

**Autors:** Adrià Crusi López i Jordi Montserrat López

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

La informació recolectada s'ha extret de la pàgina web *comuniozo.com*. Aquesta web es un portal d'un joc bastant conegut en l'àmbit del futbol.

Aquest joc permet a un grup d'amics crear lligues per a competir entre ells. Cada usuari d'una lliga, ha de crear el seu propi equip a partir de jugadors de la lliga de futbol espanyola i el funcionament es el següent:

Cada setmana, es realitzen partits reals a la lliga espanyola entre els 20 equips més importants del país i depenen del rendiment dels jugadors (gols, assistències, percentatge de passades bones, pèrdues, recuperacions i altres paràmetres subjectius) cada jugador de cadascun dels equips obtenen punts. Com millor juga el jugador, més punts obté.

És per aquest motiu que els usuaris creen equips amb diferents jugadors per tal d'obtenir els màxims punts possibles i poder guanyar la lliga.

Aquest joc, té el seu portal a *comuniozo.com* i per tal de que pugui funcionar ha d'emmagatzemar tot tipus d'estadístiques. Des de jugadors, equips, gols, assistències, punts valor de mercat dels jugadors i moltes altres.

Degut a tota la informació que incorpora, s'ha vist necessari extreure dades d'aquest portal per tal de poder-les analitzar i extreure conclusions de les mateixes.

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

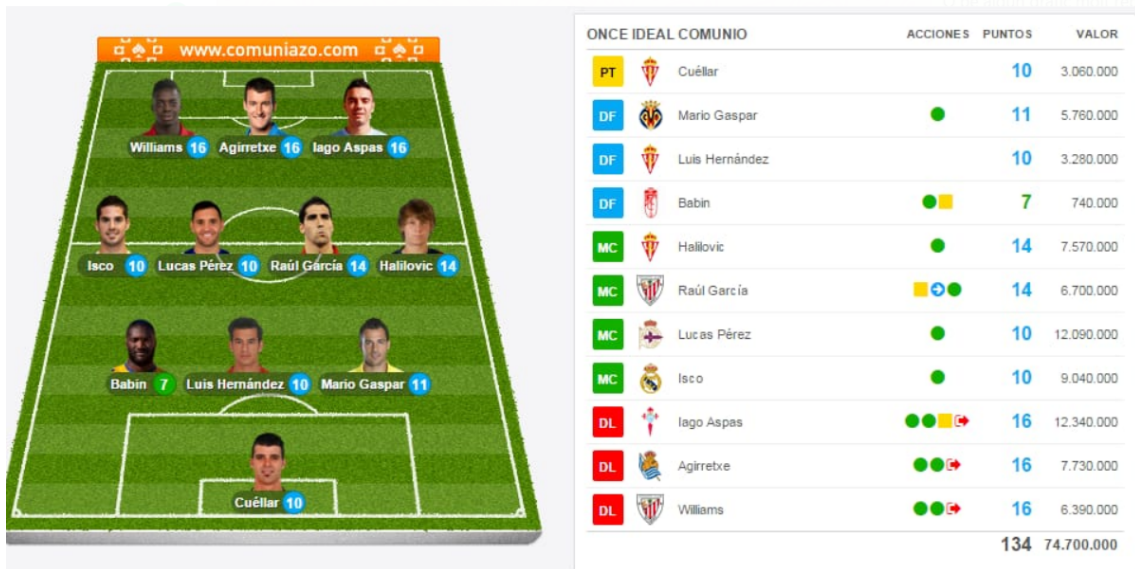
Estadístiques comunió Lliga espanyola temporada 2020-21

*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 triat).*

El *dataset* presentat, és un conjunt de dades i estadístiques dels jugadors que han participat almenys en un partit de la lliga espanyola la temporada corresponent a l'any 2020 i 2021. Donat que quan s'ha realitzat el scraping la lliga encara no ha finalitzat, les dades corresponen a les 30 primeres jornades disputades.

