Solución PRA1 – Web Scraping

NBA season players stats

Tipología y Ciclo de Vida de los Datos

David Esparcia

Rubén Revuelta

1. Contexto

El dataset generado a partir de técnicas de “web scraping” recoge datos sobre sobre las estadísticas de los jugadores de la liga de baloncesto estadounidense conocida como NBA durante cada temporada. Estos datos han sido obtenidos aplicando técnicas de “Web Scraping” en el sitio web “https://www.basketball-reference.com”. Este sitio web pertenece al grupo de webs Sport Reference que aglutinan métricas sobre diferentes licas locales estadounidenses de diversos deportes.

2. Imagen identificativa

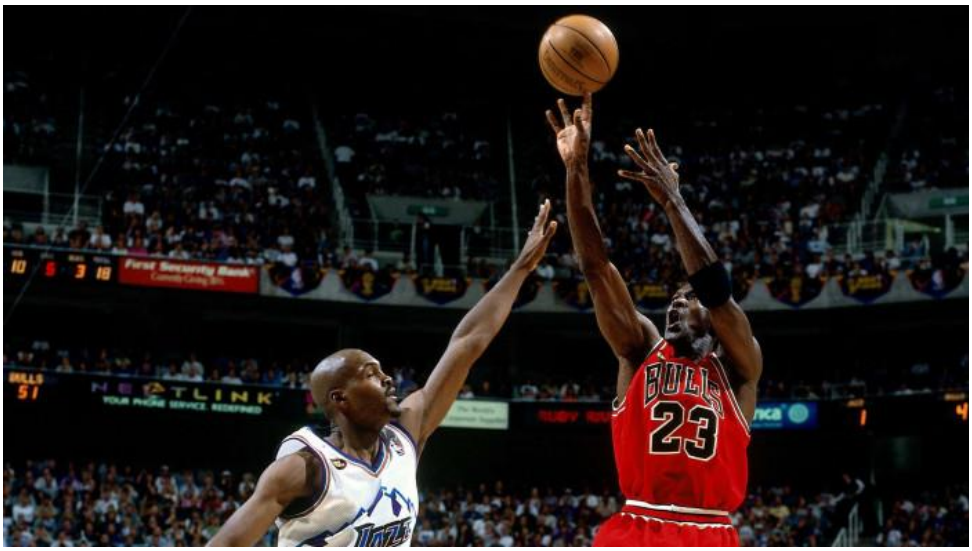

Ilustración 1: Michael Jordan, jugador histórico de la NBA, tirando a canasta.

3. Descripción del dataset

El dataset recogido contiene estadísticas individuales de los jugadores de la NBA durante el transcurso tanto de la fase regular como de la fase de “play offs” de la liga. Estas estadísticas representan diferentes aspectos básicos del juego habituales en este deporte. El dataset recoge de forma general las estadísticas, a final de temporada, de la media por partido en diferentes aspectos del deporte. En concreto, se recogen datos estadísticos del rendimiento de los jugadores desde la temporada 1946-47.

4. Contenido

Como se ha mencionado en el apartado anterior, el dataset contiene estadísticas de los jugadores desde la temporada 1946-47 de la NBA. A continuación, en la siguiente tala se muestran los campos tratados en el conjunto de datos.

CAMPO	DESCRIPCIÓN	EJEMPLO
Player	Nombre del jugador	Norm Baker
Pos	Posición del jugador dentro del campo. Si el jugador puede ocupar diferentes posiciones estas se encuentran separadas por "-"	G-F
Age	Edad	27
Tm	Equipo en el cual consiguió dichas estadísticas	DTF
G	Número de partidos jugados.	75
GS	Número de partidos que comenzó jugando de inicio.	28
MP	Media de minutos jugados por partido.	23.6
FG	Media de tiros de campo acertados por partido.	3.6
FGA	Media de intentos de tiros de campo por partido.	8.7
FG%	Porcentaje de tiros de campo acertados por partido	.439
3P	Media de tiros de tres puntos por partido	0.8
3PA	Media de intentos de tiros de tres puntos por partido	2.4
3P%	Porcentaje de tiros de tres puntos acertados por partido.	0.359
2P	Media de tiros de dos puntos por partido.	2.9
2PA	Media de intentos de tiros de dos puntos por partido.	6.1
2P%	Porcentaje de tiros de dos puntos acertados por partido.	.468
eFG%	Porcentaje efectivo de tiros de campos	.713
FT	Media tiros libres acertados por partido.	4.6
FTA	Media de tiros libres intentados por partido.	6.7
FT%	Porcentaje efectivo de tiros libres por partido.	.690
ORB	Rebotes ofensivos por partido.	3.7
DRB	Rebotes defensivos por partido.	11.0
TRB	Media total de rebotes por partido.	14.7
AST	Asistencias por partido.	1.1
STL	Robos por partido.	0.7
BLK	Bloqueos por partido.	2.1
TOV	Pérdidas de balón por partido.	1.2
PF	Faltas personales por partido.	2.1
PTS	Puntos por partido.	9.1

5. Propietario

Como hemos comentado brevemente en la introducción, los datos pertenecen a Sport Reference cuya actividad principal es la aglomeración de estadísticas de diferentes deportes y sus ligas asociadas en Estados Unidos. Entre ellos se encuentra la web "Basketball-Reference" la cual aglutina diferentes tipos de estadísticas sobre la NBA a lo largo de los años.

