

Pràctica 1 (25% nota final)

Òscar Busquets Garcia i David Malvesí José

1. **Context.** Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.



Al llarg dels darrers anys a Europa hi ha hagut equips de futbol que han destacat moltíssim. Per exemple està el Barça de Guardiola o el Bayern de Nagelsmann. En tots aquests anys s'ha recordat el nom de jugadors icònics com Lionel Messi, però també n'hi ha hagut molt bons que no han destacat tant a les portades dels diaris.

Com qualsevol director esportiu d'un club que analitza jugadors per a si interessa fitxar-los, es voldrà estudiar tot un rang de jugadors amb les seves posicions i arribar a obtenir un 11 competitiu de forma analítica i no pas passional. Actualment els entrenadors demanen jugadors després de parlar amb ells, veure'ls jugar, o pujar-los de categories inferiors. Aquestes demandes no solen basar-se mai en cap model ni justificació matemàtica.

Fixem-nos per exemple en el cas d'Mbappé, que el pretén el Reial Madrid. Aquest jugador porta les darreres temporades marcant molts gols en el Paris Saint Germain, però realment aquesta dada el fa bo? Potser és que té un bon equip darrere que fan assistències o bé només marca de pena màxima? Seria bo analitzar en profunditat la viabilitat d'aquest jugador en el Madrid abans de fitxar-lo.



Amb tot això es cerca un data set actualitzat dia a dia que engloba les dades de jugadors i equips de futbol de les grans lligues europees: Understat.

2. **Títol.** Definir un títol que sigui descriptiu pel dataset.

Estadístiques de jugadors de futbol d'Europa

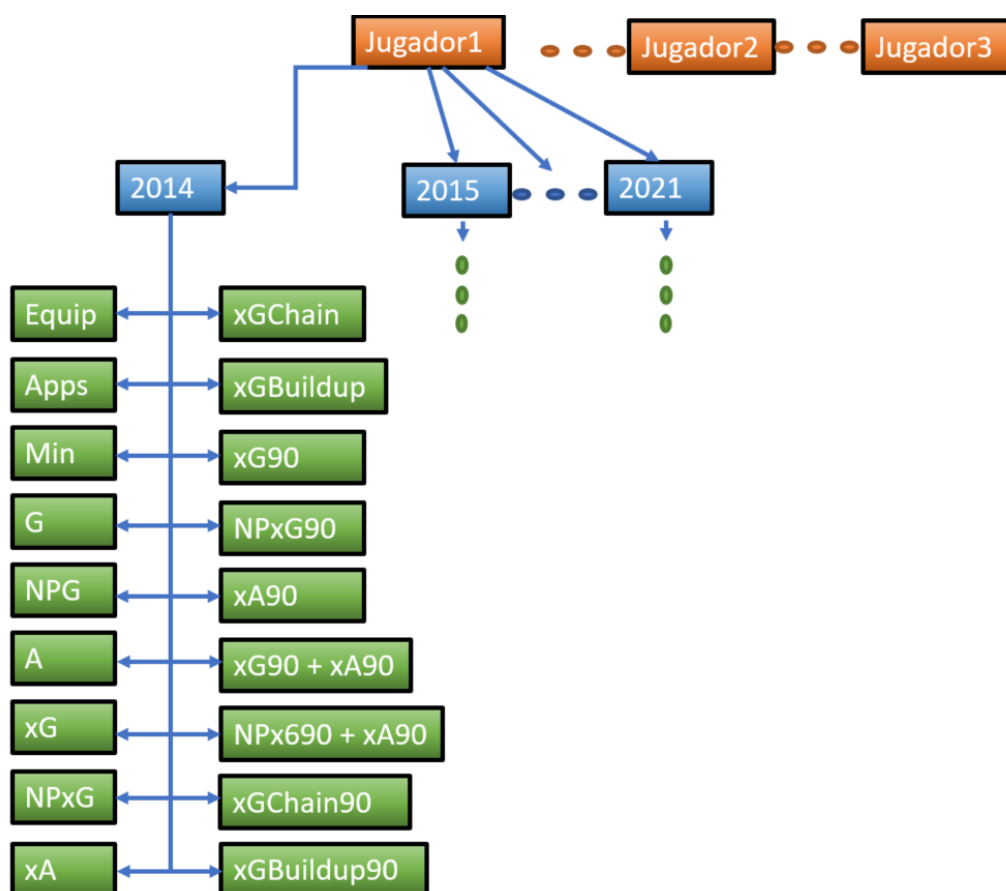
3. **Descripció del dataset.** Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.

