Placido A. López Ávila

Alex Pardo Ramos

**Tipología y ciclo de vida de los datos**

**Práctica 1**

**1. Contexto**

ESPN Deportes es un canal de televisión por suscripción estadounidense que emite eventos en español, propiedad de Walt Disney Televisión. Por esta razón, en su página web tiene información sobre eventos deportivos de todo tipo, y en este caso nos hemos centrado en el ámbito del baloncesto, en concreto en la liga estadounidense que es la NBA. De la cual dispone de toda la información posible sobre los partidos, equipos, jugadores y demás.

**2. Título**

Información NBA en intervalo de tiempo

**3. Descripción del dataset**

El dataset extraído se basa en la información de los partidos en un tiempo concreto, elegido por el usuario, y a su vez las estadísticas de los equipos y los jugadores que han participado en dichos partidos. Hay todo tipo de estadísticas relacionadas con el juego como puntos, tiros, rebotes, robos, etc. Esta información podría ser muy útil si es bien utilizada.

A modo de resumen, en la siguiente table se pueden apreciar el conjunto de columnas que encontramos en el dataset y sus descripciones:

|  |  |  |
| --- | --- | --- |
| **Columna** | **Descripción** | **Ejemplo** |
| Column1 | Número de fila | 0 |
| Date | Fecha del partido | 19/10/2021 |
| Away Team | Equipo visitante | Brooklyn |
| Home Team | Equipo local | Milwaukee |
| Result | Resultado del partido | MIL 127, BKN 104 |
| Max Winner | Jugador con mayor anotación del equipo ganador | Giannis Antetokounmpo |
| Pts Winner | Puntos del jugador del equipo ganador | 32 Pts |
| Max Loser | Jugador con mayor anotación del equipo perdedor | Kevin Durant |
| Pts Loser | Puntos del jugador del equipo perdedor | 32 Pts |
| Id Game | Id del partido | 401358773 |
| FG AT | Tiros de campo del equipo visitante | 37-84 |
| FG HT | Tiros de campo del equipo local | 48-105 |
| Field Goal % AT | Porcentaje de tiros de campo del equipo visitante | 440 |
| Field Goal % HT | Porcentaje de tiros de campo del equipo local | 457 |
| 3PT AT | Tiros de triple del equipo visitante | 17-32 |
| 3PT HT | Tiros de triple del equipo local | 17-45 |
| Three Point % AT | Porcentaje de tiros de triple del equipo visitante | 531 |
| Three Point % HT | Porcentaje de tiros de triple del equipo local | 378 |
| FT AT | Tiros libres del equipo visitante | 13-23 |
| FT HT | Tiros libres del equipo local | 14-18 |
| Free Throw % AT | Porcentaje de tiros libres del equipo visitante | 565 |
| Free Throw % HT | Porcentaje de tiros libres del equipo local | 778 |
| Rebounds AT | Rebotes del equipo visitante | 44 |
| Rebounds HT | Rebotes del equipo local | 54 |
| Offensive Rebounds AT | Rebotes ofensivos del equipo visitante | 5 |
| Offensive Rebounds HT | Rebotes ofensivos del equipo local | 13 |
| Defensive Rebounds AT | Rebotes defensivos del equipo visitante | 39 |
| Defensive Rebounds HT | Rebotes defensivos del equipo local | 41 |
| Assists AT | Asistencias del equipo visitante | 19 |
| Assists HT | Asistencias del equipo local | 25 |
| Steals AT | Robos del equipo visitante | 3 |
| Steals HT | Robos del equipo local | 8 |
| Blocks AT | Tapones del equipo visitante | 9 |
| Blocks HT | Tapones del equipo local | 9 |
| Total Turnovers AT | Pérdidas totales del equipo visitante | 13 |
| Total Turnovers HT | Pérdidas totales del equipo local | 8 |
| Points Off Turnovers AT | Puntos tras pérdidas del equipo visitante | 22 |
| Points Off Turnovers HT | Puntos tras pérdidas del equipo local | 2 |
| Fast Break Points AT | Puntos al contraataque del equipo visitante | 15 |
| Fast Break Points HT | Puntos al contraataque del equipo local | 21 |
| Points in Paint AT | Puntos en la zona del equipo visitante | 34 |
| Points in Paint HT | Puntos en la zona del equipo local | 42 |
| Fouls AT | Faltas del equipo visitante | 17 |
| Fouls HT | Faltas del equipo local | 19 |
| Technical Fouls AT | Faltas técnicas del equipo visitante | 0 |
| Technical Fouls HT | Faltas técnicas del equipo local | 0 |
| Flagrant Fouls AT | Faltas antideportivas del equipo visitante | 1 |
| Flagrant Fouls HT | Faltas antideportivas del equipo local | 0 |
| Largest Lead AT | Ventaja más amplia del equipo visitante | 2 |
| Largest Lead HT | Ventaja más amplia del equipo local | 23 |
| Team | Nombre del equipo | Nets |
| Name | Nombre del jugador | K. Durant |
| Position | Posición | AP |
| First/Substitute | Titular/Suplente | First |
| MIN | Minutos de juego | 30 |
| FG | Tiros de campo | 13-25 |
| % TC3 | Porcentaje de tiros de triple | 3-7 |
| TL A-I | Tiros libres | 3-6 |
| OREB | Rebotes ofensivos | 0 |
| DREB | Rebotes defensivos | 11 |
| REB | Rebotes totales | 11 |
| AST | Asistencias | 4 |
| STL | Robos | 0 |
| BLK | Tapones | 2 |
| PÉR | Pérdidas | 1 |
| PF | Faltas personales | 2 |
| +/- | Diferencia en el marcador estando el jugador en juego | -20 |
| PTS | Puntos anotados | 32 |

