Tipologia de dades - PRAC. 1

Dataset: Jugadors de futbol primera i segona divisió Espanyola

Universitat Oberta de Catalunya

Professora - Mireia Calvo González

Màster en Ciència de Dades

Autors:

Eduard López Fina & Marc López Vila

Abril 2023

1 Descripció dataset

El dataset generat és una recopilació de les estadístiques principals de tots els jugadors de futbol de primera divisió espanyola. Per cada jugador es recull tant informació de caràcter personal (nom complet, data de naixement, país d'origen, etc.) com de caràcter esportiu (fores de joc, targetes, gols, etc.).

2 Context

Tota la informació s'ha extret del diari esportiu Marca, concretament dels següents enllaços:

- `https://www.marca.com/futbol/primera-division/clasificacion.html`
- `https://www.marca.com/futbol/segunda-division/clasificacion.html`

El codi recorre cada equip llistat que apareix en alguna de les lligues, i per un cop dins recorre tota la plantilla per guardar la informació que s'ha considerat més rellevant.

3 Contingut

Els camps que s'inclouen en el dataset són:

- **playerName**: Nom complet del jugador.
- **height**: Alçada del jugador en *cm*.
- **weight**: Pes del jugador en *kg*.
- **birthDate**: Data de naixement i, si està disponible, la ciutat.
- **country**: País d'origen
- **url**: Enllaç de la fitxa del jugador a Marca.com.
- **penalties**: Nombre de penaltis xutats.
- **shots**: Nombre total de xuts.
- **foulsReceived**: Nombre de faltes rebudes.
- **offSides**: Nombre fores de joc comesos.

- **foulsCommited**: Nombre de faltes comeses.
- **passesCutOff**: Nombre de passades rivals tallades.
- **entrancesWon**: Nombre entrades realitzades amb èxit.
- **cards**: Nombre targetes rebudes, tant grogues com vermelles.
- **passes**: Nombre de passades totals realitzades.
- **goalAssists**: Nombre assistències de gol realitzades.
- **dribbles**: Nombre total de regatejos, tant amb èxit com sense.
- **corners**: Nombre de còrners xutats.
- **ballLosses**: Nombre pilotes perdudes.

S'ha de tenir present que **tots els atributs de caràcter esportiu només estan actualitzats en el moment d'obtenir les dades**, ja que en l'instant en què es juga un partit deixen d'estar-ho.

4 Representació gràfica

El dataset es pot representar de la següent manera:

