TIPOLOGÍA I CICLE DE VIDA DE LES DADES
PRÀCTICA 1

Web Scraping dels jugadors de futbol de la lliga professional de la primera divisió espanyola

Noms dels estudiants: **Josep Pou Mas** i **Xavier Badia Mulero**

1. Context

El conjunt de dades recollides correspon als jugadors de la lliga de primera divisió espanyola de futbol de la temporada 2021-22.

La recollida d'aquest conjunt de dades es realitza amb la finalitat d'aportar una eina que té l'objectiu de proporcionar una informació útil pels entrenadors, directius esportius o persones amb interessos professionals, que permeti contribuir a l'elecció de fitxatges, la gestió per part dels tècnics esportius de la càrrega de partits i minimitzar les possibles lesions durant la temporada.

Aquesta web proporciona tota la informació actualitzada dels jugadors de tots els equips de primera divisió. Recull les dades més rellevants de l'activitat de cada jugador dins del camp, sigui quina sigui la seva posició.

2. Títol

Dades acumulades de l'activitat dins del camp dels jugadors de futbol de primera divisió.

3. Descripció del dataset

El dataset presenta informació de les activitats executades pels jugadors de futbol professional en el camp durant un partit. En aquest dataset, apareixen tots els jugadors dels 22 equips que formen part de la lliga professional de futbol espanyola.

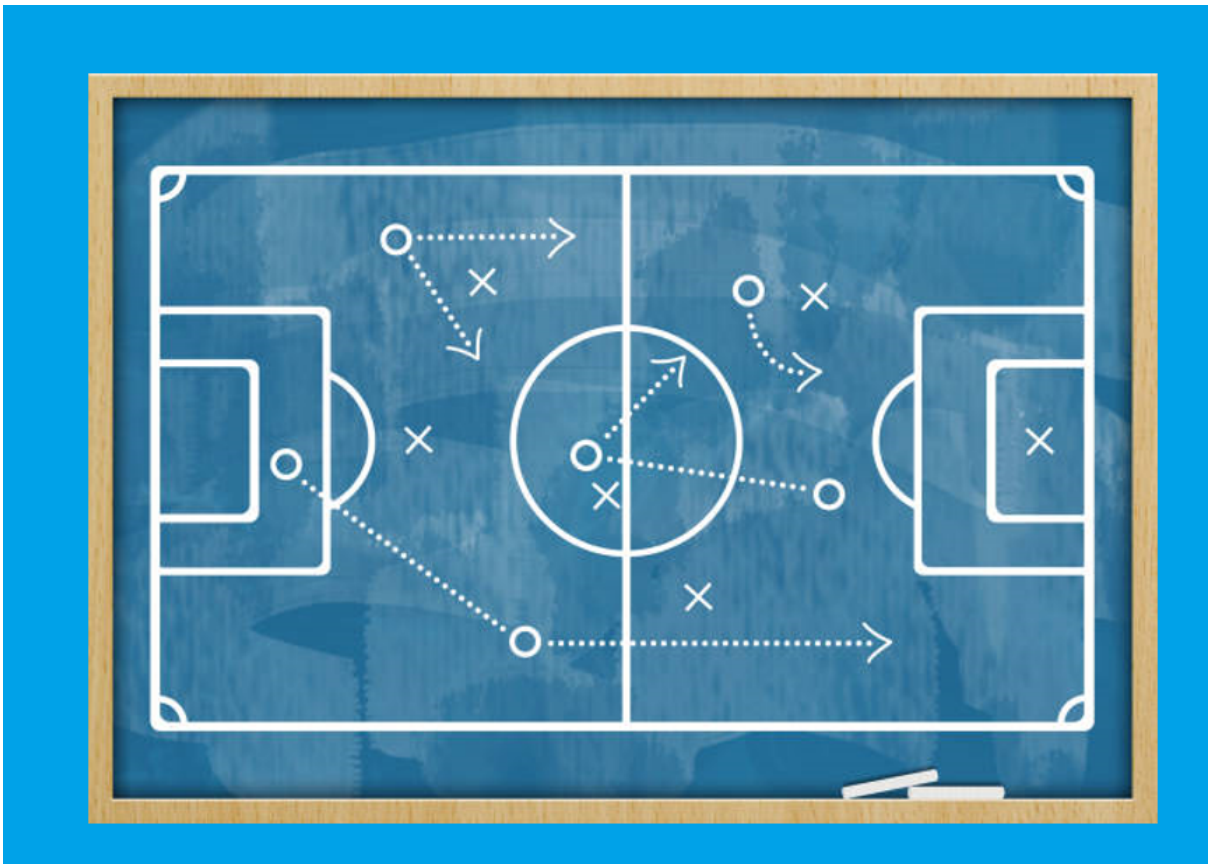
Podem identificar els jugadors per el variable nom. Altres dades descriuen les seves característiques físiques com són l'alçada i el pes. Així mateix, tenim dades personals que fan referència a l'edat i la nacionalitat.

Respecte a les dades que penalitzen els jugadors per infraccions comeses tenim els camps de les targetes grogues i targetes vermelles. Tanmateix, podem trobar dades que descriuen les infraccions realitzades o rebudes per altres jugadors com són, les faltes comeses i les faltes rebudes.

Altres dades descriuen les habilitats del joc dels jugadors com són les assistències, els gols marcats o els llançaments a porta. També podem saber la posició sobre el camp de l'esportista.

Podem saber la càrrega acumulada dels partits jugats per un jugador per les dades que fan referència a la titularitat o suplència d'aquest.

