Cas Pràctic

de Webscraping

FutStd 2018-2019

LLIGA DE FUTBOL DE 1ª Divisió

TEMPORADA 2018-2019

(font: Diari **MARCA)**

Repositori GitHUB: https://github.com/jjdiezm/Practica1

GRUP PRÀCTICA

Joaquim de Dalmases Juanet (quimdalmases)
Juanjo Díez Moya (jdiezm)

Índex de Continguts

Pràctica 1: Web scraping

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

   El context del nostre dataset és la lliga de futbol de 1ª divisió espanyola d'enguany, temporada 2018-2019. El lloc web és el del diari esportiu "MARCA". El diari proporciona aquestes dades perquè son reflex de com evoluciona la lliga de futbol, i és una manera de satisfer la demanda dels lectors, que volen saber tots els detalls estadístics al detall.

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

   Títol: "**Dades i estadístiques dels equips de futbol de 1ª divisió**"

   On **FutStd 2018-2019** seria l'alies curt proposat a mode de màrqueting, per facilitar la seva identificació.

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 es composa de varis fitxers de dades: 2 fitxers en format CSV i un en format binari.

   El fitxer '**jugadors.csv**', emmagatzema les plantilles de tots els equips. Les plantilles tot i jugar només 11 jugadors estan composades de l'ordre de 25 a 33 jugadors.

   A més a més el dataset també emmagatzema les fotografies dels jugadors ('***fotos.pickle***'). Les fotografies per optimització d'espai a disc, es guarden en format binari 'pickle'. Quan no es disposa de la fotografia d'un jugador, aleshores se li assigna un patró d'imatge no existent enfosquit.

   El dataset ens ha de permetre analitzar el rendiment de cada jugador o del seu equip si acumulem dades per plantilla. També amb les fotografies dels jugadors podem plantejar models de reconeixement de cares.

   Tanmateix el dataset ens pot permetre establir o estudiar coeficients de rendiment, per avaluar el millor jugador de la lliga sota conceptes diferents al jugador que marca més gols.

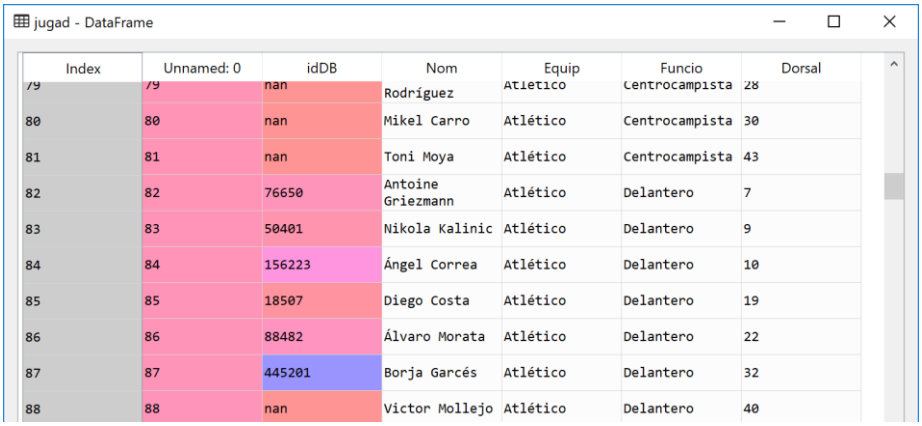
4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment

Figura 1: Logotip del dataset

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

El període de temps de les dades es correspon a la **Temporada 2018-2019**, i l'execució del procés de 'scraping' s'hauria de fer al finalitzar el campionat vigent. En el procés d'organització de les dades un cop realitzat el 'raspat', cal tenir en compte, si compensa en cost, portar a terme operacions de post-processat perquè aquest dataset té una variabilitat molt alta de la composició de les plantilles d'una temporada a l'altre. Una distribució àgil com la plantejada té en compte aquest aspecte.