El dataset Understat conté les dades principals dels jugadors de tot Europa, des del 2014 fins l'actualitat. No només mostra els gols o minuts jugats sinó que va més enllà

intentant predir quants gols marcarà o quants minuts acabarà jugant. Es tracta d'un dataset força complet que intenta mostrar les xifres dels jugadors de forma analítica.

Nº	Player	Team	Apps	Min	G	A	xG	xA	xG90	xA90
1	Karim Benzema	Real Madrid	11	975	10	7	6.29 ^{-3.71}	2.45 ^{-4.55}	0.58	0.23
2	Luis Suárez	Atletico Madrid	12	796	7	1	5.86 ^{-1.14}	0.78 ^{-0.22}	0.66	0.09
3	Raül de Tomás	Espanyol	12	1058	7	2	6.49 ^{-0.51}	0.86 ^{-1.34}	0.55	0.06
4	Vinicius Júnior	Real Madrid	12	919	7	2	4.96 ^{-2.04}	1.64 ^{-0.38}	0.49	0.16
5	Memphis Depay	Barcelona	12	1080	6	2	6.34 ^{-0.34}	3.51 ^{-1.51}	0.53	0.29
6	Mikel Oyarzabal	Real Sociedad	8	656	6	1	6.00	1.15 ^{-0.15}	0.82	0.16
7	Falcao	Rayo Vallecano	8	331	5	0	2.72 ^{-2.28}	0.01 ^{-0.01}	0.74	0.00
8	Joselu	Alaves	12	951	5	1	4.34 ^{-0.66}	0.81 ^{-0.19}	0.41	0.08
9	Juanmi	Real Betis	11	587	5	1	2.80 ^{-2.20}	0.34 ^{-0.66}	0.43	0.05
10	Iago Aspas	Celta Vigo	13	1058	5	1	4.97 ^{-0.03}	2.04 ^{-1.04}	0.42	0.17

4. **Representació gràfica.** Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.



En aquesta imatge s'observa l'esquema del dataset així com les variables.

5. **Contingut.** Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

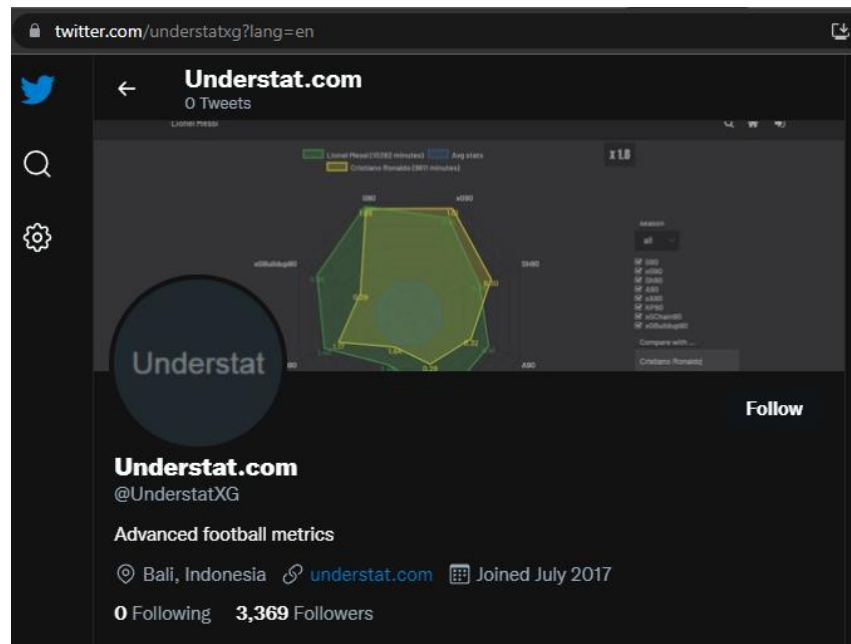
Les dades contenen per a cada jugador que ha estat en les grans lligues europees a partir del 2014 els següents camps:

