Universitat Oberta de Catalunya
uoc.edu
David Lucas Torres (dlucast@uoc.edu)
Francisco Javier Cantero Zorita (fcanteroz@uoc.edu)

# Tipología y ciclo de vida de los datos
## Práctica 1

## Descripción de la práctica

### Contexto

En el mundo actual, los datos, se han convertido en un activo fundamental para gobiernos y grandes empresas que hacen uso de estos para la toma de sus decisiones más importantes. Si nos referimos al deporte, esta situación no es diferente.

En el caso del baloncesto y en especial en la NBA, el uso de los datos se ha convertido en pieza fundamental del deporte ya que gracias a ellos los equipos pueden obtener ventajas competitivas conociendo mejor a los equipos a lo que se enfrenta además de poder conocer qué jugadores son los mejores de la liga. Esta ventaja competitiva ha hecho que en la NBA el avance con respecto a la recolección y tratamiento de los datos esté en su máximo esplendor y que tanto la liga como los equipos hagan uso de ellos.

Es por esto, que teniendo presente la importancia de los datos y el uso que se puede dar de ellos, se ha decidido realizar la recolección de las estadísticas de los jugadores en cada partido de la NBA a través del portal *https://www.basketball-reference.com/* que es, tras la página oficial de la NBA, el portal más popular a la hora de consultar las estadísticas de los jugadores actuales como históricos. Por lo tanto, se realizará la obtención de las estadísticas de los jugadores en cada partido y se realizará la suma de estos para ofrecer un dataset que pueda ser utilizado para posteriores análisis sobre este.

### Título

Como se ha explicado en el apartado anterior, el objetivo será recolectar la información de los jugadores de la NBA durante una serie de partidos. En este caso, la obtención de los datos se realizará en referencia a los partidos de la temporada 2020 - 2021 y es por esto que el dataset será llamado *NBAPlayerStatistics_20-21*, y el cual será obtenido en formato .csv por lo que el nombre final del fichero obtenido será *NBAPlayerStatistics_20-21.csv*.

### Descripción del dataset

El dataset contiene datos de cada uno de los jugadores que hayan interactuado con la NBA durante un periodo de tiempo específico (última temporada) y recoge todas las estadísticas acumuladas. Adicionalmente, el dataset incluirá la variable *PER*[1] (*player efficiency rating*) que permite evaluar la eficiencia del jugador con respecto a las estadísticas totales de la temporada. Mediante esta variable, se podrá conocer cómo ha sido el rendimiento del jugador y poder comparar con el resto de jugadores de una manera rápida y sencilla.

[1] https://es.wikipedia.org/wiki/Player_efficiency_rating

David Lucas Torres (dlucast@uoc.edu)
Francisco Javier Cantero Zorita (fcanteroz@uoc.edu)

Dicho esto, se pasa a continuación a mostrar la categorización de los jugadores en función del *PER* obtenido:

| | |
|---|---|
| La mejor temporada de la historia | 35.0+ |
| Objetivamente MVP | 30.0-35.0 |
| Fuerte candidato a MVP | 27.5-30.0 |
| Débil candidato a MVP | 25.0-27.5 |
| Fijo All-Star | 22.5-25.0 |
| Cerca de All-Star | 20.0-22.5 |
| Segunda opción ofensiva | 18.0-20.0 |
| Tercera opción ofensiva | 16.5-18.0 |
| Jugador ligeramente por encima de la media | 15.0-16.5 |
| Jugador de rotación | 13.0-15.0 |
| Jugador de no-rotación | 11.0-13.0 |
| Jugador de banquillo | 9.0-11.0 |
| Jugador que no se queda en la liga | 0 - 9.0 |