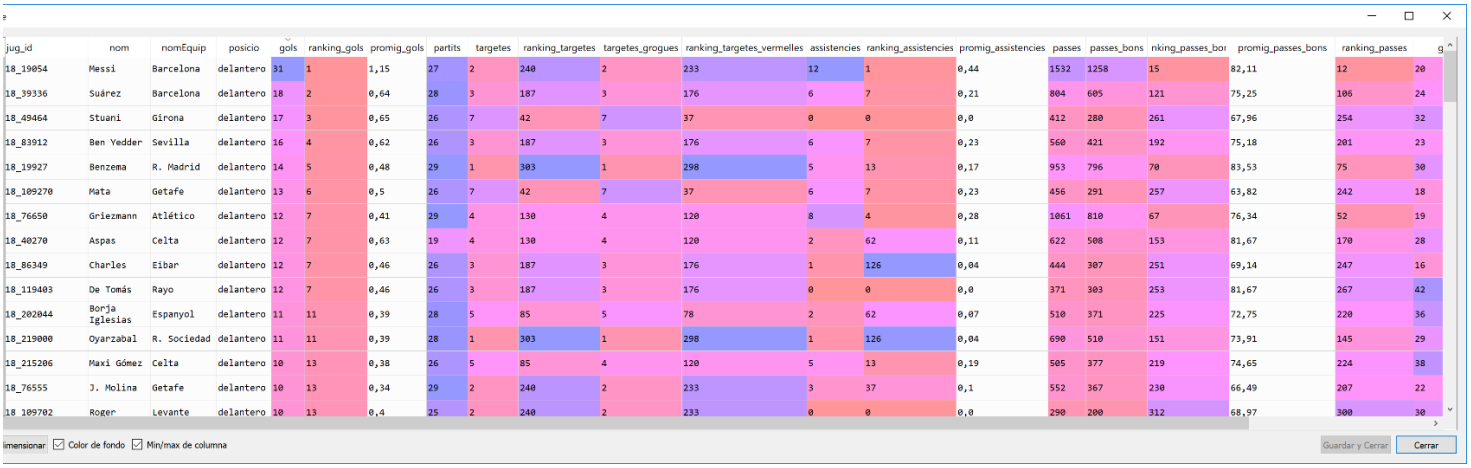
1. Dades alfanumèriques:
   Per una banda les dades personals dels jugadors com un identificador, el nom i el cognom, i dades respecte a la seva funció en l'equip al jugar els partits, com son la seva posició/funció (porter/defensa/mig-campista/davanter) i el dorsal assignat per el club.


jugad - DataFrame

Index	Unnamed: 0	idDB	Nom	Equip	Funcio	Dorsal
79	79	nan	Rodríguez	Atlético	Centrocampista	28
80	80	nan	Mikel Carro	Atlético	Centrocampista	30
81	81	nan	Toni Moya	Atlético	Centrocampista	43
82	82	76650	Antoine Griezmann	Atlético	Delantero	7
83	83	50401	Nikola Kalinic	Atlético	Delantero	9
84	84	156223	Ángel Correa	Atlético	Delantero	10
85	85	18507	Diego Costa	Atlético	Delantero	19
86	86	88482	Álvaro Morata	Atlético	Delantero	22
87	87	445201	Borja Garcés	Atlético	Delantero	32
88	88	nan	Víctor Mollejo	Atlético	Delantero	40

Figura 2: Dades de les plantilles dels equips en format tabular, fitxer **jugadors.csv**.