**4. Representación gráfica**

**ESPN NBA**

**Web Scraping**

**Extracción Partidos**

**Extracción Jugadores**

**Extracción Calendario**

**Merge**

**Dataset**

**5. Contenido**

La extracción de la información se ha realizado empleando la librería BeautifulSoap de Python, generando una clase llamada RobotScraper con toda la lógica necesaria para el proceso. Desde un main se controla el uso de los parámetros para el control de los rangos de fechas deseados en la extracción del dataset. El proceso de extracción se separa en 3 partes bien diferenciadas, las cuales deben seguir un orden concreto para la correcta generación de los datos:

* Extracción Calendario: Se trata de una primera extracción, la cual está centrada principalmente en los datos resumidos de los partidos seleccionados, extrayendo los datos del equipo visitante, el equipo local, el resultado, máximo anotador del equipo ganador, Puntos del equipo ganador, máximo anotador del equipo perdedor, puntos del equipo perdedor y la URL del partido. Esta URL cuenta con el ID del partido, identificador único que se necesita para poder seguir navegando hasta los datos más específicos del partido y jugadores en este. El nivel de detalle de este conjunto de datos es muy general, teniendo 1 fila por partido.
* Extracción Partidos: Se trata de la segunda extracción, la cual extrae información más concreta de cada uno de estos partidos, algunos de estos datos son: estadísticas tanto para el equipo visitante como para el local, tiros de campo, porcentaje de acierto en tiros de campo, tiros de tres puntos, porcentaje de acierto en tiros de tres puntos, tiros libres, porcentaje de acierto en tiros libres, rebotes, rebotes ofensivos, rebotes defensivos, asistencias… Esta información se almacena junto con el ID del partido, estando al mismo nivel de detalle que la anterior extracción, 1 fila por partido.
* Extracción jugadores: Se trata de la última extracción, la cual se compone de la información por jugador en cada uno de los partidos, en concreto esta información es el id del partido, el equipo y el nombre del jugador, posición, Titular o suplente, minutos, tiros de campo, porcentaje de triples, tiros libres anotados e intentados, rebotes ofensivos, rebotes defensivos, rebotes, asistencias, robos, bloqueos, perdidas, faltas, balance de puntos y puntos. El nivel de detalle es mucho mayor, teniendo 1 fila por jugador, de tal forma que al realizar un left join con el resto de la información las otras dos extracciones se repetirán tantas veces como jugadores en el partido. Es importante tener esto en cuenta para realizar las limpiezas de datos y el análisis de la información en el dataset.

Tras esto se realiza un merge de las tres extracciones, en concreto las operaciones de left join, puesto que hay partidos pospuestos sin datos de jugadores, pero si del partido en general. Esto se realiza para combinar la información, y generar un dataset con toda esta información.

**6. Agradecimientos**

El propietario de este conjunto de datos como ya hemos comentado antes es la compañía Disney. Hay diversos análisis en repositorios como GitHub que pueden tener algunas semejanzas con este, pero ninguno de ellos trata toda la información de partidos, equipos y jugadores en base a tiempo, por lo tanto, es difícil la comparación con cualquiera de estos. A la hora de realizar el proyecto, hemos tenido en cuenta las pocas restricciones de uso de estos datos que podemos encontrar al final de la web, para no realizar nada fuera del marco permitido.

**7. Inspiración**

Es interesante este conjunto de datos porque nos permitirán el estudio tanto de partidos, equipos y jugadores en un tiempo concreto, siendo el usuario quien elige el rango de fechas a extraer, lo cual diferencia este estudio a los anteriores a este. Podemos responder casi cualquier pregunta sobre estos tres ámbitos. Esto permitirá tener un acceso mucho más directo e incluso realizar estudios sobre la propia información que se extrae. Es decir, este conjunto de datos nos proporcionara acceso a información temporal muy detallada de la NBA.

**8. Licencia**

Relased Under CC BY-SA 4.0. License. Se tiene la libertad de compartir y adaptar los datos de los cuales se dispone. Las condiciones de esta son que se debe reconocer adecuadamente la autoría y que, aunque se mezcle, transforme o cree a partir del material, hay obligatoriedad de que la licencia sea la misma que la original.

**9. Código**

Enlace a repositorio Git.

<https://github.com/plopezavi/Web-Scraping>

**10. Dataset**

Enlace de los dataset en formato .csv en Zenodo.

<https://doi.org/10.5281/zenodo.6450203>

**Firmas**

|  |  |
| --- | --- |
| **Contribuciones** | **Firmas** |
| Investigación previa | P.L.A. , A.P.R. |
| Redacción de las respuestas | P.L.A. , A.P.R. |
| Desarrollo del código | P.L.A. , A.P.R. |