PRÀCTICA 1 – WEB SCRAPING

Obtenció d'un data set de tots els jugadors actuals de la lliga espanyola amb el seu valor de mercat a partir de les dades de Transfermarkt

NOM I COGNOMS: JORDI PUIG BENAGES

Novembre 2021

1. Context

A l'hora de realitzar la pràctica, s'ha utilitzat la font de dades Transfermarkt, a partir de la seva versió anglesa https://www.transfermarkt.co.uk. Aquesta font de dades s'ha convertit en la gran referència principal pels clubs de futbol, pels jugadors i per totes les agències de representació per a conèixer el valor de mercat dels jugadors.

Són dades completament públiques però que en alguna ocasió han causat polèmiques degut a no tenir un algoritme obert per saber les valoracions, cosa que podria, amb el gran poder que ha agafat, ser una eina d'influència en els traspassos dels jugadors, ajudant a inflar o reduir xifres.

2. Descripció del data set

El data set resultant és una base de dades de tots els jugadors que participen en equips de la Primera Divisió espanyola de la temporada 2021-2022 amb informació personal i el seu valor de mercat.

3. Representació gràfica i contingut

A partir d'un exemple de pàgina de jugador: https://www.transfermarkt.co.uk/dani-parejo/profil/spieler/59561 es mostra les dades que són interessants i que es recullen al data set final.

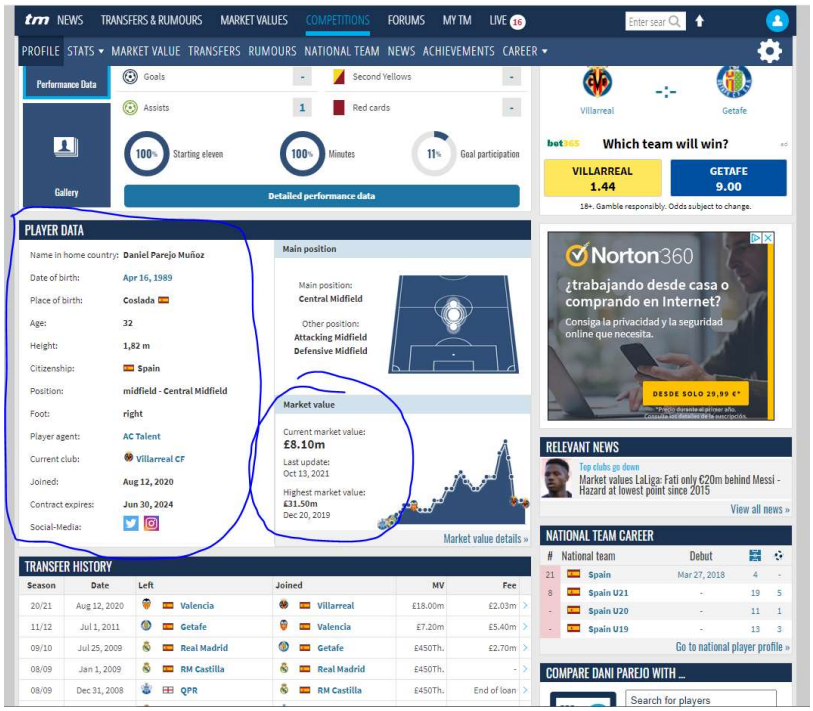

Figura 1: Exemple pàgina de la qual es fa web scraping

Es pot mostrar el procés general que es segueix per a l'obtenció del Dataset, havent d'entrar una URL referent a la competició a estudiar, per acabar arribant a la llista de tots els jugadors amb la seva informació.