Per altre banda les dades estadístiques que descriuen el rendiment del jugador:



jug_id	nom	nomEquip	posicio	gols	ranking_gols	promig_gols	partits	targetes	ranking_targetes	targetes_grogues	ranking_targetes_vermelles	assistencies	ranking_assistencies	promig_assistencies	passes	passes_bons	nking_passes_bon	promig_passes_bons	ranking_passes	g
18_19054	Messi	Barcelona	delantero	31	1	1,15	27	2	240	2	233	12	1	0,44	1532	1258	15	82,11	12	20
18_39336	Suárez	Barcelona	delantero	18	2	0,64	28	3	187	3	176	6	7	0,21	804	605	121	75,25	106	24
18_49464	Stuani	Girona	delantero	17	3	0,65	26	7	42	7	37	0	0	0,0	412	280	261	67,96	254	32
18_83912	Ben Yedder	Sevilla	delantero	16	4	0,62	26	3	187	3	176	6	7	0,23	560	421	192	75,18	201	23
18_19927	Benzema	R. Madrid	delantero	14	5	0,48	29	1	303	1	298	5	13	0,17	953	796	70	83,53	75	30
18_109270	Mata	Getafe	delantero	13	6	0,5	26	7	42	7	37	6	7	0,23	456	291	257	63,82	242	18
18_76650	Griezmann	Atlético	delantero	12	7	0,41	29	4	130	4	120	8	4	0,28	1061	810	67	76,34	52	19
18_40270	Aspas	Celta	delantero	12	7	0,63	19	4	130	4	120	2	62	0,11	622	508	153	81,67	170	28
18_86349	Charles	Eibar	delantero	12	7	0,46	26	3	187	3	176	1	126	0,04	444	307	251	69,14	247	16
18_119403	De Tomás	Rayo	delantero	12	7	0,46	26	3	187	3	176	0	0	0,0	371	303	253	81,67	267	42
18_202044	Borja Iglesias	Espanyol	delantero	11	11	0,39	28	5	85	5	78	2	62	0,07	510	371	225	72,75	220	36
18_219000	Oyarzabal	R. Sociedad	delantero	11	11	0,39	28	1	303	1	298	1	126	0,04	690	510	151	73,91	145	29
18_215206	Maxi Gómez	Celta	delantero	10	13	0,38	26	5	85	4	120	5	13	0,19	505	377	219	74,65	224	38
18_76555	J. Molina	Getafe	delantero	10	13	0,34	29	2	240	2	233	3	37	0,1	552	367	230	66,49	207	22
18_109702	Roger	Levante	delantero	10	13	0,4	25	2	240	2	233	0	0	0,0	290	200	312	68,97	300	30

imensionar | Color de fondo | Min/max de columna | Guardar y Cerrar | Cerrar

Figura 3: Dades estadístiques de rendiment, fitxer **Estadistiques_jugadors.csv**.

El conjunt atributs que descriuen totes les estadístiques és el següent:

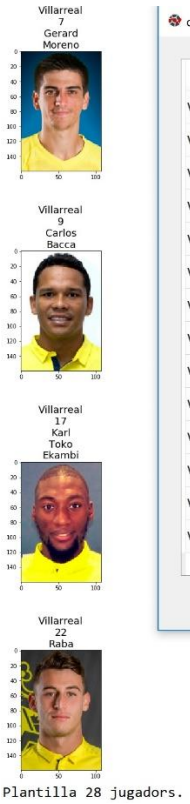
'jug_id'	: id assignat en la base de dades.	Ex: '01_0101_2018_19054'
'nom'	: nom i cognoms del jugador.	Ex: 'Messi'
'nomEquip'	: nom de l'equip on juga.	Ex: 'Barcelona'
'posicio'	: posició en el camp.	Ex: 'Delantero'
'gols'	: nº de gols marcats.	Ex: 31
'ranking_gols'	: posició en ranking de gols marcats.	Ex: 1
'promig_gols'	: promig de gols marcats.	Ex: 1.15
'partits'	: nº de partits jugats.	Ex: 27
'targetes'	: nº de targetes que li han mostrat.	Ex: 2
'ranking_targetes'	: posició en el ranking de targetes.	Ex: 240
'targetes_grogues'	: nº de targetes grogues rebudes.	Ex: 2
'ranking_targetes_vermelles'	: posició en el ranking de targetes vermelles.	Ex: 233
'assistencies'	: nº d'assistències realitzades.	Ex: 12
'ranking_assistencies'	: posició en el ranking de assistències.	Ex: 1
'promig_assistencies'	: promig d'assistències.	Ex: 0.44
'passes'	: nº de passes.	Ex: 1532
'passes_bons'	: nº de passes bons.	Ex: 1258
'ranking_passes_bons'	: posició en el ranking de passes bons.	Ex: 15
'promig_passes_bons'	: promig de passes bons.	Ex: 82.11
'ranking_passes'	: posició en el ranking de passes.	Ex: 12
'gols_ok'	: nº de gols comptabilitzats en la base de dades.	Ex: 20
'ranking_gols_ok'	: posició en el ranking de gols ok.	Ex: 170
'promig_gols_ok'	: promig de gols ok.	Ex: 0.74