Positions: Last															1 2 3 4 5 ... 45									
All	All games	Start date	End date																					
NP	Player	Team	Apps	Min	G	NPG	A	xG	NPxG	xA	xGChain	xGBuildup	xG90	NPxG90	xA90	xG90 + xA90	NPxG90 + xA90	xGChain90	xGBuildup90					
1	Mohamed Salah	Liverpool	11	990	10	9	7	7.81 ^{-2.18}	7.05 ^{-1.95}	3.67 ^{-3.33}	13.21	5.27	0.71	0.64	0.33	1.04	0.87	1.20	0.48					
2	Jamie Vardy	Leicester	11	943	7	7	1	4.41 ^{-2.58}	4.41 ^{-2.58}	0.76 ^{-0.24}	5.94	0.87	0.42	0.42	0.07	0.49	0.49	0.53	0.08					
3	Michail Antonio	West Ham	10	900	6	6	3	6.54 ^{-0.56}	5.78 ^{-0.72}	2.96 ^{-0.04}	9.87	2.81	0.65	0.58	0.30	0.95	0.87	0.99	0.28					
4	Sadio Mané	Liverpool	11	908	6	6	0	6.68 ^{-0.86}	6.68 ^{-0.86}	1.36 ^{-1.06}	10.92	4.07	0.66	0.66	0.13	0.79	0.79	1.08	0.40					
5	Raphinha	Leeds	10	828	5	5	0	2.16 ^{-2.84}	2.16 ^{-2.84}	2.00 ^{+0.00}	4.35	1.75	0.23	0.23	0.22	0.45	0.45	0.47	0.19					
6	Pierre-Emerick Aubameyang	Arsenal	10	792	4	4	1	4.91 ^{-0.91}	3.38 ^{-0.52}	0.72 ^{-0.28}	3.10	0.70	0.56	0.38	0.08	0.64	0.47	0.35	0.08					
7	Ron Keighan	Tottenham	10	882	4	4	1	3.34 ^{-0.68}	3.34 ^{-0.68}	0.98 ^{-0.04}	5.08	1.82	0.34	0.34	0.10	0.44	0.44	0.52	0.19					
8	Callum Wilson	Newcastle	7	602	4	4	0	3.01 ^{-0.99}	3.01 ^{-0.99}	0.07 ^{+0.07}	2.46	0.06	0.45	0.45	0.01	0.46	0.46	0.37	0.01					
9	Roberto Firmino	Liverpool	8	369	4	4	1	3.56 ^{-0.44}	3.56 ^{-0.44}	0.73 ^{-0.27}	6.88	2.71	0.87	0.87	0.18	1.05	1.05	1.68	0.86					
10	Wendie Renard	Crystal Palace	10	831	4	2	1	3.24 ^{-0.76}	1.72 ^{-0.78}	1.70 ^{-0.70}	5.35	2.17	0.35	0.18	0.18	0.54	0.37	0.58	0.23					
				291	271	213	317.28	248.25	229.68	16.68	2.06	1.25												

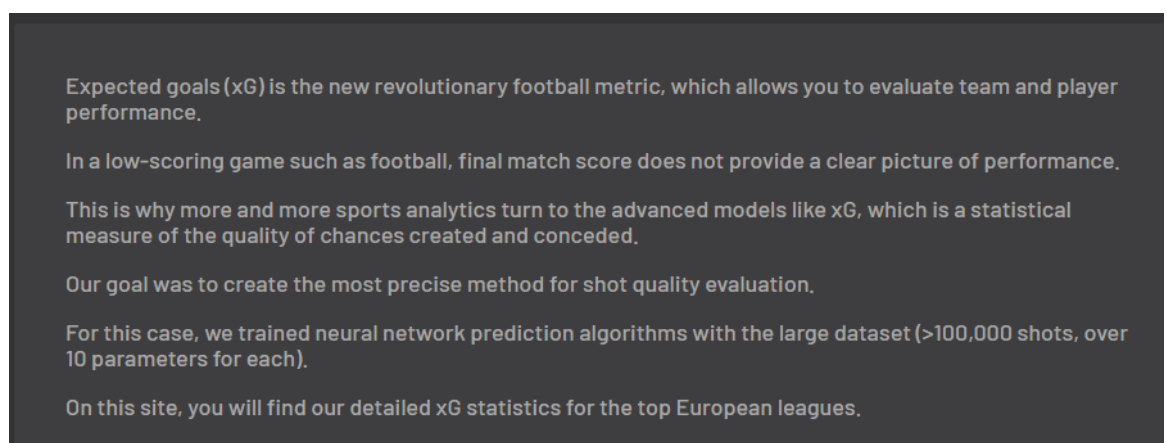
- Equip: Equip que ha jugat aquella temporada
- Apps: Partits jugats
- Min: Minuts jugats
- G: Gols marcats
- NPG: Gols marcats no de penal.
- A: Assistències
- xG: Gols esperats segons l'algoritme d'Understat.
- NPxG: Gols marcats no de penal segons l'algoritme d'Understat.
- xA: Assistències esperades segons l'algoritme d'Understat
- xGChain: Gols estimats per cada possessió en que el jugador està involucrat
- xGBuildup: Gols estimats per a cada possessió en que el jugador està involucrat sense contar assistències.
- xG90: Gols esperats per partit o 90 minuts.
- NPxG90: Gols esperats per partit o 90 minuts sense penes màximes.
- xA90: Assistències esperades per partit o 90 minuts.
- xG90+xA90: Assistències esperades més gols esperats per partit o 90 minuts.
- NPx690+xA90: Assistències esperades més gols esperats sense penes màximes per partit o 90 minuts.
- xGChain90: Gols estimats per cada possessió en que el jugador està involucrat per partit o 90 minuts.
- xGBuildup90: Gols estimats per a cada possessió en que el jugador està involucrat sense contar assistències per partit o 90 minuts.

