Geli Taberner, Albert
Grau Escolano, Jordi

Tipologia i cicle de vida de les dades - PRA 1

1. Context

L'*Sport Analytics* és una branca del món esportiu que, en els darrers anys, ha guanyat molt de pes en la presa de decisions dins d'aquest. És un camp força arrelat (per no dir molt) en disciplines com podrien ser el baseball, el bàsquet o el futbol americà, disciplines en les quals hi ha molts punts per partit i fases de joc clarament definides. En els darrers anys aquest s'ha obert camí, com no podia ser de cap altra manera, en el considerat esport rei: el futbol. Aquesta branca del *Sport Analytics* és la més recent de totes, ja que degut a la naturalesa estocàstica del joc i a la poca claredat en les fases d'aquest, el seu anàlisi es va fer una mica més de pregar. Tot i això, tal i com va comentar Patrick Lucey (Chief Data Scientist a Stats Perform) al webinar *AI for Soccer*, aquest joc és, tal i com s'ha demostrat, menys estocàstic del que sembla.

En futbol (igual que en qualsevol altre esport) sempre ha existit la figura de l'analista: aquella figura que analitzava, en el cas del futbol, estadístiques de pilota parada, pèrdues i recuperacions de pilota, gols, xuts, xuts entre els 3 pals, parades, etc. Però aquestes dades no expliquen una història, i aquesta és la funció del Data Science, no només en l'esport, sinó a cada un dels llocs on s'aplica. Així doncs, basar-se en les pilotes que perd un migcampista defensiu pot ser una estadística enganyosa ja que, com s'ha vist aquesta temporada, Sergio Busquets (jugador del F.C. Barcelona) presenta dues cares d'una mateixa moneda: mentre sembla que amb el Barça és una molèstia i un jugador que no aporta, amb la selecció espanyola s'ha convertit en el millor jugador del torneig *UEFA Nations League*. Donar explicació a aquest fenomen és una de les funcions de la ciència de dades en l'esport.

Actualment, en el futbol d'alt nivell, es recullen dades de cada esdeveniment d'un partit (passades, xuts, faltes, inicis de pressió...), qui efectua l'esdeveniment, qui n'és el beneficiari (en el cas d'una passada, per exemple), com acaba aquest esdeveniment,

distància, angles i posició del porter en una situació de xut etc. Però no només això: també es recull la posició de la pilota i els 22 jugadors en cada esdeveniment. El fet de recollir el "mapatge" del camp és el que aporta diferencialitat d'aquestes dades (anomenades tracking data) respecte les de tota la vida (event data).

Al final, amb totes aquestes dades es poden fer des d'anàlisis "més senzills" (remarco les cometes, perquè de senzill no en té res, ja que, per a un sol partit, podem tenir, de mitjana, de l'ordre de 10^5 registres d'esdeveniments) com podrien ser mapes de calor, percentatge de pressió alta en cada un dels terços en què es divideix el camp; però també s'apliquen models de Machine Learning com podria ser de clustering per a obtenir la formació mitjana d'un equip al llarg del partit (o de diferents trams). Aquestes dades són les que permeten endinsar-se en la història: si comparéssim la distribució de jugadors entorn a Busquets en el Barça i la selecció espanyola, veuríem que Busquets es troba completament sol al centre del Camp Nou (una sentència quasi de mort per un jugador que no té la velocitat per atribut principal), mentre que amb la selecció es troba compactat amb la resta de migcampistes, el que li permet exercir les funcions que l'han permès ser un dels millors migcampistes del món, sense haver de preocupar-se per recórrer grans distàncies.

Altres exemples podrien ser la creacions de mètriques com la d'*expected goals*, que és la probabilitat de que un xut acabi en gol donades una distància i angles a porteria, i la posició del porter. Però no només acaba aquí: l'Sport Analytics ha nascut també per a la capacitat de predir i evitar lesions (o conèixer patrons en aquestes) i, fins i tot, per estudiar l'entorn d'un futbolista i predir si el fitxatge triomfaria. Com es pot comprovar, l'Sport Analytics és un camp fascinant, en ple auge i amb un món sencer de possibilitats

2. Títol

El projecte s'anomena: "**Big 5 European football leagues: teams and players stats**" i els datasets es diran: "**Big 5 European football leagues teams stats**" i " **Big 5 European football leagues players stats**".

