Tipologia i cicle de vida de les dades

PRA1 – Com podem capturar les dades de la web?

Noms: Carlos Martínez Torró (cmtorro@uoc.edu), Xavier Roca Canals (xrocaca@uoc.edu)

Exercici 1. Context

En un món en el que les dades cada cop tenen un paper més protagonista, hi ha pocs àmbits que escapin de la tendència a utilitzar-les per aconseguir posicions avantatjoses. Un dels sectors que es va incorporant poc a poc al món de la ciència de dades és el de l'esport, on ja és freqüent trobar professionals de les dades als clubs més importants.

Això és especialment preponderant als Estats Units, on el culte a les dades i l'estadística ha arribat fins i tot a la gran pantalla amb Moneyball (2011), amb la història d'un executiu del beisbol, Billy Beane, que es basava en les estadístiques per a fer fitxatges. Beane, que per aquella època estadísticament no era més que un *outlier*, va ser un pioner, doncs cada cop és més habitual que les dades condicionin les decisions esportives. I no només pel que fa a fitxatges.

El joc evoluciona i ho fa de la mà de les dades, en busca d'una major eficiència, d'un avantatge competitiu. El món del bàsquet és un dels grans exemples, tot i que no l'únic. A ningú escapa que no es jugava igual als 90 o, fins i tot, als inicis dels 2000 que actualment. Què ha canviat? I, més important, *per què* ha canviat? Estan les dades relacionades amb aquesta evolució de l'esport? Podem descobrir amb aquestes dades els patrons que fan que uns equips guanyin més que d'altres?

Per intentar respondre aquestes preguntes, tenim la gran sort de comptar amb la web de Basketball Reference (https://www.basketball-reference.com/), una web de referència on es poden consultar dades de totes les temporades de la NBA i la WNBA des d'abans de 1950. Podem anar des de les estadístiques més bàsiques (tirs de camp, punts, percentatges d'encert...) fins mètriques molt més complexes (com l'eficiència dels llançaments o la duresa del calendari).

Exercici 2. Títol

Evolució estadística de la NBA (1950-2022) i la WNBA (1997-2022).

Exercici 3. Descripció del dataset

Es tracta d'un dataset de 1900 files i 53 variables que recull un bon conjunt d'estadístiques dels equips de la NBA i la WNBA, des de 1950 i 1997, quan aquestes dues lligues van començar, respectivament. Cada fila és la descripció estadística de la temporada d'un equip a la temporada regular, des de les seves victòries, als punts anotats i encaixats per partit, el percentatge de tirs de camp (o triples, un cop van aparèixer el 1979), a altres mètriques més complexes com el ritme cada 100 possessions, ràtings ofensius o defensius, eficiència dels llançaments... Per tant, és un dataset que ens dona una imatge amb diferents capes de profunditat de cada equip en un any concret, el que ens permetrà, a priori, estudiar com ha canviat l'esport amb els anys i què és el que distingeix, per exemple, als millors equips cada any.

Exercici 4. Representació gràfica

***Figura 1**. Representació gràfica de l'evolució estadística de la NBA i WNBA. Mostra els logos dels equips guanyadors dels anys definits.*