4. Representació gràfica



5. Contingut

Les següents dades corresponen a l'activitat dels jugadors en el camp:

- **Nom:** Nom del jugador.
- **Pos**: Posició on du a terme un jugador la seva activitat en el camp.
- **Edat**: Edat del jugador.
- **Est**: Alçada del jugador.
- **Nac**: Nacionalitat del jugador.
- **Ap**: Els partits que el jugador ha començat jugant com a titular.
- **Sub**: Els partits que el jugador ha participat com a suplent.

- **G:** Total de gols marcats.
- **A:** Assistències realitzades.
- **TT:** Tirs realitzats pel jugador.
- **TM:** Tirs a porteria realitzats pel jugador.
- **FC:** Faltes comeses pel jugador.
- **FS:** Faltes rebudes per un jugador
- **TA:** Targetes grogues acumulades
- **TR:** Targetes vermelles acumulades
- **GA:** Gols rebuts pels porters.

El període de temps que s'han recollit les dades equival al comprès des de la jornada 1 del campionat amb data (14, 15 i 16 d'agost del 2021) fins a la jornada 12 (5, 6 i 7 de novembre del 2021), ja que és tracta de dades que es van actualitzant cada jornada del campionat.

Mitjançant l'arxiu robots.txt, visualitzem els permisos oferits pel propietari de la web. Iniciem el scraping accedint a la web on apareixen el llistat de clubs https://espndeportes.espn.com/futbol/equipos/_/liga/ESP.1/primera-divisi%C3%B3n-de-espa%C3%B1a. En aquesta pàgina hem de recórrer cada un dels clubs que formen part de la lliga, i obtenir l'enllaç de la pàgina web on apareix el llistat de jugadors que formen part de l'equip. En aquest entorn trobem les dades de les activitats de camp dels esportistes. Aquest conjunt de dades són obtingudes pel nostre web scraping.

6. Agraïments

El propietari del conjunt de dades és **Disney Interactive Media Group** i està administrada per **Disney Interactive Media Group**. Aquest grup forma part de **The World Disney Company Limited**.

Aquesta web recull principalment les dades actualitzades de diferents esports a títol informatiu, com per exemple futbol, bàsquet, futbol americà, boxa, tenis, etc. A la seva política de permisos fa menció de la propietat de les dades amb les quals es

treballa. Com que no hi una finalitat comercial en aquest treball, no cal avisar a l'empresa propietària de les dades.

7. Inspiració

Aquest conjunt de dades és interessant per diversos motius. El fet que les dades siguin actualitzades a cada jornada, fa que puguin ser utilitzades en diferents àmbits per la presa de decisions i així obtenir un avantatge competitiu.

Un primer àmbit pots ser el del staff tècnic d'un club, en la gestió diària dels entrenaments i els dies de partit. Per exemple en el tema de les lesions dels jugadors, el fet que a les dades aparegui la càrrega de partits de cada jugador, si ha sortit com a titular o com a suplent, el cansament físic es va acumulant al llarg de la temporada en els jugadors. Aquest seguiment pot ajudar a l'equip tècnic a dosificar els jugadors segons els partits per evitar futures lesions.

En el món del mercat dels fitxatges també es poden utilitzar aquest tipus de dades, ja que el fet de tenir el seguiment de cada jugador, queden reflectides tant les seves habilitats més destacades com els seus punts febles. Per exemple, si agafem com a exemple un jugador amb les qualitats de Leo Messi, que pràcticament està arribant al final de la seva carrera esportiva, com es pot tornar a descobrir un jugador com aquest?, és pràcticament impossible, ja que és considerat un dels millors, sinó el millor jugador de la història d'aquest esport. El que podria fer la directiva i la direcció tècnica del club és, amb l'ajuda del seu historial de dades, es buscar jugadors amb habilitats semblants i que tinguin potencial per omplir el seu buit.

Dins dels clubs, també es poden fer servir aquest tipus de dades dins del departament de màrqueting, i en concret en el món del merchandising. És sabut que hi ha jugadors que venen més camisetes i tota mena d'objectes amb la seva imatge que d'altres. Això és degut al seu rendiment dins del camp que queda reflectit en les seves dades del seu rendiment en el joc, juntament amb el carisma i personalitat del mateix jugador.

En l'àmbit de la premsa esportiva, ja sigui en diaris, premsa escrita, televisió o ràdio tenir les dades actualitzades es claus. Cada grup empresarial fa la seva pròpia recollida de dades i el seu tractament posterior.

8. Llicència

Hem seleccionat la llicència Released Under CC BY-NC-SA 4.0 License pel dataset resultant. Aquesta llicència permet a altres combinar, ajustar i construir a partir del seu treball amb finalitat no comercial. Això sempre que es reconegui l'autoria i les seves noves creacions estiguin sota una llicència amb els mateixos termes.

9. Codi

El codi font escrit en llenguatge Python per realitzar el webscraping és accessible mitjançant el següent enllaç al repositori GitHub:

https://github.com/jpoumas/Practica-1---Tipologia-i-cicle-de-vida-de-les-dades/blob/main/src

10. Dataset

El dataset obtingut en format CSV està publicat a Zenodo i el DOI obtingut és el següent:

`10.5281/zenodo.5651980`

Contribucions	Signatura
Investigació prèvia	X.B.M. - J.P.M.
Redacció de les respostes	X.B.M. - J.P.M.
Desenvolupament del codi	X.B.M. - J.P.M.