Els dos datasets poden unir-se per el codi de jugador, però cal tenir en compte que no tots els elements de les plantilles tenen codi , només els jugadors identificats. Per exemple entrenadors i jugadors sense fotografia no tenen 'codi', i que els entrenadors no tenen dorsal. També caldrà tenir en compte que les estadístiques son dels 100 millors jugadors, i no de tots els jugadors de la lliga, que son en total 576.

S'han elaborat tres datasets addicionals pel seguiment dels màxims golejadors (Pichichi), millors porters (Zamora) i millors entrenadors (Miguel Muñoz)

2. Dades binaries

Les fotografies dels jugadors han estat emmagatzemades en format binari. En la següent figura podem veure el format intern de les dades un cop les recuperem de disc i les disposem en variables de Python:



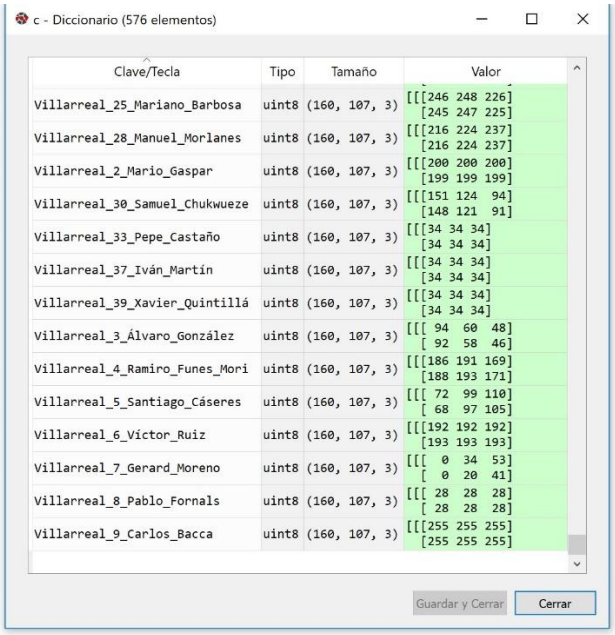


Figura 4: Imatges recuperades i matrius de píxels RGB de dimensions 160x107 píxels.

S'ha aprofitat la clau del diccionari per emmagatzemar les dades de l'equip, dorsal i nom i cognoms de cada jugador i identificar cada fotografia inequívocament.

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

   El diari MARCA és líder a Espanya en informació esportiva en format paper i digital. Cada temporada ofereix les estadístiques més rellevants que es poden obtenir, i que sense l'esforç dels mitjans de comunicació seria quasi inviable disposar d'aquestes dades.

7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

   L'interès de les dades té varies vessants, que van des de la purament lúdica per satisfer la curiositat dels aficionats i generar opinió de qui destaca o qui ha de millorar en el seu rendiment, a la estadística informativa d'un diari esportiu que basa el seu contingut en la seva publicació, i la utilitza com a font d'ingressos. L'interès des de el punt de vista de fer el scraping l'exposem a continuació.

   Les preguntes a respondre o les possibles opcions d'aplicació en posteriors projectes d'anàlisi de dades proposades per ordre d'importància son les següents:

   1. Conèixer qui fa millor i pitjor cada aspecte rellevant del joc:

      Gols, aturades, passades correctes, més minuts jugats, rànquings, mitjanes etc...

