Máster en Ciencia de Datos

Asignatura: Tipología y ciclo de vida de los datos

Práctica1 – WEB Scraping

Autores:

Andrea Marcos Vargas
Felipe Eugenio García GonzáleZ

1. Contexto

La FEB (Federación Española de Baloncesto) tiene una web, https://competiciones.feb.es/estadisticas/, donde se proporcionan estadísticas, resultados, calendarios y *rankings* de todas las competiciones de baloncesto del país, incluyendo datos de las competiciones base por Comunidades y de temporadas anteriores. Los datos no están disponibles para ser descargados como *Open Data,* sino que se muestran en tablas.

Puede ser interesante recopilar estos datos y dejarlos en un repositorio para su posterior consulta o utilización a nivel educativo, profesional o por curiosidad. A lo largo de los años y a medida que el baloncesto ha ido quemando etapas, el seguimiento estadístico de este deporte ha servido para aumentar su nivel. Las estadísticas en baloncesto son importantes, por ejemplo para determinar qué factores son determinantes en la consecución de una victoria.

Según los estudios de Ibáñez y col, (2003) y Sampaio y Janeira (2003), los equipos que anotan más tiros de 2 y capturan más rebotes defensivos ganan más partidos. Según la tesis de Sampaio y Leite (2003), estos resultados se podrían explicar por una idea defensiva de presionar a tu adversario para que no esté cómodo atacando y que falle sus tiros y, simultáneamente, por una idea ofensiva de canalizar el juego hacia posiciones más cercanas de la canasta del rival. Estos autores, analizaron en el año 2006 las estadísticas de los equipos que participaron en el Eurobasket 2005, y llegaron a la conclusión de que las estadísticas más discriminantes entre vencedores y perdedores fueron los tiros de 3 puntos fallados y las asistencias. Es decir, los equipos ganadores planteaban situaciones ofensivas para poder desequilibrar a la defensa, en las que el último pase conduce a un tiro de elevadas probabilidades de acierto, minimizaban riesgos en el tiro exterior eligiendo las mejores opciones de lanzamiento, la mayoría de ellas después de una circulación de balón correcta. En la línea general de los autores anteriores, un estudio de Hierro (2002), tras analizar la liga ACB y la NBA, concluye que los equipos con mejores porcentajes de tiros de 2 y de 3, mayor número de faltas recibidas, tapones y asistencias, obtenían más victorias; no considerando el número de pérdidas de balón y los tiros libres como aspectos definitivos.

La web de la FEB es https://www.feb.es, al ejecutar la función whois para ver el propietario de las webs, en ambos casos aparecen todos los campos vacíos, lo que indica que la FEB ha optado por no proporcionar información de registro de dominio. Sin embargo, al utilizar la herramienta BuiltWith se puede determinar que el sitio web de la FEB utiliza tecnologías como Microsoft ASP.NET, Handlebars, Moment.js y jQuery, y está alojado en un servidor IIS con sistema operativo Windows Server. Además, utiliza Google AdSense como red de publicidad.

whois.whois("https://www.feb.es/")

{'domain_name': None,
'registrar': None,
'whois_server': None,

'referral_url': None,
'updated_date': None,
'creation_date': None,
'expiration_date': None,
'name_servers': None,
'status': None,
'emails': None,
'dnssec': None,
'name': None,
'org': None,
'address': None,
'city': None,
'state': None,
'registrant_postal_code': None,
'country': None}

builtwith.builtwith(url)

{'web-servers': ['IIS'],
'operating-systems': ['Windows Server'],
'advertising-networks': ['Google AdSense'],
'javascript-frameworks': ['Handlebars', 'Moment.js', 'jQuery'],
'web-frameworks': ['Microsoft ASP.NET']}

2. Título

El título elegido es: **Estadísticas del baloncesto español, 2012-2022** .