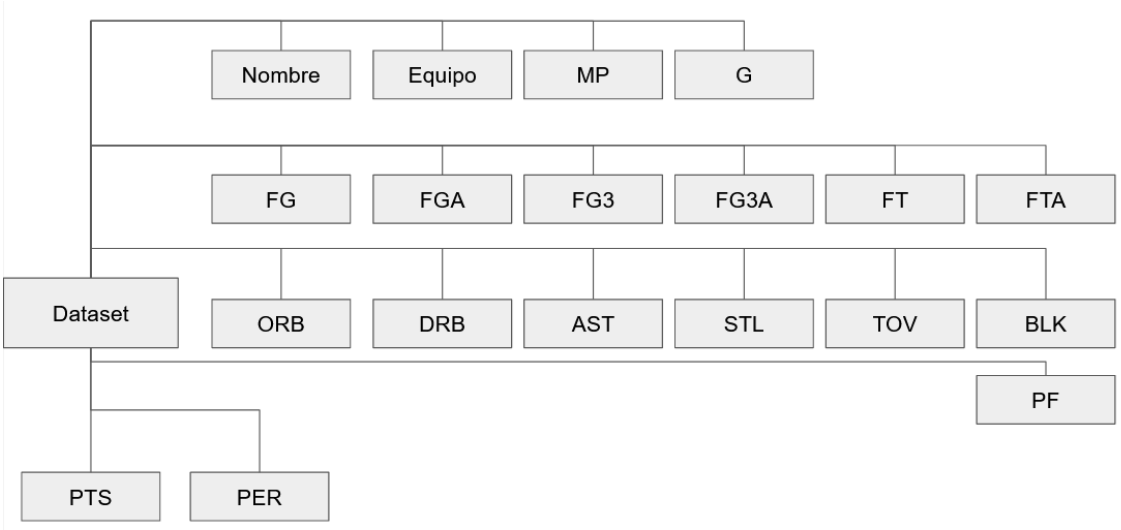
En su conjunto, este dataset pretende explicar, tanto de forma resumida como específica, lo que ha aportado un jugador a lo largo de un período de tiempo.

## Representación gráfica

Con respecto a la representación gráfica, en primer lugar, el ciclo de vida de los datos sería el siguiente:


Como se puede ver en el esquema anterior, los datos están almacenado en la página https://www.basketball-reference.com/ y es sobre la que tras realizar la ejecución del scraping se obtiene el dataset que se muestra a continuación:

David Lucas Torres (dlucast@uoc.edu)
Francisco Javier Cantero Zorita (fcanteroz@uoc.edu)




En este diagrama, se puede ver en la primera rama las variables que permite identificar el jugador y los minutos y partidos totales jugados por él, en la segunda todas las acciones de tiro realizadas por este, en la tercera estadísticas que engloban acciones ofensivas y defensivas, y por último las estadísticas valorativas del rendimiento del jugador. A partir, de estas variables el objetivo final del proyecto es identificar a los mejores jugadores en un periodo de tiempo indicado al igual que se realiza en la NBA:

David Lucas Torres (dlucast@uoc.edu)
Francisco Javier Cantero Zorita (fcanteroz@uoc.edu)

## Contenido

El dataset incluye un total de 19 campos, que se podrían dividir en 3 grupos: identificación del jugador, métricas de rendimiento y resumen de rendimiento.

- **Nombre:** Indica el nombre del jugador.
- **Equipo:** Indica el nombre del equipo al que pertenece el jugador.
- **MP:** (Minutes played) Indica los minutos que ha jugado en total un jugador.
- **FG:** (Field goals) Indica los tiros de campo anotados (sin contar los tiros libres).
- **FGA:** (Field goal attempts) Indica los tiros de campo realizados (sin contar los tiros libres).
- **FG3:** (3-point field goals) Indica los tiros de tres puntos anotados por el jugador.
- **FG3A:** (3-point field goal attempts) Indica los tiros de tres puntos realizados por el jugador.
- **FT:** (Free throws) Indica los tiros libres anotados por el jugador.
- **FTA:** (Free throw attempts) Indica los tiros libres realizados por el jugador.
- **ORB:** (Offensive rebounds) Indica los rebotes ofensivos capturados por un jugador.
- **DRB:** (Defensive rebounds) Indica los rebotes defensivos capturados por un jugador.
- **AST:** (Assists) Indica las asistencias realizadas por un jugador.
- **STL:** (Steals) Indica los robos realizados por  jugador.
- **BLK:** (Blocks) Indica los tapones realizados por el jugador.
- **TOV:** (Turnovers) Indica las pérdidas realizadas por el jugador.
- **PF:** (Personal fouls) Indica las faltas personales realizadas por el jugador.
- **PTS:** (Points) Indica el total de puntos anotados por el jugador.
- **G:** (Games) Indica los partidos totales jugados por el jugador.
- **PER:** (Player efficiency rating) Detalla la eficiencia del jugador haciendo uso de las estadísticas anteriores.

El periodo de tiempo de los datos, tal y como se ha comentado anteriormente, corresponde a la temporada 2020-2021 de la NBA (22 de diciembre de 2020 a 16 de mayo de 2021).