Aquestes característiques poden ser analitzades individualment o per equips.

Podem arribar a determinar qui es el millor jugador o el millor equip (col·lectivament per l'acumulació d'estadístiques). Com a repte es podria intentar desenvolupar un índex o coeficient de valoració de rendiment més complex, que el nº de gols marcats, per decidir qui és el millor jugador de la lliga.

2. Construir models i fer prediccions sobre el que pot arribar a fer un equip o jugador, en una temporada següent.
3. Amb algorismes com les regles d'associació, generar models adients per intentar esbrinar quins jugadors juguen millor plegats, ja sigui per posicions d'equip (defensa, mig de camp o davantera) o per rendiment global d'equip.
4. Aplicar algorismes de 'clustering' i agrupar els jugadors per similitud de característiques en tasques no supervisades.
5. Elaboració de gràfiques per informes de rendiment individual o col·lectiu.
6. Disponibilitat de dades per creuar-les amb **altres datasets** i conèixer quan es juga millor o com condicionen o incideixen altres aspectes com la climatologia, malalties, aspectes mentals o psicològics en la pràctica del futbol.
7. Utilitzar les fotografies per l'entrenament de models de xarxes neuronals artificials i disposar d'un reconeixedor de cares dels jugadors de futbol, per la classificació automàtica d'informació o la generació d'altres projectes d'interactivitat en les xarxes socials, App's per telèfons mòbils, on els aficionats puguin gaudir dels ídols futbolístics o del seu propi equip de futbol.
8. Si a més a més volem anar més enllà aquest dataset en un futur podria comparar-se, establir connexions (inicialment no esperades) amb altres campionats en altres països, o continents.

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

Mirant el robots.txt:

```
User-agent: *
Disallow: /s/
Disallow: /corporativo/aviso-legal.html
Disallow: /corporativo/contacto.html
Disallow: /corporativo/ayuda.html
Disallow: /multimedia/en-tu-movil/listado/index.html
Disallow: /social/
Disallow: /edicion/
Disallow: /eltiempo/
Disallow: /deporte/futbol/primera-division/2010-2011/panel-de-fichajes/*
Disallow: /eventos/marcador/futbol/2013_14/*
Disallow: /eventos/marcador/futbol/2012_13/*
Disallow: /eventos/marcador/futbol/2011_12/*
Disallow: /encuentros/roberto-palomar/2016/03/29/*
Disallow: /2012/11/03/en/football/spanish_football/1351965522.html
Disallow: /2012/11/03/futbol/1adivision/1351945508.html
```

No se'n deriven restriccions per a bots de les pàgines que emprem.

Però de l'anàlisi de l'avís legal: https://www.marca.com/corporativo/aviso-legal.html
Queda clar que no es pot fer-ne ús de cap casta de les dades que conté, per lo que la publicació de les dades haurà de ser el màxim de restrictiva.

De les proposades:

Llicència	Característiques
Released Under CC0: Public Domain License	Llicència sense drets reservats, es pot fer el que es vulgui amb el codi i les dades.
Released Under CC BY-NC-SA 4.0 License	Aquesta reserva el dret de: **Reconeixement** — Heu de reconèixer l'autoria de manera apropiada, proporcionar un enllaç a la llicència i indicar si heu fet algun canvi. Podeu fer-ho de qualsevol manera raonable, però no d'una manera que suggereixi que el llicenciador us dóna suport o patrocina l'ús que en feu. **NoComercial** — No podeu utilitzar el material per a finalitats comercials. **Compartir Igual** — Si remescleu, transformeu o creeu a partir del material, heu de difondre les vostres creacions amb la mateixa llicència que l'obra original.
Released Under CC BY-SA 4.0 License	**Reconeixement** — Heu de reconèixer l'autoria de manera apropiada, proporcionar un enllaç a la llicència i indicar si heu fet algun canvi. Podeu fer-ho de qualsevol manera raonable, però no d'una manera que suggereixi que el llicenciador us dóna suport o patrocina l'ús que en feu. **CompartirIgual** — Si remescleu, transformeu o creeu a partir del material, heu de difondre les vostres creacions amb la mateixa llicència que l'obra original. **No hi ha cap restricció addicional** — No podeu aplicar termes legals ni mesures tecnològiques que restringeixin legalment a altres de fer qualsevol cosa que la llicència permet.
Database released under Open Database License, individual contents under Database Contents License	**ODbL** és una llicència de copyleft tipus (compartir igual) que permet a altres el lliure ús, compartició i modificació mentre mantinguin aquesta llicència a la obra derivada.

