TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRÁCTICA 1: WEB SCRAPING

Gabriel Quintairos Rial

1. **Contexto:** se pretende obtener datos sobre el valor de mercado de los jugadores de los equipos de la Liga Santander, así como alguna información adicional que aparezca con el dato que queremos. El conjunto de datos se ha obtenido de una página dedicada a mostrar información sobre el valor de mercado de jugadores de fútbol.
   Hoy en día la Liga Santander es una competición deportiva muy seguida a nivel mundial, siendo los jugadores los principales protagonistas y artífices de este espectáculo. Conocerlos y tener datos sobre ellos es fundamental para multitud de campos, desde profesionales del deporte (directores deportivos, ojeadores, agentes de futbolistas, etc.) hasta aficionados que desean conocer mejor a su equipo o a sus rivales o jugar al conocido Comunio [1].
   Es por esto por lo que tener datos actualizados sobre los jugadores es algo muy importante, ya que dichos datos tienen importancia para un gran número de personas y por tanto adquieren mucho valor.

2. **Título:** Información sobre el valor de mercado de los jugadores de la Liga Santander.

3. **Descripción:** el conjunto de datos contiene información sobre los jugadores de la Liga Santander, en concreto: equipo al que pertenece, su edad y fecha de nacimiento, su valor de mercado, su posición, su dorsal y su nacionalidad. Todos los datos son cadenas de caracteres, salvo el dorsal que es un número entero.
   Se han descargado los jugadores según el orden que aparecen los equipos en la web (orden alfabético) y dentro de cada equipo por posición. En total se han obtenido datos de 495 jugadores. La fecha de descarga ha sido el 10 de abril de 2019, con lo cual el valor de mercado corresponde a dicha fecha. El script se puede ejecutar en cualquier momento para actualizar dichos valores, ya que el archivo se sobrescribirá. Sería interesante tener planificada la ejecución de dicho script con algún tipo de periodicidad para mantener un evolutivo de los datos y tener siempre una última copia actualizada.
   Los datos descargados podrían necesitar una posible limpieza o transformación una vez completada la descarga. Por una parte, podría ser necesario convertir el campo fecha de nacimiento y edad a dos campos distintos, uno que contenga la fecha de nacimiento convertida a timestamp o dividida en 3 campos diferentes (día, mes y año) y otro que contenga la edad como número entero. También podría ser necesario limpiar la cadena que contiene el valor de mercado, eliminando el texto que corresponde a la unidad y quedándose únicamente con el valor numérico (teniendo especial cuidado ya que en algunos casos son millones y en otros casos miles de euros y eso habría que tenerlo en cuenta al convertirlo a número).
   Un posible inconveniente de este script es que la página web deje de estar operativa o la URL donde aparecen estos jugadores se modifique, por lo cual es necesario estar atentos por si necesitamos hacer alguna modificación. Si no se da el caso, los datos descargados serán siempre consistentes.

4. **Representación gráfica:**

Referencia: https://www.marca.com/futbol/primera-division/especial/2018/08/16/5b7191b3268e3ea22c8b45ac.html

5. **Contenido:** para cada jugador se recogen los siguientes datos:
   a. **Equipo:** equipo al que pertenece.
   b. **Dorsal:** dorsal que porta en la temporada en curso en los partidos de la Liga Santander.
   c. **Posición:** ubicación que tiene dicho jugador en el terreno de juego.
   d. **Nombre:** nombre del jugador.
   e. **Fecha de nacimiento:** fecha de nacimiento del jugador.
   f. **Edad:** edad del jugador.
   g. **Nacionalidad:** nacionalidad del jugador
   h. **Valor de mercado:** valor actual de mercado del jugador (en millones de euros).

   Se ha hecho uso del lenguaje Python, de las librerías BeautifulSoup y request para realizar web scraping y obtener la información alojada en las páginas HTML de cada equipo y de la librería csv para escribir los datos en un fichero .csv.

   Para ello en primer lugar se analizó el código fuente de la página utilizando la funcionalidad del navegador que permitía hacerlo. Cuando supimos en qué campo estaba la información que queríamos, realizamos la codificación.

   Se incluye en primer lugar una funcionalidad para examinar el archivo robots.txt, viendo que deshabilita a robots del tipo Exabot y que no incluye nada para el resto de user-agents.

Se utilizó la librería request para hacer un GET y obtener el código fuente de la página principal, el cual fue leído utilizando la librería BeautifulSoup. De dicho código se obtuvieron los enlaces correspondientes a la página de cada equipo, guardándolos en un array. Finalmente, se iteró sobre ese array y se volvió a utilizar request para hacer un GET y obtener el código fuente de la página de cada equipo, para de nuevo leerlo con la librería BeautifulSoup y obtener los datos de cada jugador. Los datos fueron guardados en un fichero csv ytilizando la función csv.writer de la librería csv.

6. **Agradecimientos:** el conjunto de datos ha sido extraído de la web de Transfermarkt (https://www.transfermarkt.es/).
   Esta página web ya ha sido utilizada para estudios anteriormente, como por ejemplo en este artículo Beya Acero (2019) [2]. En él se utilizan los valores de mercado dados por Transfermarkt para realizar una comparativa entre los dos clubes más importantes de la Liga Santander, el F.C. Barcelona y el Real Madrid C.F. En este artículo se comparan los valores en millones de euros de varios jugadores y también se analizan algunas tendencias.
   Otro artículo que ha utilizado los datos de esta página web es el de La Prensa de Honduras (2019) [3]. En él se hace un repaso de los futbolistas centroamericanos con mayor valor de mercado.