## Agradecimientos

Todos los datos que se recolectan para construir el dataset han sido obtenidos de la página web https://www.basketball-reference.com/, que como se ha comentado anteriormente, es la página no oficial de la NBA más utilizada para promover las estadísticas de los jugadores y equipos de la NBA desde el año 1946 hasta el dia de hoy.

Con respecto a los análisis similares, en el repositorio https://github.com/vishaalagartha/basketball_reference_scraper se puede acceder a un scraper de todos los elementos a los que se puede tener acceso en *Basketball reference*, y el cual como se indica en la documentación del repositorio, es el primer scraper a disposición de los usuarios para obtener la información de la página. Adicionalmente, también existe el repositorio https://jaebradley.github.io/basketball_reference_web_scraper/ el cual permite realizar la obtención de las estadísticas de los jugadore y de los equipos en un día definido en la ejecución.

Con respecto a lo ético, para asegurar que no se obtenía información no permitida, se ha revisado el archivo robots.txt de basketball-reference. De esta forma, se ha limitado a hacer scraping de las direcciones posibilitadas.

## Inspiración

El conjunto de datos obtenido, en el ámbito que se ha presentado durante el primer apartado, permite a los usuarios poder conocer qué jugadores de la NBA han obtenido mejores estadísticas y cuales componen el elenco de los mejores jugadores de la liga y que optan por el galardón de jugador con más valor, es decir, el MVP. Es por esto que las preguntas que responde el conjunto de datos son:

- ¿Cuales han sido las estadísticas totales de los jugadores durante un periodo de tiempo?
- En ese periodo de tiempo, ¿qué jugadores han tenido una mayor eficiencia?

Con respecto a los ejemplo indicados en el apartado anterior, el conjunto de datos se diferencia de ambos ya que ninguno de ellos permite obtener un conjunto de datos en el que se pueda comparar la eficiencia de los jugadores en un periodo de tiempo sino que mediante estas herramientas se puede obtener o las estadísticas totales o las estadísticas avanzadas en las que se incluye el PER pero no un conjunto de datos con ambas.

## Licencia

La licencia atribuida para el dataset resultante sería la de Released Under CC BY-SA 4.0 License, puesto que la intención de ayudar a realizar cualquier tipo de cálculos donde se necesiten las estadísticas por separado o la global del PER, requiere que se pueda copiar y redistribuir el material por cualquier medio y formato. Además, también se podrá adaptar para cualquier propósito, aunque sea comercial, puesto que los datos que obtenemos son de carácter público.

Esta licencia además, provoca que se den créditos, se indiquen los cambios que se han realizado y que se mantenga la misma licencia cuando se vuelva a compartir, por lo que se impide en cierta forma que un externo se lucre de este trabajo.

Además, hemos utilizado la web de https://ufal.github.io/public-license-selector/ para obtener una recomendación de licencia a utilizar, donde nos ha indicado la misma que habíamos seleccionado y nos ayuda a confirmar que ha sido una buena elección.

## Código

El código para realizar el scraping de la página web y crear el dataset ha sido realizado en el lenguaje de programación Python, utilizando las librerías *BeautifulSoup, pythonrequests y pandas.*

Se adjunta el link de Github para poder acceder al repositorio donde se encuentra el código.

*https://github.com/dlucas98/PRA1-Tipologia*

Para llevar a cabo este scraping se han tenido en cuenta una serie de buenas prácticas. Entre ellas, se ha evitado parsear el HTML manualmente mediante el uso de la librería BeautifulSoup, también se ha tenido en cuenta la posibilidad de que la página web caiga teniendo un tiempo máximo de espera para timeouts y se han modificado el user agent y otras cabeceras HTTP para evitar ser bloqueados por el sitio web.

## Dataset

Tras publicar el dataset final en formato csv en la web de Zenodo, se muestran a continuación el link de la web y el enlace del DOI.

Enlace en Zenodo para obtener el dataset: *https://zenodo.org/record/6425685#.YlBqaZlBxD8*

DOI: `10.5281/zenodo.6425685`

Enlace del DOI: *https://doi.org/10.5281/zenodo.6425685*

## Tabla de contribuciones

| Contribuciones | Firma |
|---|---|
| Investigación previa | David Lucas, Javier Cantero |
| Redacción de las respuestas | David Lucas, Javier Cantero |
| Desarrollo del código | David Lucas, Javier Cantero |