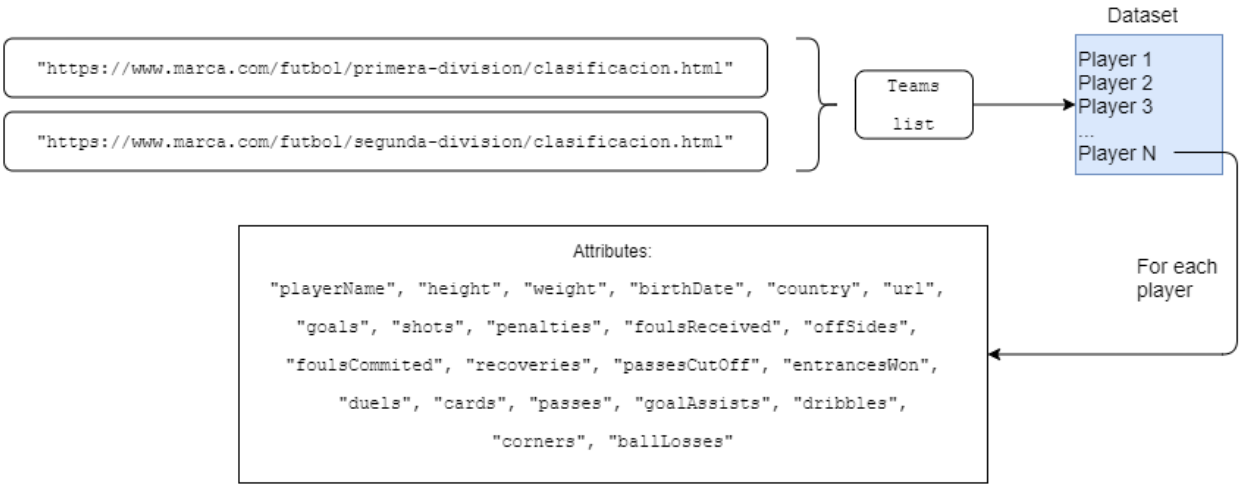

Figure 1: Diagrama del procés d'obtenció dataset

5 Propietari

Primerament hem analitzat l'arxiu ROBOTS.TXT per veure si teniem alguna restricció al fer web scraping:

```
User-agent: *
Disallow: /s/
Disallow: /corporativo/aviso-legal.html
Disallow: /corporativo/contacto.html
Disallow: /corporativo/ayuda.html
Disallow: /multimedia/en-tu-movil/listado/index.html
Disallow: /social/
Disallow: /edicion/
Disallow: /eltiempo/
Disallow: /deporte/futbol/primera-division/2010-2011/panel-de-fichajes/*
Disallow: /eventos/marcador/futbol/2013_14/*
Disallow: /eventos/marcador/futbol/2012_13/*
Disallow: /eventos/marcador/futbol/2011_12/*
Disallow: /encuentros/roberto-palomar/2016/03/29/*
Disallow: /2012/11/03/en/football/spanish_football/1351965522.html
Disallow: /2012/11/03/futbol/1adivision/1351945508.html
```

Figure 2: Document *robots.txt* de la web Marca.com

Podem veure el propietari no ens ha restringit l'accés a cap de les pàgines que volem consultar. També s'ha consultat qui és el propietari del domini utilitzant la llibreria *whois* de python.

>> print(whois.get("https://www.marca.com"))

Aquesta comanda ens dona molta informació, entre la que destaca:

- name: marca.com
- tld: com
- registrar: EuroDNS S.A.
- registrantCountry: ES

Un cop vista tota la informació, arribem a la conclusió que no tindrem problemes a l'hora de fer web scraping i utilitzar les dades, ja que no tenim cap limitació per part de Marca i són dades públiques no sensibles de jugadors.

6 Inspiració

Aquest anàlisi és interessant per tota la gent que segueix el futbol i que es dedica a aquest sector. Com que agafem les estadístiques actuals dels jugadors, les dades ens serveixen no només com un històric, sinó com una eina pels equips a l'hora de preparar els següents partits o mercats de fitxatges.

De fet, aquestes dades les haguéssim pogut extrete d'altres webs, però s'han extret de Marca perquè hem vist que el format i estructura html de la web ens ho permetia. Altres opcions podien haver estat:

- `www.laliga.es`
- `www.transfermarkt.com`
- `www.fbref.com`

Un cop extretes les dades, el dataset podrà resoldre preguntes d'aficionats com:

- Qui és el màxim golejador del meu equip?
- Quins jugadors hi han al meu equip?
- De quina nacionalitat són els jugadors del meu equip?

Fent un estudi més profund i relacionant les dades entre si, també pot resoldre preguntes més estratègiques pels entrenadors actuals com:

- Quines són les debilitats estratègiques del meu equip?
- Quines posicions hauríem de reforçar amb fitxatges?
- Quin és l'estil de joc del següent rival?
- Quins són els punts dèbils del rival?

7 Llicència

La llicència seleccionada pel nostre dataset ha estat la *CC BY-NC-SA 4.0 License* per les raons següents:

- S'ha de proveir el nom de l'autor del conjunt de dades generades indicant els canvis que s'han fet.
- Permet l'ús comercial, fent així que hi hagin més probabilitats de que utilitzin el nostre dataset.
- Les contribucions realitzades a posteriori hauran de mantenir la mateixa llicència. Amb això aconseguirem que no es pugui canviar la llicència del nostre treball.

8 Codi

El codi està disponible al repositori (`https://github.com/eduardfina/laliga-players`).

Per programar s'ha utilitzat Python3.9 i les llibreries externes que hem instal·lat són les següents:

- beautifulsoup4==4.12.0
- bs4==0.0.1
- selenium==4.8.3
- urllib3==1.26.15

El lloc web triat no ha presentat gaires dificultats per la seva bona estructura.

Al inici el programa agafa els links dels equips, aquí és on ens hem trobat la màxima dificultat. No podíem agafar directament el codi html perquè els links dels equips es trobaven dins de codi no compilat a una #shadow-root. Per això vam haver d'utilitzar selenium perquè obris la pàgina a través del Chrome i agafés la informació de la shadow root. Aquest procés es pot trobar a la funció __get_team_links().

Un cop tenim els links dels equips, trobem el link de la plantilla utilitzant BeautifulSoup a la funció __get_squad_link().

Ja tenint el link de les plantilles fem el mateix procés per trobar els links dels jugadors amb la funció __get_player_links().

Finalment amb la funció __get_player_basic_info() agafem la informació bàsica del jugador i amb la funció __get_player_stats_link() trobem el link d'on agafem informació més avançada dels jugadors amb la funció __get_stats().

Aquesta informació llavors la passem a csv cridant a la funció data2csv().

9 Dataset

Enllaç del DOI del dataset: `https://doi.org/10.5281/zenodo.7795945`.

10 Video

Enllaç al vídeo explicatiu:

`https://drive.google.com/file/d/1V3-F5EE0-Jj1Yu1Ol9NRv9zu3Ij_BgLi/view?usp=sharing`.