Índex de Continguts

[1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació. 2](#_Toc5038420)

[2. Definir un títol pel dataset. Triar un títol que sigui descriptiu. 2](#_Toc5038421)

[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). 2](#_Toc5038422)

[4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment 2](#_Toc5038423)

[5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit. 3](#_Toc5038424)

[6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha). 4](#_Toc5038425)

[7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. 4](#_Toc5038426)

[8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció: 5](#_Toc5038427)

[9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R. 5](#_Toc5038428)

[10. Dataset. Presentar el dataset en format CSV. 5](#_Toc5038429)

**Pràctica 1: Web scraping**

# 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**.**

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

Títol: **“Dades i estadístiques de plantilles de jugadors de futbol de 1ª divisió”**. -> **FutStd\_2018-2019.**

# 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 de la majoria de jugadors. Les fotografies per optimització d’espais 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 enfosquida.

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

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



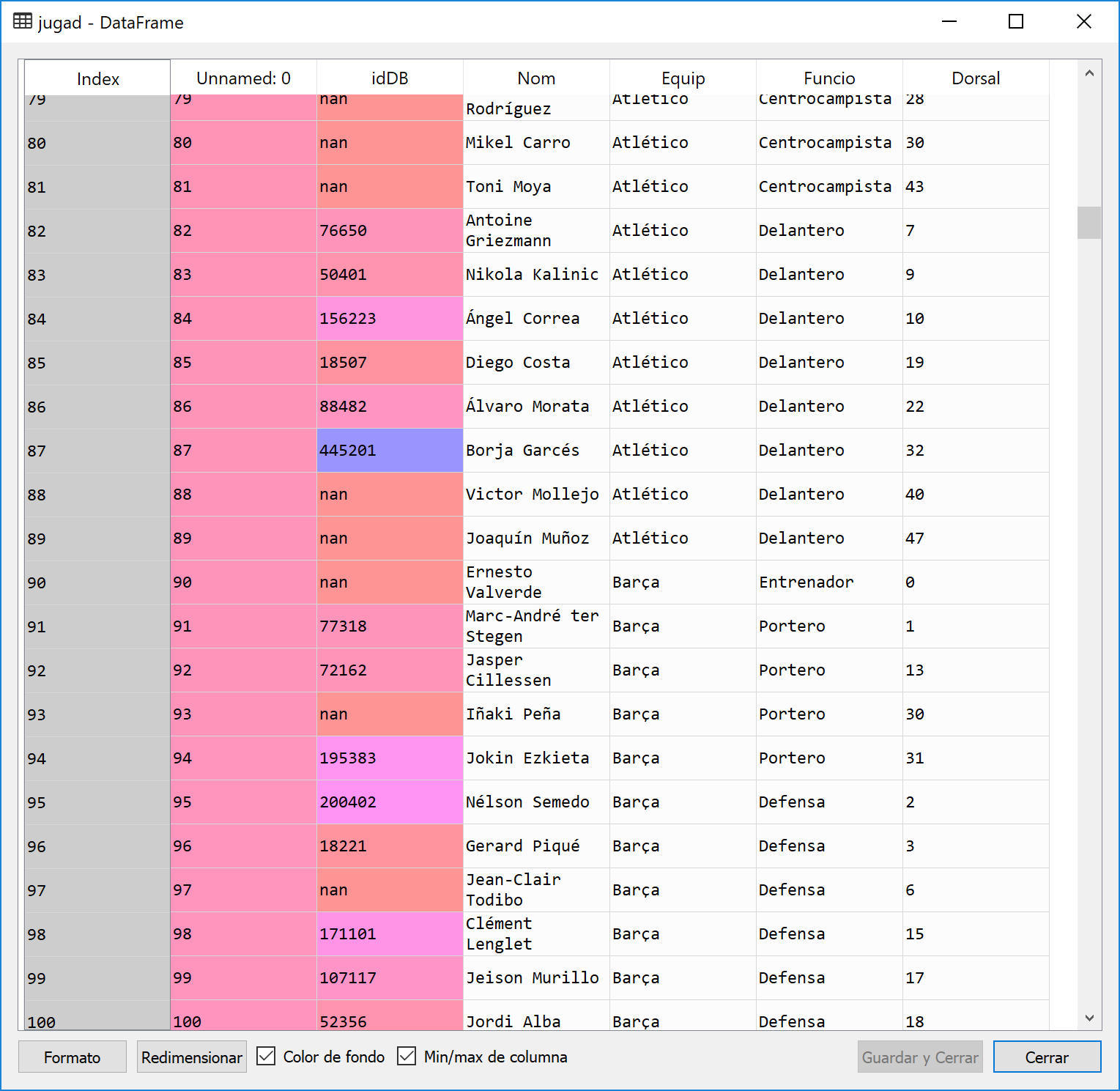
**Figura** **1**: Logotip del dataset

# 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 compensen 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.

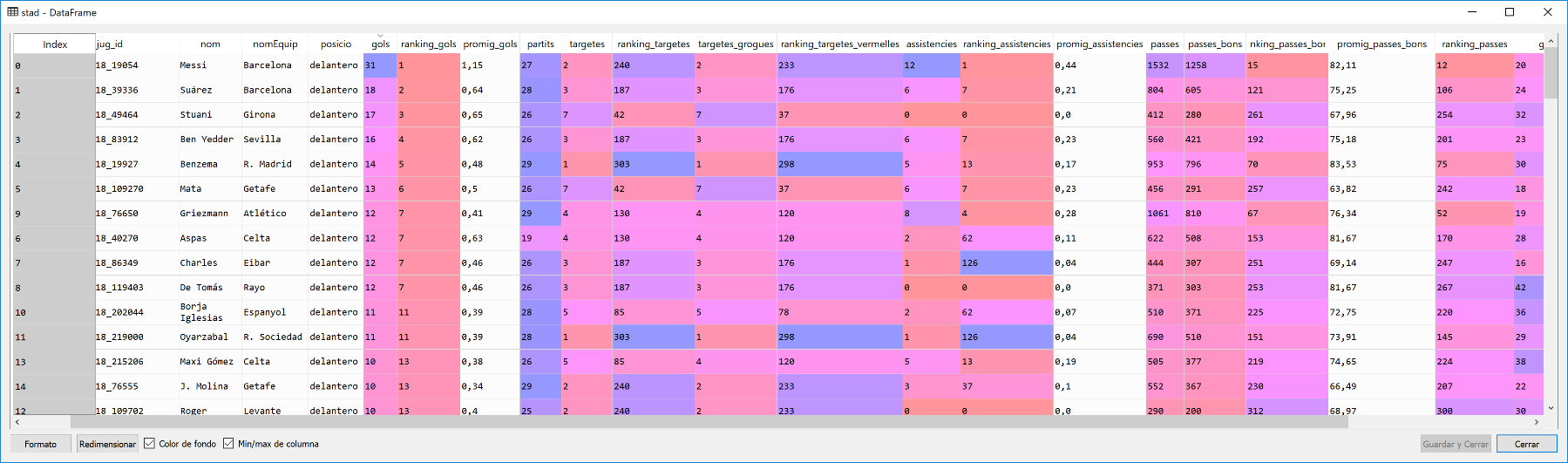
1. Dades alfanumèriques:

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.



**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:



**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

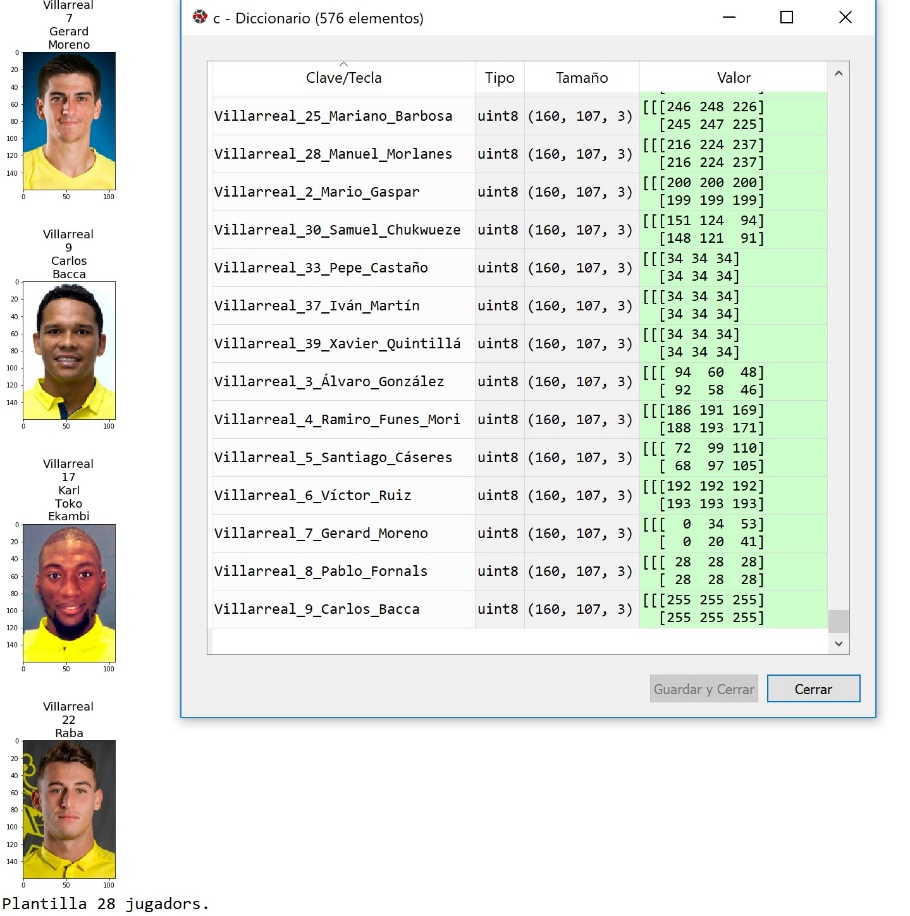