3. Descripción del dataset

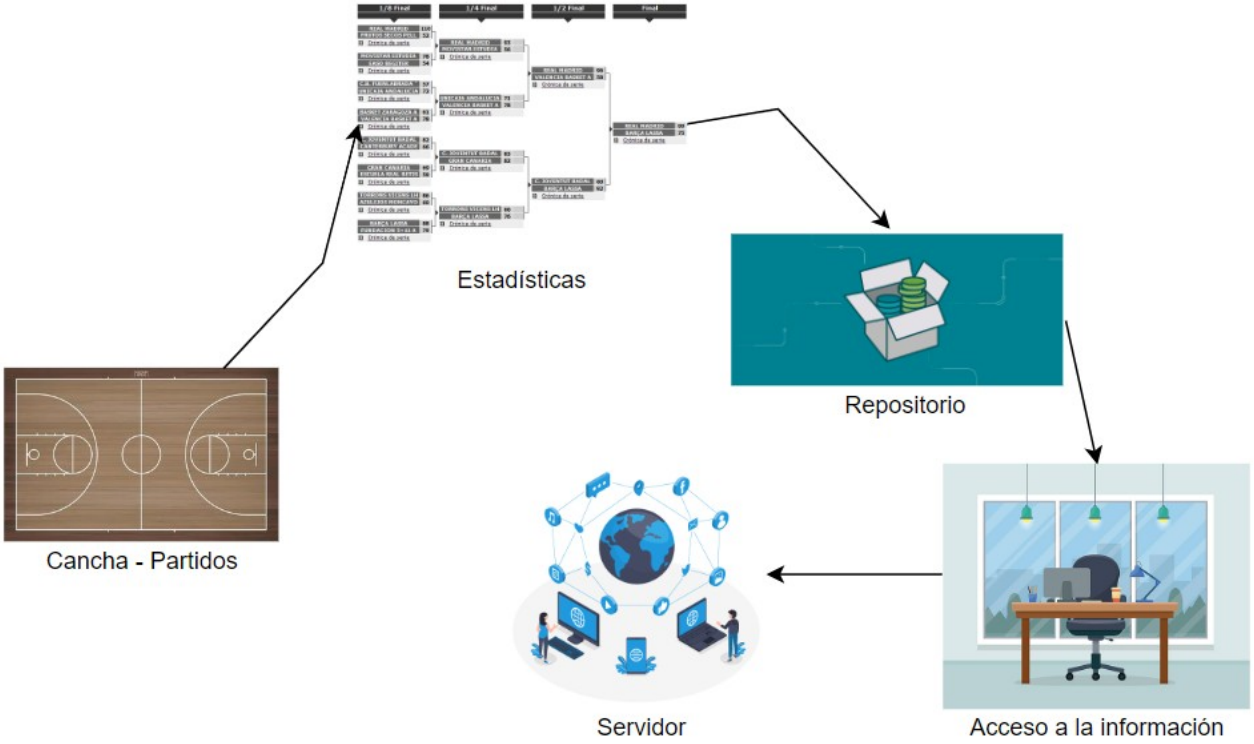
El dataset que se menciona en este contexto se trata de un conjunto de datos que contiene información detallada y estadísticas de los equipos que han participado en las ligas de baloncesto **LF Endesa**, **Liga LEB Oro** y **Liga LEB Plata** durante los últimos 10 años completos. Esto significa que el dataset incluye información desde la temporada regular de la Liga 2012-2013 hasta la temporada regular de la Liga 2021-2022.

Este conjunto de datos es una herramienta muy valiosa para aquellos interesados en el análisis estadístico del baloncesto. Los datos están disponibles en un formato estructurado y organizado que permite su fácil manipulación y análisis. El dataset proporciona información sobre una amplia variedad de estadísticas y métricas relacionadas con el rendimiento de los equipos y jugadores, como por ejemplo, puntos anotados, rebotes, asistencias, tiros de campo, porcentaje de acierto en tiros de campo, tiros libres, entre otras.