3. Descripció del dataset

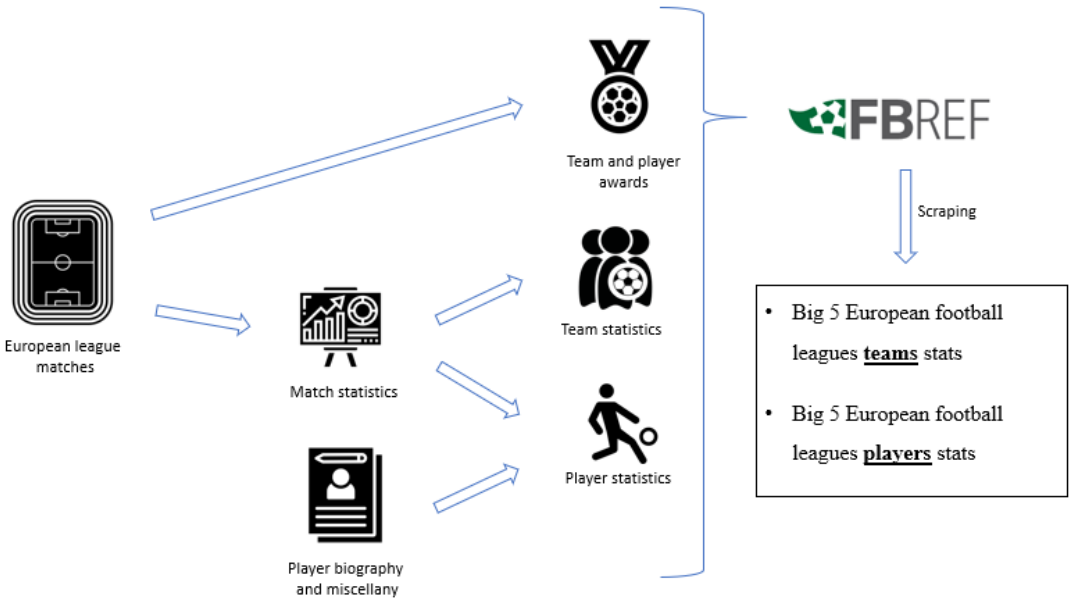
Amb la intenció de facilitar la extracció del coneixement i la presa de decisions que pugui crear un avantatge competitiu a un equip de les 5 grans lligues europees del futbol masculí (Premier League, La Liga, Ligue 1, Bundesliga i Serie A), s'han recollit de la pàgina web https://fbref.com/ dos tipus d'informacions que provenen del període inclòs entre les temporades 2010-2011 i la 2020-2021 (ambdues incloses).

Per una banda es crearà una taula on les dades estaran centrades en la classificació general de les temporades de cada lliga. La informació que hi podrem trobar inclou aspectes com la classificació final de la competició i estadístiques de rendiment dels equips.

Per l'altra banda es crearà una altra taula de jugadors on s'inclogui exclusivament la informació centrada en les temporades i les lligues del projecte. El tipus de dades que hi trobarem serà d'estadístiques de rendiment, de temps de joc, l'equip i la temporada, i a més alguns camps de tipus biogràfic. És important destacar que la presència dels camps temporada i equip en aquesta taula permetrà poder relacionar el rendiment d'un equip amb les contribucions personals de cada jugador.

4. Representació gràfica

Seguidament es mostra un gràfic d'on provenen les dades del projecte:



5. Contingut

La informació referent a la taula de jugadors ha estat capturada de la pàgina web https://fbref.com/en/players/ mitjançant les llibreries Urlib, BeautifulSoup i Pandas en el 29 de setembre del 2021. S'inclouen dades des de la temporada 2010-2011 a la 2020-2021 (ambdues incloses) per a les 5 grans lligues europees masculines (Premier League, La Liga, Ligue 1, Bundesliga i Serie A). Els camps de la taula de jugadors són:

1. **name (string):** nom no complet
2. **c_name(string):** nom complet
3. **position(string):** posició del camp on juga
4. **footed(string):** cama bona (dreta/esquerra)
5. **height(string):** alçada
6. **weight(string):** pes
7. **birth_year(integer):** any de naixement
8. **player_country(string):** país d'origen del jugador
9. **Season(string):** temporada del registre

A partir d'aquest punt tots els camps següents són respecte la temporada en el camp Season.

10. **Age(integer):** edat del jugador
11. **Squad(string):** equip en el que va jugar
12. **Country(string):** país en el que va jugar
13. **Comp(string):** lliga en la que va jugar
14. **LgRank(string):** classificació final de l'equip del jugador en la temporada
15. **MP(integer):** partits jugats
16. **Starts(integer):** partits jugats com a titular
17. **Min(string):** minuts jugats
18. **90s(float):** minuts jugats dividit per 90
19. **Gls(integer):** gols marcats
20. **Ast(integer):** assistències
21. **G-PK(integer):** gols marcats que no provenen de penals
22. **PK(integer):** gols marcats de penals
23. **PKatt(integer):** xuts de penal totals (marcats i no marcats)
24. **CrdY(integer):** targetes grogues
25. **CrdR(integer):** targetes vermelles