Other (specified above)	Nosaltres triem una altra la CC BY-NC-ND 4.0 com expliquem a continuació.
Unknown License	Sense cap llicència coneguda, el que és un risc ja que no hi ha el parany d'una llicència amplament reconeguda, o simplement es lliura sense cap llicència.

Repassades les llicències proposades, nosaltres optem per triar una llicència CC que no es proposa a l'enunciat per ser **encara mes restrictiva**:

CC BY-NC-ND 4.0 License: https://creativecommons.org/licenses/by-nc-nd/4.0/deed.ca

Perquè requereix de:

- **Reconeixement** — Cal reconèixer l'autoria de manera apropiada, proporcionar un enllaç a la llicència i indicar si s'ha fet algun canvi.
- **No Comercial** — No permet utilitzar el material per a finalitats comercials i evitem fer de facilitadors de materials de Marca a terceres empreses.
- **Sense Obra Derivada** — No permet difondre el material modificat i ens assegurem que s'empra com és.

9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

El codi de raspat s'ha generat en llenguatge **Python**, i està contingut en els fitxers:

1. Webscraping1.py
2. Webscraping2.py
3. Webscraping3.py

utilitzant les llibreries següents:

- '***Requests***' (gestió de peticions al web),
- '***BeautifulSoup***' (anàlisi de documents HTML i navegació per l'arbre d'elements del document per el raspat d'informació),
- '***lxml*** (per poder fer servir el parser lxml a BeautifulSoup)
- '***numpy***' (biblioteca de funciones matemàtiques d'alt nivell per operar amb aquests vectors i matrius),
- '***pandas***' (bàsicament per la gestió de '***DataFrames***'),
- '***json***' (gestió del format '***JSON***' per l'obtenció de les dades),
- '***Pickle***' (compressió binaria d'informació en fitxers),
- '**skimage'** (lectura i visualització d'imatges),
- '***matplotlib.pyplot***' (biblioteca de funcions per el control de la visualització gràfica '***matplotlib***' per Python).

Per que sigui més entenedor s'han inserit comentaris en quasi bé la totalitat de les línies i s'han inclòs descripcions dels procediments implementats en els scripts presentats. En el fitxer '***readme.md***' de la '***wiki***' i directori '***master***' de ***GitHub*** es disposa també d'explicacions concises del contingut i organització de fitxers que conformen la pràctica de webscraping.

Webscraping1.py implementa un raspat senzill sobre marques HTML. ***Webscraping2.py*** ha implicat un raspat més avançat on s'ha resseguit el codi font '***Javascript***' per trobar la '***URL***' que realitza la consulta de dades estadístiques i permet raspar les dades visualitzades en taules. **Webscraping3.py** tracta a la vegada tres pàgines del diari (Trofeu Pichichi, Trofeu Zamora i Trofeu Miguel Muñoz) el repte en aquest cas es adaptar-se amb una sola funció a recórrer taules de diferent nombre de columnes i elaborar en conseqüència tres csv diferents.
En els tres fitxers s'han implementat en el codi, solucions per superar en cas de produir-se bloquejos del servidor. En concret la utilització de '***proxies***' per variar la IP, i la modificació del '***User-Agent***' reconstruint capçaleres diferents per tal de variar el format de la peticions web. Si l'execució dels scripts no presenten problemes de bloquejos, no es tenen en compte aquestes mesures i només s'aplica un criteri de 'fair-play' o no saturació del servidor aplicant un 'timeout' de 1 segon entre peticions.