Aquestes dades van des de característiques del jugador com són el nom, edat, posició, equip, alçada ... fins aquelles que permeten veure com està sent el seu rendiment segons el joc; gols, assistències, puntuació, valor dins del joc...

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



The image shows a virtual football team formation on a pitch and a detailed player statistics table. The pitch shows the following players: Cuéllar (Goalkeeper), Williams (Defender), Agirretxe (Defender), Iago Aspas (Defender), Isco (Midfielder), Lucas Pérez (Midfielder), Raúl García (Midfielder), Halilovic (Midfielder), Babin (Midfielder), Luis Hernández (Midfielder), Mario Gaspar (Midfielder), and Cuéllar (Goalkeeper).

	ACCIONES	PUNTOS	VALOR
PT Cuéllar		10	3.060.000
DF Mario Gaspar		11	5.760.000
DF Luis Hernández		10	3.280.000
DF Babin		7	740.000
MC Halilovic		14	7.570.000
MC Raúl García		14	6.700.000
MC Lucas Pérez		10	12.090.000
MC Isco		10	9.040.000
DL Iago Aspas		16	12.340.000
DL Agirretxe		16	7.730.000
DL Williams		16	6.390.000
		134	74.700.000

Aquesta és una imatge és una captura de pantalla d'una jornada a comuniazo, i per tant em considerat que és representativa sobre el nostre dataset.

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

El dataset inclou 378 observacions corresponent als jugadors que han disputat almenys un partit, i un total que 20 variables. Els camps que inclou són els següents:

- **Non de jugador.** Aquesta és la més important, ja que informa del nom del jugador en qüestió
- **Posició.** Aquesta variable indica si un jugador es porter, defensa, migcampista o davanter. És una variable molt important ja que depèn de la posició les estadístiques tenen un pes diferent.
- **Punts (totals, a casa i a fora):** Depenen del rendiment dels jugadors, obtenen més o menys punts.
- **Mitjana de punts:** Permet quins jugadors punten millor per partit jugat
- **Gols:** El número de gols que porta.
- **Assistències:** El número d'assistències que porta.
- **Targetes:** Nombre de targetes vistes.
- **Valor:** Els valor dels jugadors s'estableix en base al que el conjunt de jugador paga per ells en una subhasta a cegues. Generalment, els que fan més punts, tenen més valor i per tant costa més diners comprar-los.
- **Edat**
- **Altura**
- **Pes**
- **Nacionalitat**
- **Valor real:** El valor de mercat del jugador fora del joc.

El període temps de les dades és des de la jornada 1 fins a la 30 de la lliga espanyola de futbol.

A continuació veurem una petita captura de les primeres observacions del dataset:

Nombre	Edad	Altura	Peso	Nacionalidad	Equipo	Posición	Partidos	Minutos por partido	Goles	Asistencias	Targetas	Puntos	Puntos local	Puntos visitante	Media total	Media local	Media visitante	Valor de Comunio	Valor de mercado
Messi	33	1.70	72	Argentina	barcelona	delanteros	28	85	23	8	4	335	174	161	12.0	12.4	11.5	32 140 000	87
Gerard Moreno	29	1.80	75	España	villarreal	delanteros	25	84	19	5	3	246	109	137	9.8	9.1	10.5	17 950 000	34
Benzema	33	1.84	79	Francia	real-madrid	delanteros	26	88	19	6	2	227	120	107	8.7	9.2	8.2	16 650 000	25
Kroos	31	1.83	78	Alemania	real-madrid	centrocampistas	27	75	3	8	6	224	109	115	8.3	8.4	8.2	13 470 000	48
Luis Suárez	34	1.82	86	Uruguay	atletico	delanteros	26	79	19	2	5	213	134	79	8.2	9.6	6.6	17 220 000	14
Iago Aspas	33	1.76	67	España	celta	delanteros	25	88	10	10	5	210	113	97	8.4	9.4	7.5	16 380 000	10
Marcos Llorente	26	1.84	71	España	atletico	centrocampistas	29	78	9	8	5	202	108	94	7.0	7.2	6.7	11 370 000	72
Parejo	31	1.80	74	España	villarreal	centrocampistas	28	89	3	2	4	192	98	94	6.9	7.0	6.7	11 350 000	10
Caseiro	29	1.85	80	Brasil	real-madrid	centrocampistas	27	85	5	2	10	191	102	89	7.1	7.3	6.8	12 400 000	64