7. **Inspiración:** como amante del fútbol, me ha parecido muy interesante este conjunto de datos, ya que proporciona información acerca de todos los jugadores de la Primera División del fútbol español.
   Con este dataset se puede, por una parte, conocer el dorsal que tiene cada jugador en su equipo, lo cual puede ser útil a la hora de elaborar camisetas o de tener identificados a los jugadores.
   También podremos saber su posición, lo que nos puede servir para conocer en qué zona del campo se desenvuelven.
   Podremos conocer también su nacionalidad, lo que es de utilidad cuando hay campeonatos de selecciones nacionales saber por cuál de los países puede ser seleccionado cada jugador.
   Por último, se obtiene el valor de mercado dado por la web, el dato que me parece más interesante. Si se planifica una descarga automática de estos datos, se podría tener un histórico de la evolución temporal del valor de mercado de los futbolistas. Estos datos podrían posteriormente ser analizados para detectar algún tipo de tendencia, como por ejemplo una época del año en la que el valor de mercado suba o baje. Al tener además la posición y la edad de cada jugador, también podremos ver si existe una relación entre el valor de mercado de cada jugador y su puesto sobre el campo o los años que tiene.
   Existen algunos estudios que tratan de determinar el valor de mercado de los jugadores y establecer alguna forma de calcularlo. En un estudio de Fútbol Finanzas (2017)[4] nos presentan el sistema MERC, el cual se basa en varios indicadores para hallar el valor, como son: posición, edad del jugador, su

repercusión mediática y su calidad futbolística. También analiza cómo influye la oferta y la demanda en la fluctuación de los precios.
En otro estudio de Samaniego, Juan F. (2018)[5], se habla de una evolución temporal del valor de los jugadores y su reciente alza. Intenta buscar una explicación de este hecho, mencionando que podría ser por el incremento de los derechos televisivos que ingresan los clubes de fútbol. También cita la influencia de la ley de la oferta y la demanda e incide en la importancia que es tener datos. De hecho, indica que se están buscando modelos científicos que puedan explicar el comportamiento que tiene el valor de mercado de los jugadores.
Pienso que la base de datos creada en esta práctica puede agregar valor a estudios del estilo de los dos que acabo de mencionar. En primer lugar, si se realiza una descarga de estos datos de forma continuada en el tiempo se puede tener registrada la tendencia de precios, pudiendo de esta forma aplicarle algoritmos de minería de datos o Machine Learning para poder extraer conclusiones y obtener información relevante.
Por otra parte, el hecho de tener también la posición, edad y nacionalidad de los futbolistas permite añadir más variables que podrían aportar información a la hora de generar un posible modelo.
Quizá sería interesante en una versión posterior de este dataset añadir información nueva, como por ejemplo el número de partidos jugados por cada jugador, su pierna más hábil o el número de tarjetas que recibe en la temporada.
El hecho de tener el script disponible también permite tener la información fresca y una copia reciente y actualizada del valor de mercado de todos los jugadores.

8. **Licencia:** la licencia escogida es Released Under CC BY-NC-SA 4.0 License por los siguientes motivos:
    a. **Atribución:** de esta forma se reconoce el trabajo original y los cambios que se han realizado, permitiendo adaptar el dataset y mejorarlo con otras ideas distintas.
    b. **Uso no comercial:** este trabajo me parece más didáctico y dedicado a la investigación que algo comercial. Pienso que en el mundo del fútbol ya hay demasiados elementos comerciales como para añadir más.
    c. **Misma licencia:** todos los trabajos realizados a partir de este conjunto de datos deberán mantener la misma licencia, respetando de este modo los motivos que llevaron a escogerla.

9. **Repositorio:** el enlace al repositorio es el siguiente: https://github.com/gquintairos/transfermarketLigaSantander1819

10. **Contribuciones:**

Contribuciones	**Firma**
Investigación previa	Gabriel Quintairos Rial

Redacción de las respuestas	Gabriel Quintairos Rial
Desarrollo código	Gabriel Quintairos Rial

11. Referencias:

[1] Comunio: https://www.comunio.es/home

[2] Beya Acero, Pau. (01 de marzo de 2019). ¡Ningún jugador del Real Madrid entre los cinco con más valor de la Liga! Recuperado de: https://www.culemania.com/merengadas/ningun-jugador-real-madrid-entre-cinco-mas-valor-liga_226407_102.html

[3] La Prensa de Honduras. (26 de marzo de 2019). ¡Con 7 hondureños! Los futbolistas centroamericanos más caros que están en Europa, según Transfermarkt. Recuperado de: https://www.laprensa.hn/fotogalerias/deportes/1269745-411/con-7-hondure%C3%B1os-los-futbolistas-centroamericanos-m%C3%A1s-caros-que-est%C3%A1n-en-europa

[4] Redacción Fútbol Finanzas (27 de febrero de 2017). Así se calcula el valor de mercado de los futbolistas de élite – Sistema MERC. Recuperado de: https://futbolfinanzas.com/calculo-del-valor-de-mercado-de-los-futbolistas-de-elite/

[5] Samaniego, Juan F. (21 de septiembre de 2018). ¿Cómo llega un futbolista a costar 200 millones de euros? Así fluctúa el valor de los jugadores. Recuperado de: https://hablemosdeempresas.com/empresa/calcular-el-precio-de-un-futbolista/