Además, el conjunto de datos también incluye información sobre el rendimiento de los equipos en diferentes aspectos del juego, como en defensa y en ataque, lo que permite realizar análisis detallados y comparativos sobre el desempeño de los equipos y los jugadores a lo largo del tiempo.

En resumen, este conjunto de datos es una herramienta muy útil para aquellos que buscan estudiar el baloncesto desde una perspectiva analítica, ya sea para fines educativos, profesionales o simplemente por curiosidad. Los datos están disponibles en un formato fácilmente accesible y organizado, lo que permite una fácil manipulación y análisis de los mismos.

4. Representación gráfica



5. Contenido

Como ya se indicó en la descripción, el *dataset está compuesto por datos y estadísticas de los equipos de las ligas de baloncesto* ***LF Endesa, Liga Leb Oro, LF Chagenger, Liga Leb Plata y Liga Eba****,* en sus temporadas regulares de los últimos 10 años completos, esto es desde la temporada 2012-2013 hasta la temporada 2021-2022. El dataset obtenido tiene un tamaño de 22 columnas y 458 filas

Los datos y estadísticas disponibles para cada equipo, con la nomenclatura utilizada en el dataset son:

- EQUIPO: Donde se indica el equipos al que pertenecen las estadísticas.
- Cantidad de partidos jugados: PART,
- Minutos jugados en total por el equipo: MIN
- Puntos Anotados por todos los jugadores del equipo: PT
- Puntos anotados con canastas de dos puntos: T2
- Puntos anotados con canastas triples: T3
- Puntos anotados con tiros de campo: TC
- Puntos anotados con tiros libres: TL
- Total Rebotes Ofensivos : RO
- Total Rebotes Defensivos: RD
- Rebotes totales: RT
- Total Asistencias: AS
- Total Balones Robados: BR
- Total Balones Perdidos: BP
- Tapones a Favor: TF
- Tapones en Contra: TC
- Mates: MT
- Faltas Cometidas: FC
- Faltas Recibidas: FR
- Valoración: VA

6. Propietario

La información del propietario de la página web de la Federación Española de Baloncesto (FEB) no está disponible públicamente en su sitio web ni en los registros de dominio WHOIS. Sin embargo, la FEB es una organización deportiva nacional que se encarga de gestionar las competiciones de baloncesto en España, y es responsable de la administración y organización de los equipos, jugadores y entrenadores de baloncesto del país.

Consideramos entonces que el propietario debe ser la propia Federación Española de Baloncesto, aunque en la web no se indica nada al respecto.

```
whois.whois("https://www.feb.es/")
```

```
{'domain_name': None,
'registrar': None,
'whois_server': None,
'referral_url': None,
'updated_date': None,
'creation_date': None,
'expiration_date': None,
'name_servers': None,
'status': None,
'emails': None,
'dnssec': None,
'name': None,
'org': None,
'address': None,
'city': None,
'state': None,
'registrant_postal_code': None,
'country': None}
```

7. Inspiración

El conjunto de datos que se compone de información y estadísticas de los equipos de las ligas de baloncesto LF Endesa, Liga LEB Oro y Liga LEB Plata en sus temporadas regulares de los últimos 10 años completos, es interesante por varias razones.

En primer lugar, el baloncesto es uno de los deportes más populares en todo el mundo y su seguimiento estadístico ha demostrado ser de gran importancia para el análisis y comprensión del juego. Los datos recopilados en este conjunto de datos permiten conocer aspectos clave de los equipos, como su rendimiento en casa y fuera, los jugadores más destacados y su desempeño individual, los estilos de juego preferidos por los entrenadores y los resultados finales de las competiciones.

En segundo lugar, estos datos pueden ser utilizados para realizar análisis y estudios sobre el rendimiento y estrategias de los equipos, lo que puede resultar útil tanto para los entrenadores como para los jugadores. Estos datos también pueden ser de interés para periodistas, aficionados y otras personas interesadas en el mundo del baloncesto.

Las preguntas que se pueden responder utilizando este conjunto de datos son diversas, algunas de las cuales son:

- ¿Cuáles son los equipos con mejor rendimiento en casa o fuera de casa?