*6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-les, justificar aquesta cerca amb anàlisis similars.*

El propietari de les dades és l'empresa comuniao SL registrada a Sevilla amb el NIF B90127275. Així mateix, val a dir que per realitzar aquest treball no s'ha pogut partit de cap altre anàlisis anterior sobre aquests tipus de dades ja que no hem pogut localitzar-ne cap. No obstant, si hem pogut fer-nos una idea del que podem esperar gracies a l'anàlisi realitzat sobre algunes estadístiques del portal fa uns 5 anys. Aquest es pot trobar al següent ellanç:

Miralles, J. M., (2016). Web Scraping, PCA y K-means para sacar todo el potencial a @LaLiga [en línea]. <https://almeriarusers.wordpress.com/>. [Consultado el 1 de abril de 2021]. Disponible en: <https://almeriarusers.wordpress.com/2016/10/11/web-scraping-pca-y-k-means-para-sacar-todo-el-potencial-a-laliga/>

*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.*

Actualment el món de l'esport està intentant aprofitar la capacitat per recollir i analitzar grans volums de dades per tal de poder analitzar el que succeeix durant la pràctica del mateix i buscar formes de millorar el rendiment en base als estudis fets amb aquestes. Aquesta "febre" per les dades no és exclusiva dels propis clubs, els espectadors cada cop més s'està acostumant al tractament de certes dades i interpretació de certes estadístiques relacionades l'esport. Fruit d'això per exemple és el creixement de tots el jocs anomenats "fantasy" on gestionen un equip en base als rendiments del jugador a la vida real, i fa que tinguem informació sobre moltes dades que abans no podíem tenir.

En el nostre cas en particular, el conjunt de dades és important ja que mostra la major part d'estadístiques més importants en el campionat de lliga de futbol. Gràcies a aquestes dades es poden fer múltiples anàlisis per respondre a moltes qüestions dins de l'àmbit de l'esport.

A més, a part d'aquestes dades, que provenen del rendiment real dels jugadors de la lliga de futbol espanyola, també s'hi troben dades i estadístiques provinents del joc 'comunio'. Per tant, amb aquest conjunt de dades es pot fer un doble anàlisi.

Entre els possibles anàlisis a realitzar sobre el dataset podem destacar per exemple les relació entre els equips i el valor, tant a nivell de joc com real per veure fins quin punt els jugadors amb més nom són també els millors dins el lloc, podríem utilitzar també un anàlisi enfocat a veure com influeix la categoria a les estadístiques observades, etc.

Per exemple, en contraposició a l'anàlisi comentat a l'anterior apartat on no s'ha profunditzat en el clúster ni PCA podríem intentar aprofitar aquestes metodologies per realitzar models predictius per posició, equip, valor real de mercat etc.

*8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:*

Hem decidit escollir la llicència de **Mozilla Public License 2.0**, doncs considerem que donada la naturalesa d'aquest projecte és la més adient en la mesura que permet distribuir, modificar les dades, distribuir comercialment les dades i possibles estudis derivats, així com un ús privat d'aquestes, a canvi de que les pròximes contribucions realitzades a posteriori s'hauran de realitzar sota aquesta mateixa llicència, per lo que ens assegurem que el nostre dataset original es seguirà distribuint sota les mateixes condicions. Així mateix, garanteix que el codi font sigui també sigui distribuït.

Contribucions	Signa
Recerca prèvia	ACL i JML
Redacció de les respostes	ACL i JML
Desenvolupament codi	ACL i JML

El codi i el DOI del dataset es poden trobar al següent enllaç:

<https://github.com/jordim14/Comuniazio-Scraping>