6. **Agraïments.** Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

El propietari del dataset és totalment desconegut i només es sap que són de Bali, Indonèsia. El seu Twitter també està buit:



En la pròpia pàgina web hi ha un petit text que explica que conté la base de dades que exporten.

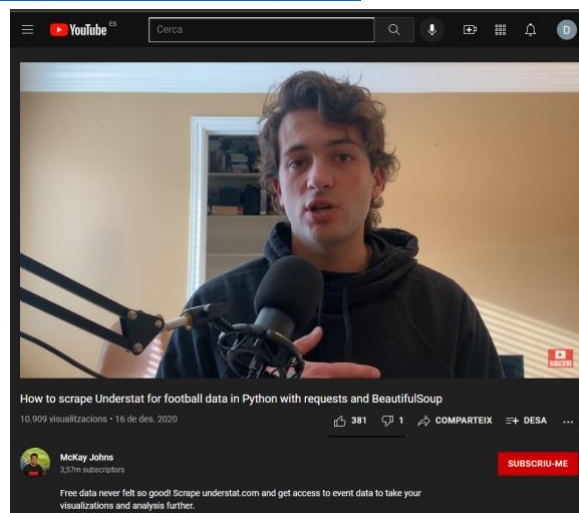


El cas és que en base a possiblement l'altruisme dels propietaris, es coneix molt poc sobre ells.

Malgrat això, a la xarxa hi ha multitud d'exemples d'usuaris d'internet que expliquen com han scrapejat aquesta pàgina i inclús com les han analitzat. Per la nostra part ens hem inspirat en aquests dos exemples de dos internautes que de forma docent mostren com tractar-les.

El primer exemple és l'usuari d'internet McKay Johns que en aquest vídeo de 23 minuts explica com treballar amb la pàgina web. Aquest usuari té un canal de Youtube amb gran quantitat de material tant per a iniciats a l'scrapping com per altres temes de programació.

<https://www.youtube.com/watch?v=lsR5FrjNmro>



L'altre usuari al que li agraïm la inspiració és en Sergi, que el juny del 2019 va redactar en un post com treballar també amb aquesta web. En Sergi però, ho explica d'una forma molt més tècnica i emprant la llibreria que finalment hem usat nosaltres, la BeautifulSoup.

<https://www.sergilehkyi.com/es/web-scraping-advanced-football-statistics/>



7. **Inspiració.** Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

Recordar el que s'ha explicat en la introducció. Actualment l'anàlisi dels jugadors es fa molt visual, els reclutadors tenen un 'sisè' sentit que no sempre funciona. El que es pretén fer amb aquesta base de dades és analitzar profundament els jugadors de forma que es visualitzi les millors opcions i fins i tot treure un onze ideal, i no pas el sempre polèmic 11 que gent de l'organització tria per fanatisme.

Dels anàlisis es mostrats en l'apartat 6 únicament es mostrava com scrapejar les dades sense acabar de definir l'objectiu d'elles. L'enunciat marcat ha estat inspirat per els dos estudiants amb la finalitat d'anàlisi de les dades.

A més a més de l'anàlisi jugador a jugador, també es podria plantejar si analíticament hi havia equips com el Madrid de Zinade si tenien la millor plantilla de jugadors quan va guanyar les copes d'Europa. Un anàlisi també que podria demostrar que no sempre es té la millor plantilla i potser a vegades es qüestió de sort. De ben segur però, la plantilla no era de les pitjors.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció:
 - Released Under CC0: Public Domain License
 - Released Under CC BY-NC-SA 4.0 License
 - Released Under CC BY-SA 4.0 License
 - Database released under Open Database License, individual contents under Database Contents License
 - Other (specified above)
 - Unknown License

Com s'ha comentat en l'apartat 6, el propietari de la base de dades és desconegut i no marca cap llicència sobre les dades. Degut a això, es decideix marcar l'opció de 'Unknown License'. Com que l'ús de les dades serà purament acadèmic i també es presuposa que les dades estan pensades per a aquest propòsit, es procedirà a treballar amb elles.