'ranking\_gols\_ok' : posició en el ranking de gols ok. Ex: 170

'promig\_gols\_ok' : promig de gols ok. Ex: 0.74

Els dos datasets poden unirse 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’. També caldrà tenir en compte que les estadístiques son dels 100 millors jugadors no dels jugadors de tota la lliga de 1ª divisió de futbol que son en total 576.

1. 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 pixels.

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.

# 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. 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.

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

Les preguntes a respondre en posteriors projectes d’anàlisi de dades son per ordre d’importància:

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

Gols, aturades, passes correctes, més minuts jugats 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’acumulat d’estadístiques).

1. Construir models i fer prediccions sobre el que pot arribar a fer un equip o jugador, en qualsevol moment en el temps d’una temporada.
2. Amb models adients 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.
3. Agrupar els jugadors per similitud de característiques.
4. Elaboració de gràfiques per informes de rendiment individual o col·lectiu.
5. 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.
6. Utilitzar les fotografies per l’entrenament de reconeixement de cares.

# Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar 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.

# 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, utilitzant les llibreries següents:

Requests, BeautifulSoup, numpy, pandes, json, Pickle, skimage, matplotlib.pyplot.

# Dataset. Presentar el dataset en format CSV.

Tant el codi usat com el dataset queden adjunts en la localització de GitHub facilitada.

<https://github.com/QuimDJ/Practica1/wiki>