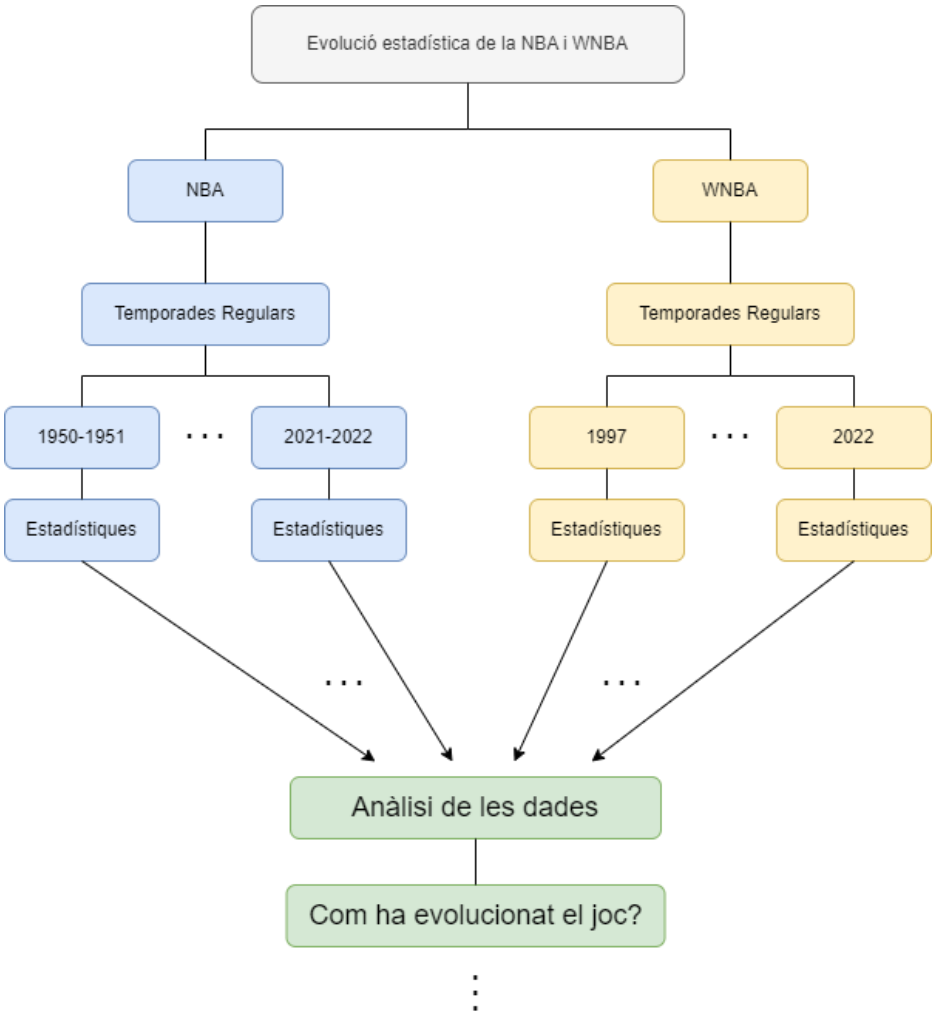

Figura 2. Diagrama del projecte de web scraping. Estructuració de les dades a recollir.

Exercici 5. Contingut

Per cada fila del conjunt de dades creat, és a dir, estadístiques d'un equip durant una temporada en concreta, es recullen els següents camps:

- **Season:** Any de la temporada regular.
- **League:** Lliga on juga l'equip.
- **Team:** Equip de bàsquet.
- **Wins:** Victòries aconseguides durant la temporada regular.
- **Losses:** Derrotes aconseguides durant la temporada regular.
- **Win_loss_pct:** Percentatge de la proporció entre les victòries i derrotes aconseguides durant la temporada regular.

- **Gb:** *Games Behind,* indica quants partits té cada equip per darrere del rival que ocupa el primer lloc segons les divisions.
- **Pts_per_g:** Mitjana de punts anotats per partit.
- **Opp_pts_per_g:** Mitjana de punts que encaixa per partit aconseguits pels rivals.
- **Playoffs:** Indicador sobre els equips que han aconseguit entrar a *playoffs*.
- **Age:** Mitjana d'edat de tots els jugadors que formen l'equip.
- **Wins_pyth:** Victòries pitagòriques, és a dir, victòries esperades en funció dels punts anotats i encaixats.
- **Losses_pyth:** Derrotes pitagòriques, és a dir, derrotes esperades en funció dels punts anotats i encaixats.
- **Mov:** Marge de victòria.
- **Sos:** Robustesa del calendari, aquesta classificació es defineix segons els punts per damunt/per sota de la mitja.
- **Srs:** Sistema de classificació simple, aquesta classificació té en compte la diferència entre la mitja de punts i la robustesa del calendari.
- **Off_rtg:** Classificació ofensiva, es basa en una estimació dels punts produïts pels jugadors o anotats pels equips per cada 100 possessions.
- **Def_rtg:** Classificació defensiva, es basa en una estimació dels punts encaixats per cada 100 possessions.
- **Net_rtg:** Classificació de cistelles, es basa en la diferencia de punts per cada 100 possessions.
- **Pace:** Estimació de les possessions per 48 minuts.
- **Fta_per_fga_pct:** Nombre d'intents de tir lliure per intent de tir de camp.
- **Fg3a_per_fga_pct:** Percentatge de tirs intentats des de la línia de tres punts segons els tirs intentats.
- **Ts_pct:** Percentatge de tir real, es tracta d'una mesura de la eficàcia dels tirs que té en compte tots els tirs.
- **Efg_pct:** Percentatge efectiu de tirs de camp, ajusta el fet de que un tir de camp de tres punts val un punt més que un tir de dos punts.
- **Tov_pct:** Percentatge de pèrdues de pilota, es basa en una estimació de les pèrdues de pilota per cada 100 jugades.
- **Orb_pct:** Percentatge de rebots ofensius, es basa en una estimació dels rebots ofensius aconseguits per un jugador durant el temps que estava en la pista.
- **Ft_rate:** Tirs lliures per intent de tir.
- **Opp_efg_pct:** Percentatge efectiu de tirs de camp de l'oponent.