- ¿Qué jugadores han sido los más destacados en cada temporada?
- ¿Cuáles son los estilos de juego más utilizados por los equipos?
- ¿Existen patrones en los resultados de las competiciones a lo largo de las temporadas?
- ¿Cuál es el desempeño de los equipos en términos de anotación, defensa, rebotes, asistencias y otras métricas clave?
- ¿Existen diferencias significativas en el rendimiento de los equipos de las diferentes ligas?

8. Licencia

La selección de una licencia adecuada para el dataset resultante es una tarea importante ya que define los términos y condiciones de uso del dataset por parte de terceros. Existen diferentes tipos de licencias que se pueden utilizar, pero es importante seleccionar una que sea coherente con los objetivos del proyecto y que permita su reutilización sin restricciones injustas.

Una licencia adecuada para el dataset resultante podría ser la Licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0). Esta licencia permite a otros distribuir, remezclar, adaptar y construir sobre el dataset en cualquier medio o formato, incluso para fines comerciales, siempre y cuando se otorgue el crédito adecuado al propietario original del dataset.

La elección de esta licencia se justifica por varios motivos. En primer lugar, permite la reutilización del dataset con fines comerciales y no comerciales, lo que puede ser importante para la promoción del baloncesto y la investigación en este ámbito. Además, esta licencia permite la adaptación y creación de obras derivadas, lo que puede promover la innovación y el desarrollo de nuevas herramientas y técnicas para el análisis de datos de baloncesto.

En segundo lugar, la licencia CC BY 4.0 es ampliamente utilizada y reconocida, lo que puede facilitar su comprensión y adopción por parte de terceros interesados en utilizar el dataset. Además, al requerir la atribución adecuada al propietario original del dataset, se garantiza la integridad de la información y se evita la apropiación indebida de los datos.

En resumen, la Licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0) sería una buena opción para la distribución del dataset resultante, ya que permite su reutilización con fines comerciales y no comerciales, la creación de obras derivadas y garantiza la atribución adecuada al propietario original del dataset.

9. Código

El user-agent es ***'Mozilla/5.0 (X11; Linux x86_64; rv:109.0) Gecko/20100101 Firefox/111.0'***

El código se encuentra disponible en el repositorio GitHub https://github.com/fgarciagonzalez2/TD_Practica1.git

Las librerías y versiones utilizadas son:

aiofiles==22.1.0 ; python_version >= '3.7' and python_version < '4.0'
aiosqlite==0.19.0 ; python_version >= '3.7'
anyio==3.6.2 ; python_full_version >= '3.6.2'
argon2-cffi==21.3.0 ; python_version >= '3.6'
argon2-cffi-bindings==21.2.0 ; python_version >= '3.6'
arrow==1.2.3 ; python_version >= '3.6'
asttokens==2.2.1
async-generator==1.10 ; python_version >= '3.5'
attrs==23.1.0 ; python_version >= '3.7'
babel==2.12.1 ; python_version >= '3.7'
backcall==0.2.0
beautifulsoup4==4.12.2 ; python_full_version >= '3.6.0'
bleach==6.0.0 ; python_version >= '3.7'
bs4==0.0.1
builtwith==1.3.4
certifi==2022.12.7 ; python_version >= '3.6'
cffi==1.15.1
charset-normalizer==3.1.0 ; python_full_version >= '3.7.0'
comm==0.1.3 ; python_version >= '3.6'
debugpy==1.6.7 ; python_version >= '3.7'
decorator==5.1.1 ; python_version >= '3.5'
defusedxml==0.7.1 ; python_version >= '2.7' and python_version not in '3.0, 3.1, 3.2, 3.3, 3.4'
distlib==0.3.6
exceptiongroup==1.1.1 ; python_version < '3.11'
executing==1.2.0
fastjsonschema==2.16.3
filelock==3.12.0 ; python_version >= '3.7'
fqdn==1.5.1
future==0.18.3 ; python_version >= '2.6' and python_version not in '3.0, 3.1, 3.2, 3.3'
h11==0.14.0 ; python_version >= '3.7'
idna==3.4 ; python_version >= '3.5'
importlib-metadata==6.5.0 ; python_version < '3.10'
iniconfig==2.0.0 ; python_version >= '3.7'
ipykernel==6.22.0 ; python_version >= '3.8'
ipython==8.12.0 ; python_version >= '3.8'