9. **Codi.** Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
10. **Dataset.** Publicar el dataset obtingut(*) en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.

(*) Si existeix qualsevol impediment per publicar el dataset real, s'haurà de justificar aquesta situació i publicar a Zenodo un dataset simulat. En aquest cas, el dataset real es comunicarà al professor de forma privada (p.ex., enllaç de Google Drive).

Criteris de valoració

Tots els apartats són obligatoris. La ponderació dels exercicis és la següent:

Apartat	1	2	3	4	5	6	7	8	9	10
Punts	0,25	0,25	0,25	0,5	1	1,5	1,25	1	2	2

Criteris que es tindran en compte per a l'avaluació de la pràctica són:

- Idoneïtat de les respostes (hauran de ser clares i completes).
- Complexitat del lloc web triat per a l'extracció.
- Síntesi i claredat, a través de l'ús de comentaris, del codi resultant.
- Presentació adequada de les dades.
- Organització i claredat dels documents de lliurament final.
- Completitud dels documents requerits per al lliurament final.
- Seguiment de recomanacions per al bon ús del web scraping.




Format i data de lliurament

Durant la setmana **del 25 al 29 d'octubre**, el grup podrà fer un lliurament parcial opcional. Aquest lliurament parcial és molt recomanable per rebre assessorament sobre la pràctica i verificar que la direcció presa és la correcta. Es lliuraran comentaris als estudiants que hagin efectuat el lliurament parcial, però no comptaran per a la nota de la pràctica. En el lliurament parcial els estudiants hauran de lliurar per correu electrònic, al professor encarregat de l'aula, l'enllaç al repositori Git amb allò que hagin avançat.

En referència al lliurament final, es demana:

- Un únic document** (.txt, .pdf, .docx) que contingui **l'enllaç al repositori Git** del projecte (apartat b) i **l'enllaç al vídeo del projecte** (apartat c). Aquest document es lliurarà a l'espai de Lliurament i Registre d'AC de l'aula.
- Un repositori Git** amb les solucions de la pràctica. El repositori Git es crearà a Github (<https://github.com/>), i podrà ser un repositori públic o privat, a elecció del grup. Si s'utilitza un repositori privat, s'haurà de facilitar accés al professor, mitjançant el nom d'usuari que s'indicarà al Tauler de l'aula o per email. **El repositori no es podrà modificar passada la data de lliurament**, i haurà de contenir:
 - Una **Wiki** o **README.md** on estiguin els noms dels components del grup, una descripció dels fitxers i el DOI de Zenodo del dataset generat.
 - Un **document PDF** amb les respostes als apartats 1-10 i els noms dels components del grup. **L'extensió d'aquest document no ha de superar les 20 pàgines**. A més, al final del document, ha d'aparèixer la següent

taula de contribucions al treball, la qual ha de signar cada integrant del grup amb les seves inicials. Les inicials representen la confirmació per part del grup que l'integrant ha participat en aquest apartat. Tots els integrants han de participar en cada apartat, per la qual cosa, idealment, els apartats haurien d'estar signats per tots els integrants.

Contribucions	Signatura
Investigació prèvia	 Òscar Busquets i David Malvesí
Redacció de les respostes	 Òscar Busquets i David Malvesí
Desenvolupament del codi	 Òscar Busquets i David Malvesí

b.3. Una carpeta amb el **codi Python** o **R** generat per obtenir les dades.

- c. Un **breu vídeo** amb la participació dels dos components del grup, on es realitzarà una presentació del projecte, destacant els punts més rellevants. El vídeo s'haurà de compartir mitjançant un enllaç del Google Drive de la UOC o incloure-ho al repositori Git. **La durada d'aquest vídeo no ha de superar els 10 minuts.**

El document del lliurament final s'ha de pujar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59 CET del dia 8 de novembre**. No s'acceptaran lliuraments fora de termini.

Si s'estima oportú, el professor sol·licitarà als integrants del grup una entrevista remota (de manera conjunta o individual) mitjançant Google Meet, en referència a la pràctica realitzada, en un dia i hora acordats.