L'extracció d'informació per a la creació de la taula d'estadístiques dels equips també ha estat obtinguda mitjançant les llibreries *urllib* i *BeautifulSoup*, a més a més de requerir el mòdul *re* per l'extracció de patrons en els links de cada temporada.
Aquesta taula recull 28 estadístiques de cada equip de les 5 grans lligues per cada una de les 11 temporades mencionades. Cada registre fa referència a dades d'un sol equip en un context "espai - temps" de competició i temporada. Aquestes són:

1. **competition (string):** Nom de la competició (o lliga)
2. **season (string):** Temporada de la competició
3. **rank (integer):** Posició de la lliga
4. **squad (string):** Nom de l'equip
5. **games (integer):** Nombre de partits disputats

6. **wins (integer):** Nombre de partits guanyats
7. **draws (integer):** Nombre de partits empatats
8. **losses (integer):** Nombre de partits perduts
9. **goals_for (integer):** Nombre de gols a favor
10. **goals_against (integer):** Nombre de gols en contra
11. **goal_diff (integer):** Diferència de gols
12. **points (integer):** Punts assolits aquella temporada
13. **Notes (string):** Comentaris sobre classificacions a competicions europees, descensos, etc.
14. **players_used (integer):** Número de jugadors utilitzats
15. **assists (integer):** Total d'assistències
16. **pens_made (integer):** Gols de penal
17. **pens_att (integer):** Nombre de penals llançats
18. **cards_yellow (integer):** Nombre de targetes grogues
19. **cards_red (integer):** Nombre de targetes vermelles
20. **shots_on_target_against (integer):** Xuts a porteria en contra
21. **saves (integer):** Nombre de parades
22. **clean_sheets (integer):** Nombre de partits amb porteria a zero
23. **shots_on_target (integer):** Nombre de xuts a porteria contrària
24. **games_started (integer):** Nombre de partits començats per cada jugador
25. **games_complete (integer):** Nombre de partits completats per jugador
26. **games_subs (integer):** Nombre de partits no començats com a titular per jugador
27. **unused_subs (integer):** Nombre de partits en què no s'ha fet servir un jugador (per jugador)
28. **points_per_match (float):** Mitjana de punts per partit

6. Agraïments:

Mitjançant els mòduls *builtwith* i *whois* determinem la tecnologia emprada en la web i el propietari d'aquesta. Fent ús de la funció *whois* del segon mòdul (*whos.whois("https://fbref.com/en/")*) descobrim que els propietaris de la web

decideixen romandre anònims; així que no podem agrair-los-hi (almenys citant-los per nom i cognom), de part de tots els aficionats a aquest esport, la gran feina realitzada en la gestió de dades de tantes competicions (lligues, copes de club nacionals, copes de club internacionals, competicions internacionals de seleccions com JJOO, mundials, etc. tant masculines com femenines) procedents de les confederacions i associacions de futbol de tot el mon: europeu, sud-americà, asiàtic, africà i centre i nord-americà. A banda de recollir aquesta informació sobre tants equips, també ho fa pels jugadors que han militat alguna vegada en un d'aquests. En definitiva: una bestialitat; es podria quasi considerar la viquipèdia de les estadístiques esportives.

Tot i no ser l'única pàgina que es dediqui a la publicació i organització d'aquest tipus de dades (com podria ser el web *soccerstats.com,* que fan el mateix), ens hem decantat per *fbref.com* ja que aquesta és la que, al nostre parer, oferia un nivell d'informació més detallat i extens; tot i que això hagi suposat no seguir els suggeriments del fitxer *robots.txt* en què es restringeixen les pàgines corresponents a jugadors i competicions (entre moltíssimes altres).

És justament aquest fet el que ens ha fet escollir una llicència que permeti la descàrrega, ús i manipulació dels datasets creats, però no la remuneració a partir de l'ús d'aquesta amb l'objectiu de reconduir el nostre comportament a l'hora d'ignorar els suggeriments que se'ns fan respecte el procés de *scrapping*. Ja que nosaltres hem decidit tirar endavant amb el procés d'extracció de dades ja que hem considerat que, en un camp en tanta expansió (*Sport Analytics*) dins d'un món tant fraternal (*Data Science*) era important que la gent pogués disposar de les dades per experimentar, hem cregut convenient també no limitar-ne l'ús i modificació, però sí establir que ningú se'n pugui beneficiar: escollim desobeir els suggeriments per incentivar i fer avançar la recerca, no per monetitzar-la.