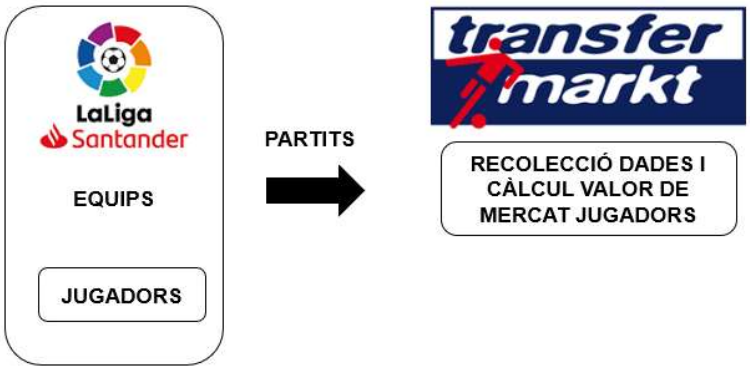

Figura 2: esquema del fluxe de dades entre LaLiga i Transfermarkt

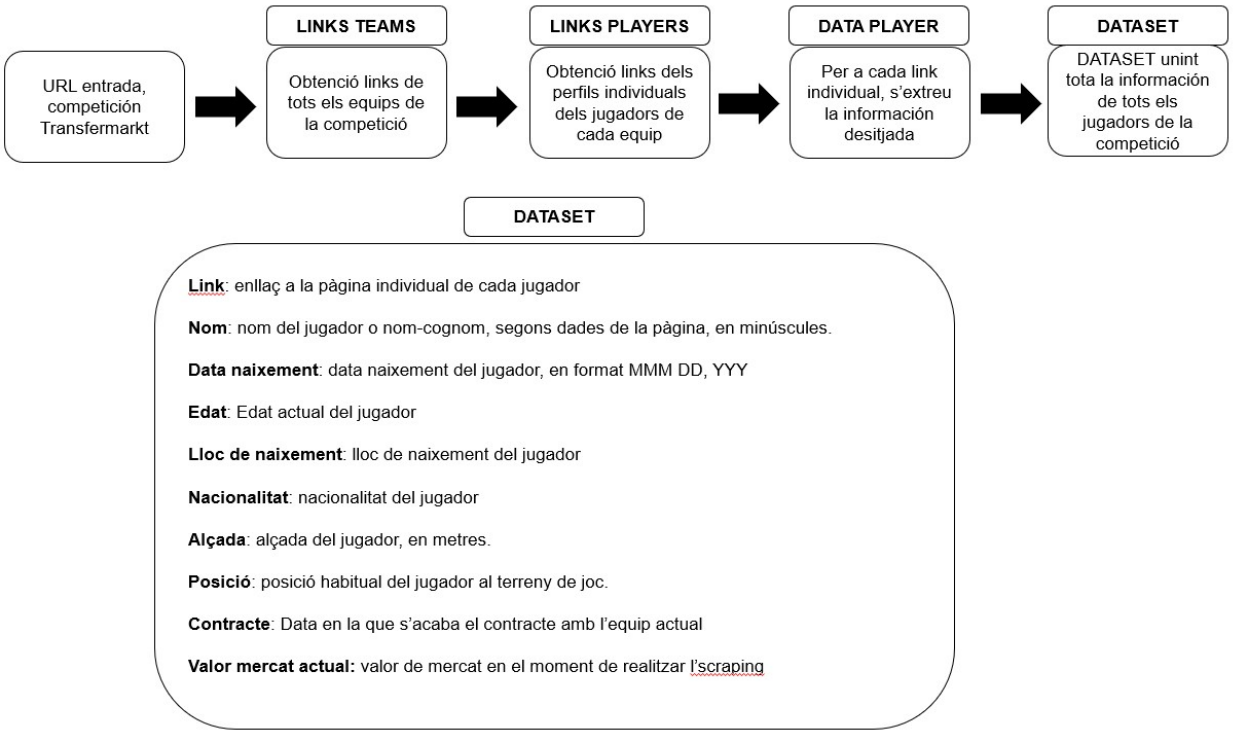

Figura 3: esquema del projecte i descripció del contingut

Les dades recollides, que són totes de tipus string, són recollides en el moment actual. Cada cop que s'executi pot haver diferències ja que el valor de mercat pot ser diferent. No obstant, com a mostra s'ha fet per a tots els jugadors de la Primera Divisió espanyola, extraient primer tots els equips que es troben a la competició de la temporada actual, 2021-2022. Si executem el mateix codi en 2 anys, tornaríem a extreure la informació de tots els jugadors dels equips que es troben a la competició actual, tot i que en 2 anys estiguin a categories inferiors. Per tant, la informació dels jugadors seria la del moment de l'extracció però no es compliria que estiguessin tots els equips de la Primera Divisió, si no que serien els de la temporada 2021-2022.

El codi, tanmateix, funcionaria igualment si introduïm altres URL's d'altres competicions o de la mateixa competició per altres temporades (només canviant el 'saison_id=2021' de la URL per un altre valor.

Per últim, és important esmentar que el projecte, en la meva opinió, hauria de ser la part inicial d'un projecte més gran on el codi oferís una visió més completa i acurada de la informació del jugador.

Per un costat, oferir una visió més completa amb informació més detallada de cada jugador, afegint per exemple l'agència de representació, informació de partits jugats per temporada, un històric de lesions...i tota la resta d'informació que pugui ajudar al consumidor final del data set a tenir major coneixement dels jugadors.