ipython-genutils==0.2.0
isoduration==20.11.0
jedi==0.18.2 ; python_version >= '3.6'
jinja2==3.1.2 ; python_version >= '3.7'
json5==0.9.11
jsonpointer==2.3
jsonschema==4.17.3 ; python_version >= '3.7'

El código tiene un punto de entrada principal en el módulo **main.py.** Primero se realiza una llamada a los procedimientos internet_ok y response_ok de la clase Enlace, este código se desarrolla en el módulo enlace_app.py, con esto comprobamos que tenemos conexión a internet y que la página web es accesible. Posteriormente se realiza una llamada a una instancia de la clase WebScrap cuyo código se desarrolla en el módulo bal_est.py. Todos los módulos excepto main.py están en la subcarpeta bal_est.

Dentro de la subcarpeta bal_est tenemos el resto de módulos.

- El indicado **enlace_app.py** para las conexiones.
- El módulo **config.py** donde configuramos los valores de las variables, de entre ellas destacamos la booleana “obtener_medias”, por defecto está en *False*, si la configuramos como *True* el programa nos extrae las medias de las estadísticas. También tenemos tres listas con los nombres de las competiciones, esto es así porque de año a año varían, son las misma competición, pero a veces se escribe con tilde, otras no, unas veces se le llama “Liga Regula Único”, otras “Liga Regular Este”, etc.
- El módulo **bal_est.py**, es donde hacemos el web scraping sobre la web elegida. El procedimiento se ha realizado con **selenium.** Se ha creado una clase WebScrap con varios procedimientos, primero extraemos las cabeceras de los datos, después tenemos un procedimiento para cada una de las tres ligas de donde vamos a extraer los datos. Para la extracción de datos se va seleccionando mediante selenium la página en concreto donde se encuentra la tabla, para ellos primero hay que seleccionar la liga (LF Endesa, LEB Oro y LEB Plata), después elegimos la temporada desde un desplegable, y después elegimos de otro desplegable la competición en concreto. Para obtener los datos de las últimas 10 temporadas tenemos que ir repitiendo el procedimiento hasta obtener los datos de las últimas 10 temporadas completas para cada una de las tres ligas.
- El módulo **datos.py** es donde convertimos los datos extraídos mediante el módulo bal_est-py en un único archivo csv, y realizamos el guardado en el disco duro.

10. Dataset

Se adjuntan los datos en formato CSV en la carpeta dataset del repositorio. Los campos de separación son ‘,’ y el nombre del dataset es, siendo el nombre del mismo “***baloncesto_est.csv***”.

El dataset se ha publicado en Zenodo el DOI es **10.5281/zenodo.7799717** y el enlace del mismo es https://zenodo.org/record/7799717#.ZD7n1HbMI2w.

11. Vídeo

Los videos se han incorporado al drive en las siguientes direcciones:

Contribuciones	Firma
Investigación previa	F.E.G.G. - A.M.V.
Redacción de las respuestas	F.E.G.G. - A.M.V.
Desarrollo del código	F.E.G.G. - A.M.V.
Participación en el video	F.E.G.G. - A.M.V.

Referencias

https://www.fbcv.es/blog/2018/06/la-influencia-la-estadistica-juego-formacion/

http://investigaciones.uniatlantico.edu.co/revistas/index.php/REDFIDS/article/view/3174

https://www.todamateria.com/basquetbol/

Wikipedia (2023). https://en.wikipedia.org/wiki/Spider_trap (Consultado: 11 de marzo de 2023)