotas de mi equipo el año pasado?
Por otro laso, en el terreno lucrativo, se podrían elaborar modelos predictivos que ayudasen a realizar apuestas deportivas, o dentro del periodismo deportivo, la figura cada vez más extendida del analista deportivo que trata de dar claves de resultados o rachas deportivas de un equipo, entre otras muchas cosas.

8 En el repositorio se puede acceder a la prueba realizada con la librería GeoPy (get_mediciones_clima_aemet_rango_fechas_test)

8. Licencia.

La licencia elegida para nuestro dataset es **CC BY-SA 4.0 License**.

Las razones que nos han llevado hasta esta licencia son:

a) Permite que el dataset sea adaptado o utilizado como base para otros, según se necesite y se puede compartir.
b) No impide que sea utilizado con fines comerciales.
c) A cambio de su uso, se debe mencionar siempre al licenciante del dataset
d) Cualquier resultado cuyo origen sea este dataset, debe ser distribuido bajo la misma licencia del original.

9. Código.

El código creado para dar solución a la práctica, así como el documento pdf con las respuestas, se encuentran publicadas en el repositorio público de GitHub: https://github.com/hulkolarry/TCVD-Practica1

10. Dataset.

11. Anexos:

https://www.bdfutbol.com/robots.txt

```
User-agent: *
Disallow: /c/archivesko.html
Disallow: /c/archivesok.html
Disallow: /c/apiko.html
Disallow: /c/apiok.html
Disallow: /es/c/archivesko.html
Disallow: /es/c/archivesok.html
Disallow: /es/c/apiko.html
Disallow: /es/c/apiok.html
Disallow: /en/c/archivesko.html
Disallow: /en/c/archivesok.html
Disallow: /en/c/apiko.html
Disallow: /en/c/apiok.html
Disallow: /bo
```

12. Recursos

Se han utilizado los siguientes recursos para la realización de la práctica:

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Scraping the Data.

13. Tabla de contribuciones:

Contribuciones	Firma
Investigación previa	*rcotillas*, *aruizpla*
Redacción de las respuestas	*rcotillas*, *aruizpla*
Desarrollo código	*rcotillas*, *aruizpla*