Per altra banda, una visió més acurada ja que s'haurien d'introduir mecanismes per evitar que la informació extreta no es correspongui als camps adients. En el data set final es pot veure com hi ha jugadors que no tenen alguna informació i, degut a la metodologia utilitzada en el codi, aquest no és capaç de veure-ho i, enlloc d'afegir 'NA' a la informació, afegeix la informació següent, fet que crea, en alguns casos, que la informació als camps no es correspongui amb el què s'hauria de trobar a aquella columna.

4. Agraïments

El propietari de les dades és Transfermarkt, una web alemanya creada al 2000 per Matthias Seidel i que al 2008 va ser adquirida per l'empresa Axel Springer, que ofereix informació futbolística, tant històrica com actual, sobre aspectes molt variats: resultats de competicions, historial futbolístic dels jugadors, disposicions tàctiques... és a dir, és una web que té una enorme base de dades de competicions, equips, jugadors i entrenadors de futbol. Les seves dades més valorades, en qualsevol cas, són el valor de mercat de cada jugador, producte de la participació comunitària (reben feedback dels usuaris i d'experts i van realitzant variacions segons rendiment, edat, posició, etc.), sense disposar d'un algoritme.

5. Inspiració

Com a apassionat dels esport, especialment del futbol, intento seguir de forma propera el món de les dades al futbol. Cada cop hi ha més ús de les dades i eines com el web scraping poden ser molt importants. Transfermarkt pot ser molt important pels clubs, ja que disposa de molta informació, i particularment obtenir de forma ràpida informació referent al valor de mercat actual pot ser una eina d'ajuda a l'hora de realitzar informes per possibles fitxatges, negociar traspassos o contractes amb els jugadors.

De la mateixa forma, es poden resoldre ràpidament preguntes com:

- Quins són els porters menors de 25 anys que tenen un valor inferior de 2M€?
- Quins jugadors acaben contracte aquest 30 de Juny i, per tant, podem començar a negociar amb ells el dia 1 de gener sense haver de pensar en pagar un fitxatge?
- Quins davanters són els més alts?
- Quins són els 10 jugadors amb un valor de mercat més alt?

A més a més, és un codi que funciona per cada URL que li posem de les competicions de Transfermarkt, de forma que podríem completar el Dataset amb informació de la Premier League, de la Bundesliga o de la Lliga Francesa i tenir una base de dades gratuïta que ens ajudi a complementar la informació que puguem tenir d'altres bases de dades.

En definitiva, ens pot resoldre moltes preguntes de forma senzilla.

En el procés de desenvolupament del codi, he consultat diferents fonts de projectes similars, alguns dels quals han deixat de ser operatius pels canvis que les pàgines web han anat introduint. En primer lloc, un projecte propi (https://github.com/jordip6/Virtual-Sports-Director) en el que realitzava scraping de Transfermarkt i amb dades de rendiment dels jugadors del proveïdor Wyscout, podia obtenir un data set que tingués molta més informació de KPI's de rendiment. Per altra banda, hi ha molts recursos interessant per poder fer web scraping a Transfermarkt o a pàgines proveïdores de dades futbolístiques com fbref.com.

Alguns dels consultats han sigut *Introduction to Scraping Data from Transfermarkt* (https://fcpython.com/scraping/introduction-scraping-data-transfermarkt) i *FBref_EPL* (https://github.com/chmartin/FBref_EPL) .

6. Llicència

S'ha escollit la llicència MIT, ja que és la més oberta de totes, permet fer un ús comercial, distribuir o modificar el codi i fer un ús privat, mentre que les condicions i limitacions són les lògiques. Personalment, crec que en tractar-se d'una font de dades oberta en un sector com el futbol, hi haurà solucions més completes que la que he realitzat jo, així que qualsevol ús i feedback del codi serà més que benvingut.

7. Codi

El codi es troba al meu repositori personal: https://github.com/jordip6 amb el nom de 'Webscraping-Transfermarkt'

8. Dataset

El Dataset obtingut s'ha pujat a Zenodo en el següent enllaç:

https://zenodo.org/record/5651520#:~:text=Target%20URL-,https%3A//doi.org/10.5281/zenodo.5651520,-License%20(for%20files

```
https://doi.org/10.5281/zenodo.5651520
```

9. Vídeo

El vídeo explicatiu es troba al següent enllaç del Drive:

https://drive.google.com/file/d/187s-sHjXlRRIO4waQQf5vfrX4TgHt1EN/view?usp=sharing

10. Contribució

En realitzar la pràctica de forma individual, tots els apartats han sigut realitzats per mi mateix.

Contribucions	Signatura
Investigació prèvia	JPB
Redacció de les respostes	JPB
Desenvolupament del codi	JPB

Taula 1: contribucions de la pràctica