El nivell de detall de la informació oferida ha fet d'aquesta pàgina una de les recurrents en l'anàlisi futbolístic com podria ser l'evolució de gols esperats per partit del Manchester United (https://maramperninety.medium.com/yes-powerpoint-e0e1fc6bbd3f), anàlisis visuals de xifres de gols i assistències per veure quins jugadors són els més decissius (https://thechelseaspot.com/2020/05/21/how-to-create-a-football-data-graph-in-2-minut

es/), un anàlisi comparatiu entre gols i gols esperats per veure quin jugador actua amb un nivell superior al qual s'esperaria (el que es coneix com *over-performance*) (https://www.invertedwinger.com/football-analytics-using-r-and-fbref-data-part-1/), una visualització de percentatge de pilotes recuperades aplicant pressió enfront el nombre d'accions de pressió aplicades (https://www.google.com/imgres?imgurl=https%3A%2F%2Fpbs.twimg.com%2Fmedia%2FEVpY8mkXYAANHUi.jpg&imgrefurl=https%3A%2F%2Ftwitter.com%2Ffbref%2Fstatus%2F1250845187217281025%3Flang%3Dfi&tbnid=MXk-r_XYKXPSfM&vet=12ahUKEwj0jJW1wYT0AhUQRBoKHe25BXsQMygGegQIARA1..i&docid=GVPgybP_aPMjlM&w=940&h=829&q=football%20data%20visualization%20fbref&ved=2ahUKEwj0jJW1wYT0AhUQRBoKHe25BXsQMygGegQIARA1) o gràfics de radar per a la comparativa de jugadors (https://medium.com/the-amateurs/comparing-liverpools-full-backs-using-statsbomb-s-radar-chart-fc64c934d76b).

7. Inspiració

Aquest joc de dades és interessant perquè ens permetrà fer alguns anàlisis com els presentats (com podria ser el gràfic de dispersió gols-assistències per veure els jugadors decissius de cada equip), modificacions d'alguns dels vistos com l'evolució del nombre de gols esperats al llarg de les temporades (podria servir per veure la davallada d'equips com el Barcelona des de que van marxar Pep Guardiola i Luis Enrique); així com infinitat de gràfics analítics com dispersions de xuts a porta (a favor i en contra) enfront de gols (a favor i en contra) per observar el rendiment cara a porteria (cas a favor) i el paper dels porters (cas en contra), també es pot modificar els gràfics de radar perquè compari equips (com rivalitats Barça - Madrid, Milan - Inter de Milan, City - United…) A banda d'aquests anàlisis de caire més avançats, podríem mirar qui acostuma a regnar en les 5 grans lligues, podríem també aplicar tècniques de *clustering* per trobar perfils de jugadors similars per al *scouting* (fitxatge de jugadors) i, fins i tot, la possibilitat de crear models per a predir la possible classificació a competicions europees, possibles descensos, etc. segons el rendiment ofert. També es pot estudiar la distribució d'estadístiques en les lligues per veure si factors com l'edat dels jugadors de la lliga els

gols, targetes, xuts, etc. influeixen en el rendiment col·lectiu i/o en aconseguir millors resultats. Veiem, aquí també, una gran possibilitat d'anàlisis diferents.

8. Llicència

Aquest projecte serà publicat a un repositori públic Git sota la llicència "Attribution-NonCommercial-ShareAlike 4.0 International". Un cop aplicada aquesta llicència a un treball (en aquest cas els datasets), es podrà utilitzar, manipular i construir sobre el producte del projecte sempre i quan es faci referència a l'autor del projecte, i el més important, no es podrà obtenir un rendiment econòmic d'un producte que derivi del projecte amb aquesta llicència. A més, el nou producte haurà d'estar publicat sota els mateixos termes en quant a llicència.

La raó per la que deixem aquest projecte el màxim de públic possible és perquè creiem que és gràcies a la col·laboració de tota la comunitat de data science que el camp està avançant tant en els últims anys. Sense l'existència d'organitzacions com kaggle, UCI, observablehq, OpeanAI, biosoftics, etc la transferència de coneixements, consells, eines i datasets no hagués estat possible, i per tant moltes persones no haguessin pogut aprendre i col·laborar en el constant avanç en que es troba avui en dia el camp de data science. Seguint aquesta filosofia de col·laboració, hem decidit posar a disposició de tothom els datasets creats en aquest projecte.

Per altra banda també hem escollit aquesta llicència per a evitar l'ús comercial de la informació que hi apareix. D'aquesta manera es podrà evitar que s'obtingui un rendiment comercial a una informació que l'arxiu *robots.txt* de la pàgina web demana que no s'extragui mitjançant scraping.

Taula de contribucions

Contribucions	**Signatures**
Investigació prèvia	AGT, JGE
Redacció de les respostes	AGT, JGE

Desenvolupament del codi	AGT, JGE