PRÀCTICA 1. Tipologia i Cicle de vida de les dades

Manel Benavides Palos (manelbenavides)
Fèlix Ribera Forés (friberaf)
Màster Universitari en Ciència de Dades

Pregunta 1

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

El context en el qual s'engloba la informació recolectada en aquesta pràctica és el camp de l'anàlisi de dades esportives. En concret, el context esportiu que s'estudia es focalitza en el món del futbol, de la 1a divisió espanyola (Liga Santander actualment). En la pregunta 7 s'exposa en més detall l'origen d'aquesta disciplina que pot arribar a ser reconeguda en si mateixa.

Quan es parla d'esport i informació, el primer que ve al cap és els diaris esportius de referència (As, Marca, Sport i Mundo Deportivo serien una bona representació, tot i que existeixen molts més amb matisos específics per atraure diferents segments de població).

Per aquest motiu, la informació s'ha extret dels llocs web de referència en el context que ens trobem, que han estat les webs dels dos diaris mencionats anteriorment amb seu a Madrid, com són As i Marca (de fet, la importància de Marca és tal, que els premis que otorga han arribat a prendre un renom especial, com són el Trofeo Pichichi al jugador que ha anotat més gols en una temporada i el Premio Zamora al porter que menys gols ha encaixat en una temporada).

Pregunta 2

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

A partir d'aquesta pràctica s'obtenen dos datasets diferents:

- Històric de classificacions lligueres
- Històric de partits lliguers

Pregunta 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).

Històric de classificacions lligueres: Conté la classificació final de cada temporada des que existeixen dades a la font de dades que s'ha fet servir (2001-2002) fins al moment actual. Aquest dataset conté per cada temporada el conjunt d'equips amb la seva corresponent posició

i estadístiques rellevants com els punts aconseguits i gols (tant el total, com el conjunt d'estadístiques aconseguides com a local i com a visitant).

Històric de partits lliguers: Conté el conjunt de tots els partits que s'han disputat des de la mateixa temporada que el dataset de classificacions (2001-2002) fins la temporada anterior a la actual.

Tal i com està fet el codi de la pràctica, es troba preparat per actualitzar-se automàticament i que cada nova temporada es puguin anar recollint les noves dades de les noves temporades.

Pregunta 4

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

El Barça rebent un aplaudiment després de proclamar-se campió matemàticament.

Pregunta 5

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

Dataset Històric de partits lliguers:

season: string
Indica la temporada a que pertanyen les dades. Format 2001_2002.

match_week: int
Jornada en la que es juga el partit.

home: string
Nom de l'equip local.

visitor: string
Nom de l'equip visitant

result: string
Resultat del partit. Format: *home_score-visitor_score*

Dataset Històric de classificacions lligueres:

season: string
Indica la temporada a que pertanyen les dades. Format 2001_2002.

name: string
Nom de l'equip.

position: int
Posició en la classificació de l'equip.

points: int
Punts totals de l'equip.

played: int
Número total de partits jugats.

won: int
Número total de partits guanyats.

drawn: int
Número total de partits empatats.

lost: int
Número total de partits perduts.

gf: int
Gols a favor totals.

ga: int
Gols en contra totals.

gd: int
Diferència entre gols a favor i gols en contra.

h_points: int
Punts totals de l'equip guanyats com a local.

h_played: int
Número de partits jugats com a local.

h_won: int
Número de partits guanyats com a local.

h_drawn: int
Número de partits empatats com a local.

h_lost: int
Número de partits perduts com a local.

h_gf: int
Gols a favor com a local.

h_ga: int
Gols en contra com a local.

h_gd: int
Diferència entre gols a favor i gols en contra com a local.

a_points: int
Punts de l'equip guanyats com a visitant.

a_played: int
Número de partits jugats com a visitant.

a_won: int
Número de partits guanyats com a visitant.

a_drawn: int
Número de partits empatats com a visitant.

a_lost: int
Número de partits perduts com a visitant.

a_gf: int
Gols a favor com a visitant.

a_ga: int
Gols en contra com a visitant.

a_gd: int
Diferència entre gols a favor i gols en contra com a visitant.

El període de temps de les dades és des de l'any 2001, temporada 2001-2002, fins a l'actualitat, temporada 2020-2021. Al no haver acabat encara la temporada actual, les dades poden variar segons el moment de l'execució de l'scraping.

L'extracció de dades s'ha dividit en dos blocs.

En el primer bloc s'ha extret les dades històriques de les classificacions dels equips així com puntuacions i estadístiques considerades rellevants.
Al voler-se recuperar informació de múltiples anys, s'ha trobat que l'estructura de la pàgina web variava cap a versions més antigues i diferents de l'actual, pel que l'algoritme s'ha anat adaptant segons el format de la informació de cada any.

Pel que fa el segon bloc, s'ha extret les dades de cada una de les jornades d'una temporada, quins equips han jugat, quin ha estat el resultat, al camp de qui es jugava, etc.
Un cop més s'ha trobat diferències en l'estructura de la pàgina en la que s'ha extret les dades segons s'avançava cap a les dades més antigues.

Malauradament no s'ha trobat cap API que retornés aquesta informació, pel que tota l'extracció s'ha fet mitjançant web scraping.

Pregunta 6

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

Al ser el futbol un esport molt seguit es van trobar varies pàgines amb informació sobre la classificació dels equips actual, com per exemple la web del diari *Sport*[1] o la pròpia pàgina de *LaLiga*[2], però que només contenien la informació de la temporada actual o de molts pocs anys enrere.