Respecto al uso de los datos facilitados por Sport-Reference, tal y como se indica en sus términos de uso, permiten el uso y publicación de los datos siempre y cuando se cumplan bajo sus términos. A continuación, se destacan los puntos principales que pueden afectar de alguna manera al desarrollo de la práctica:

- Se identifique a Sport Reference como fuente y propietario original de los datos.
- Los datos extraídos no sean utilizados para el desarrollo de ningún tipo de herramienta o servicio que pretenda o pueda competir con la actividad desarrollada por Sport Reference.
- Se permite el uso de técnicas de Web Scraping siempre y cuando estas no causen un perjuicio a la disponibilidad o rendimiento de los servicios que Sport-Reference ofrece. Sport-Reference se reserva el derecho de bloqueo ante la detección de técnicas agresivas de obtención de datos.

Para cumplir con estos términos éticos y legales, durante el desarrollo de la práctica se han tomado las siguientes medidas:

- Tanto en la publicación de los datos recogidos como en la presente memoria, se reconoce a Sport-Reference como legítimo propietario de los datos.
- Los datos han sido extraídos con propósitos únicamente educacionales y sin la intención alguna de desarrollar cualquier tipo de producto o servicio, que hiciera uso de los datos, bajo el desarrollo de la actividad educativa.
- Se ha ejecutado la actividad de Web Scraping cumpliendo con los límites permitidos evitando así hacer un uso abusivo y agresivo del recurso web. Para se ha realizado la actividad respetando los umbrales fijados por Sport Reference de menos de 20 peticiones por minuto.

6. Inspiración

Los datos aquí recogidos pueden ser utilizados con multitud de finalidades. Las más directa son las relacionadas con el estudio y modelado del rendimiento deportivo de los propios jugadores. Por ejemplo, se podrían crear modelos que permitan ayudar a los equipos a tomar decisiones durante el "*draft*" (similar a la fase de mercado de fichajes) basadas en la evolución de los jugadores durante sus distintas temporadas, predecir quien pudiera ser el siguiente MVP de la NBA o detectar futuras promesas. También puede ayudar a los cuerpos técnicos a estudiar y conocer mejor a sus rivales, su propio equipo, realizar simulaciones en función de los jugadores en pista, etc.

Sin embargo, no solo tiene una aplicación directa sobre lo que viene siendo el deporte en sí mismo. Este conjunto de datos puede ser utilizado dentro del periodismo, más concretamente el deportivo, y en lo que se conoce cada vez más como *periodismo de datos*. Estos datos permitirían al periodista tener una posición más informada frente a las actuaciones y rendimiento de cada jugador.

7. Licencia

De acuerdo con lo estipulado en los términos de uso marcados por Sport Reference, creemos que la licencia que más se ajusta a este conjunto de datos es **CC BY-NC-SA 4.0 License**. La elección de esta licencia se ha realizabo bajo los siguientes tres pilares de este tipo de licencias.

- Como cualquier otra licencia Creative Commons, se debe **citar en todo momento la autoría original de la obra**. Este punto respeta lo definido en los términos de uso de los datos en los cuales se menciona que se debe reconocer a Sport Reference como autor y propietario original de los datos.
- Al igual que el punto anterior, las licencias Creative Commons garantizan que los trabajos derivados de la obra sigan reconociendo a su autor original, en este caso Sport Reference.
- La licencia estipulada permite un **uso no comercial** de los datos cumpliendo con lo estipulado por Sport Reference en los términos de uso. Sport Reference permite el uso de los datos siempre y cuando no sean para el desarrollo de un producto o servicio que entre en competencia con la actividad desarrollada por estos.

8. Código

En el siguiente enlace se puede encontrar el repositorio en GitHub que aloja el proyecto:

https://github.com/rrevuelta/NBA-Scraper

Una de las dificultades que se encontraron durante el desarrollo del scraper fue que la ruta para aceptar las cookies en la web es dinámica pudiendo variar entre dos opciones. Por ello, se incluyeron ambas opciones en el código.

El otro condicionante era no hacer un uso intensivo de web-scraping sobre la web para poder cumplir con los términos de uso. Por ello, Se han hecho multitud de pruebas contando con el número de peticiones realizadas y el tiempo de adquisición de los datos garantizando que no se sobrepasa en ningún momento el umbral de más de 20 peticiones por minuto.

Por este mismo motivo, el de evitar hacer un uso intensivo de la fuente de datos y cumplir con los términos de uso, se ha optado por descargar únicamente los datos asociados a las 12 últimas temporadas de la NBA. Sin embargo, el código está implementado para poder sacar todas ellas.

9. Dataset

A continuación, se muestra la referencia DOI al conjunto de datos:

David Esparcia, & Rubén Revuelta. (2022). NBA season player stats [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7347478

10. Vídeo

Se facilita el enlace de acceso al vídeo con la explicación:

https://drive.google.com/file/d/1R3Tl0V10ZG8OO_NNflpmPe0GPpozjy9i/view?usp=sharing

11. Participación

Contribuciones	Firma
Investigación previa	D.E / R.R
Redacción de las respuestas	D.E / R.R
Desarrollo del código	D.E / R.R
Participación en el vídeo	D.E / R.R

12. Referencias

- https://creativecommons.org
- https://www.selenium.dev/documentation/
- https://www.sports-reference.com
- https://www.basketball-reference.com/?__hstc=213859787.896116061e3d423283f1364246948fb2.1668277532548.1668277532548.1668421388982.2&__hssc=213859787.1.1668421388982&__hsfp=1114887993#site_menu_link