A banda del codi de raspat, s'han elaborat dos fitxers auxiliars:

- **Llegir_CSV.py** fitxer bàsic per a obrir els csv amb ***pandas*** i poder revisar els fitxers, aporta dues funcions per fer-hi una ullada ràpida als CSV.

- **Visualitza_Juagadors.py** fitxer amb dues funcions per a obtenir les imatges dels jugadors

    - d'una amb una:

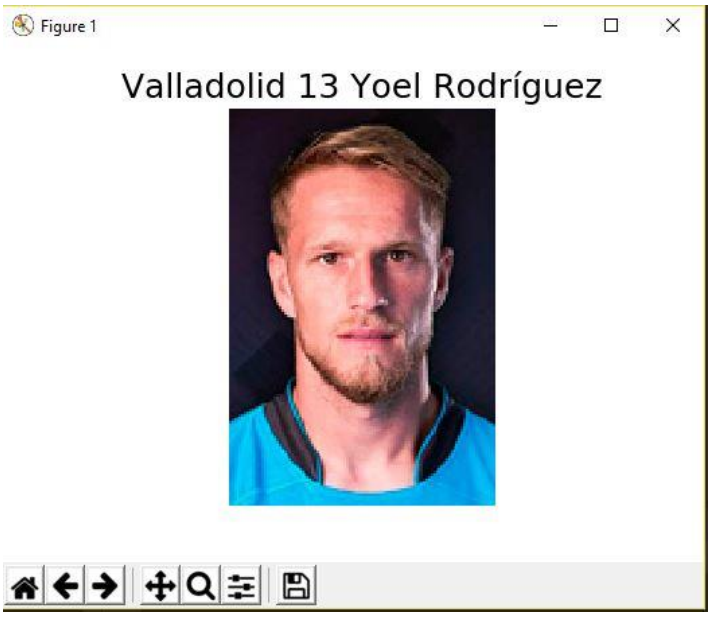


Figura 5: Exemple de visualització de les imatges de jugadors d'un en un.

- O en format d'anuari (Facebook) com es mostra a la imatge següent:

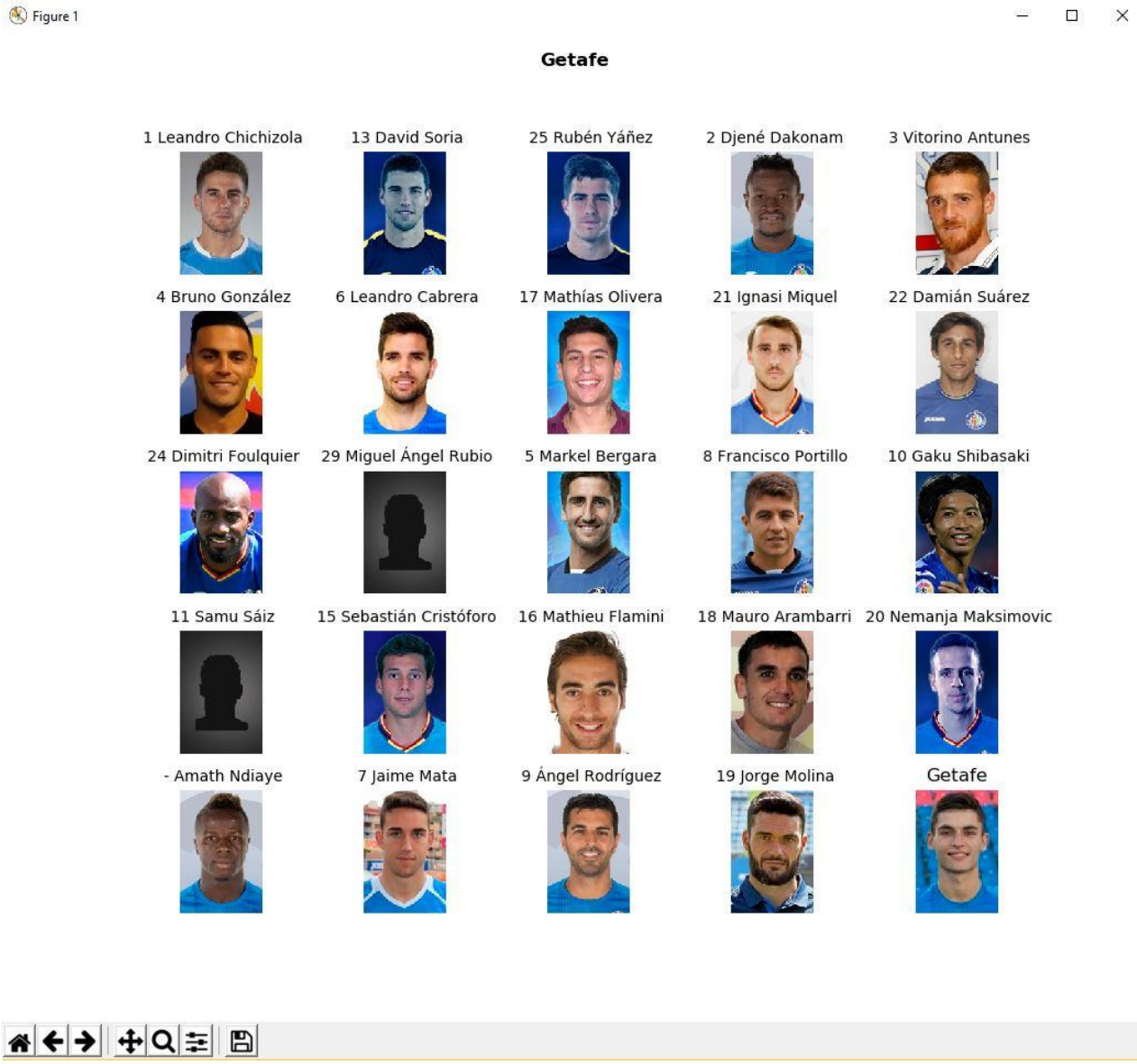


Figura 6: Exemple de visualització de les imatges d'una plantilla amb format anuari.

10. Dataset. Presentar el dataset en format CSV.

El dataset és un dataset composat per els fitxers:

jugadors.csv – format text CSV

estadístiques_jugadors.csv – format text CSV

pichichi_stats.csv – format text CSV

zamora_stats.csv – format text CSV

miguel-munoz_stats.csv – format text CSV
Fotos.pickle – format binari.

El format CSV utilitza la coma '**,**' com a separador de camps i com a caràcter decimal en valors numèrics. Per tant aquests últims queden delimitats per el caràcter '**"**'. En sistemes operatius on el punt decimal és un punt, i generi problemes, en el codi es pot variar la instrucció actual:

(fitxer '***estadístiques_jugadors.csv***')

```
td_jug.to_csv("estadístiques_jugadors.csv", decimal="," , index=False)
```

per:

```
td_jug.to_csv("estadístiques_jugadors.csv", decimal="." , index=False)
```

En el fitxer '***jugadors.csv***' no tenim que preocupar-nos ja que no es treballa amb números decimals. La codificació o 'encoding' usat és '***UTF-8'***. Per tant si utilitzem programari com pot ser 'Excel' que per defecte utilitza '1252-Europeu occidental(Windows)' o similar, caldrà especificar '***UTF-8***' en el procés d'importació de les dades.

Tant el codi usat com el dataset queden adjunts en la localització de GitHub:
https://github.com/jjdiezm/Practica1

Contribucions	Signatura
Recerca Prèvia	QDJ, JDM
Redacció de les respostes	QDJ, JDM
Desenvolupament del codi	QDJ, JDM