Es va acabar trobant la web de l'*as*[3] de la qual es va extreure l'històric de classificacions.
L'as és un diari esportiu centrat principalment amb el futbol, és per això que es va decidir fer la recerca aqui. D'aquest diari es va extreure l'històric de classificacions.

Per altre banda, del diari *Marca*[4] es va extreure les dades de cada jornada dels últims 20 anys.
Marca igual que l'as és un diari esportiu centrat en el futbol.

Pregunta 7

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

Durant les últimes dècades, hi ha hagut un *boom* en l'anàlisi de successos empírics observats.

L'esport en general i el futbol en particular, no han estat aliens a aquest *boom* i ho han fet usant un creixent nombre de gadgets que permeten disposar de moltes dades que ajuden a monitoritzar tot el què succeeix en cada instant de joc, en cada jugador i en cada posició de l'entorn del terreny de joc. Actualment en un equip professional, es monitoritza cada instant d'entrenament de cadascun dels jugadors per a optimitzar el seu rendiment i estat físic a l'hora de competir. A més l'anàlisi també es pot aplicar sobre estadístiques grupals d'equip, amb eines per a predir quants gols hauria d'haver marcat un equip donades les ocasions que han tingut durant un partit.

Tot això ha propiciat que l'estadística i l'anàlisi de dades orientat a l'esport permetin la creació d'una disciplina tan concreta com interessant i que crea un gran nombre de llocs de treball.

En aquesta pràctica, s'ha volgut donar un caire d'aquesta disciplina, tot i que a un nivell molt inferior a la monitorització a la que arriben els equips professionals (entre altres motius, no es disposen de les dades necessàries per a recolectar, ja que són dades molt privades i valuoses).

[1] https://www.sport.es/es/resultados/la-liga/clasificacion/
[2] https://www.laliga.com/laliga-santander/clasificacion
[3] https://resultados.as.com/resultados/futbol/primera/2020_2021/clasificacion/
[4] https://www.marca.com/estadisticas/futbol/primera/

S'ha intentat agafar un conjunt de dades d'accessibilitat relativament senzilla, on el problema fós el scraping de les dades, doncs aquest era l'objectiu de la pràctica.

Exposats tots aquests motius, s'ha agafat el resultat de tots els partits disputats i de l'històric de classificacions de la lliga, ja que són les dades més importants i visibles sobre les quals es podria mesurar l'èxit o fracàs d'un projecte esportiu.

Mitjançant aquestes dades, es podria aconseguir correlacionar dades sobre com aconseguir fer una temporada exitosa. Hi ha diverses hipòtesis:

- Cal ser un equip regular en el joc i sempre aconseguir resultats ajustats (tant guanyant com perdent), per tant segurament s'hauria d'apostar per un plantejament defensiu? Potser els equips més exitosos, ofereixen un joc més vistós (i per tant més obert) on, quan es perd un partit es perd per molts gols?
- Els equips exitosos es mouen per ratxes de victòries consecutives? O més aviat per ratxes de partits sense perdre (per tant, encadenant empats i victòries)?
- Cal fer molts gols per guanyar? O saber optimitzar els gols (i que el ratio de gols/punt a la classificació sigui baix)? S'ha de tenir en compte que un golejador és la posició més cara al futbol, i que cal optimitzar costos.
- Un equip fort a casa pot ser exitós? O són els equips que marquen la diferència fora els que realment acaben en bones posicions a la classificació?
- Com d'important és començar la lliga amb una bona ratxa? I acabar-la encadenant bons resultats? Orientat a aquesta pregunta, es podria optimitzar la càrrega de treball als jugadors per arribar en les millors condicions al tram decisiu.

Totes aquestes preguntes i moltes més poden marcar les bases d'un projecte esportiu, doncs es pot valorar si en el futbol modern (dels últims 20 anys) els equips que triomfen són ofensius o defensius, marquen molts gols o optimitzen el ratio de gol/partit, en quin tram de la temporada tenen una ratxa de bons resultats, en quin tipus de partit t'has d'esforçar encara més que en la resta, etc. I en base a les respostes, signar un entrenador amb un cert estil de joc, jugadors que encaixin en els perfils idonis per a l'equip i un cos tècnic preparat per a executar les ordres necessàries per aconseguir un èxit esportiu.

Pregunta 8

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**

Així com en el diari *as* permet el tractament de dades sempre i quan no se'n faci un ús comercial[5], el diari marca prohibeix entre d'altres extreure o copiar part o la totalitat del seu contingut[6].

És per això que s'ha decidit que el dataset Històric de classificacions lligueres, provinent de l'as tingui una llicència CC BY-NC-SA 4.0 ja que d'aquesta manera s'estipula un ús no comercial de les dades ni dels seus derivats.

Pel que fa al dataset Històric de partits lliguers, al provenir de marca cal demanar permís per distribuir les dades, pel que pel moment està protegit i no es pot utilitzar.

Pregunta 9

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

El codi es pot trobar al següent enllaç de github:

https://github.com/friberaf/web-scrapping

[5] Punt 4: https://as.com/diarioas/aviso_legal.html
[6] Apartat 3.4: https://www.marca.com/corporativo/aviso-legal.html

Pregunta 10

Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

Aquests datasets contenen un històric de les classificacions de la lliga de primera divisió espanyola desde la temporada 2001-2002 fins a la temporada actual i un històric de tots els partits disputats a la primera divisió espanyola desde la temporada 2001-2002 fins la temporada anterior a la actual.

DOI: 10.5281/zenodo.4263439

TAULA DE CONTRIBUCIONS

Contribucions	Signa
Recerca prèvia	MBP, FRF
Redacció de les respostes	MBP, FRF
Desenvolupament del codi	MBP, FRF