- **Opp_tov_pct:** Percentatge de pèrdues de pilota del rival.
- **Drb_pct:** Percentatge de rebots defensius, es basa en una estimació dels rebots defensius aconseguits per un jugador durant el temps que estava en la pista.
- **Opp_ft_rate:** Tirs lliures de l'oponent per intent de tir.
- **G:** Nombre de partits jugats.
- **Mp:** Minuts jugats.
- **Fg:** Mitja de tirs de camp.
- **Fga:** Mitja d'intents de tirs de camp.
- **Fg_pct:** Percentatge de tirs de camp segons els intents i els aconseguits.
- **Fg3:** Mitja de tirs de tres punts.
- **Fg3a:** Mitja d'intents de tirs de tres punts.
- **Fg3_pct:** Percentatge de tirs de tres punts segons els intents i els aconseguits.
- **Fg2:** Mitja de tirs de dos punts.
- **Fg2a:** Mitja d'intents de tirs de dos punts.
- **Fg2_pct:** Percentatge de tirs de dos punts segons els intents i els aconseguits.
- **Ft:** Mitja de tirs lliures.
- **Fta:** Mitja de tirs lliures intentats.
- **Ft_pct:** Percentatge de tirs lliures segons els intents i els aconseguits.
- **Orb:** Mitja de rebots ofensius.
- **Drb:** Mitja de rebots defensius.
- **Trb:** Mitja de rebots totals.
- **Ast:** Mitja d'assistències.
- **Stl:** Mitja de pilotes robades.
- **Blk:** Mitja de bloquejos.
- **Tov:** Mitja de pilotes perdudes.
- **Pf:** Mitja de faltes personals.

Cada una d'aquestes dades es van afegint i actualitzant durant el transcurs de les temporades. Com el que es vol és recollir les dades de les temporades regulars finalitzades, aquestes no presentaran actualitzacions. Per tant, l'última actualització que s'haurà realitzat és al final de temporada o en el moment en què es va crear la pàgina.

Exercici 6. Propietari

El propietari de les dades és Sports Reference, una web on es recullen estadístiques de diferents esports nord-americans. Amb l'objectiu de respectar les seves normes pel que fa a l'ús de dades recollides de la seva web, vam estar revisant la seva política respecte a les dades, que es pot trobar en aquest enllaç. En l'enllaç, se'ns emplaça a no utilitzar mètodes automàtics (com podrien ser *scrapers* com el nostre) d'una forma que pugui afectar tant el rendiment de la pàgina com l'accés d'altres usuaris a aquesta. Per evitar-ho i, també seguint les recomanacions respecte al tràfic amb *scrapers*, hem implementat uns *timers* al nostre codi que s'ajusten sobradament al que es demana (menys de 20 *requests* per minut). També hem implementat uns *headers* que tracten d'imitar el tràfic humà a la web, tot i que, com s'indica en l'enllaç anterior, el que limita l'accés és el nombre de peticions per minut.

A més, també se'ns prohibeix generar noves pàgines webs o eines basades en les dades que acabem de crear. Aquesta no és la nostra intenció, ja que la nostra finalitat és purament analítica i no pretenem competir amb ells ni generar benefici a partir d'aquest codi o l'anàlisi que en sorgeixi a posteriori. Per altra banda, el fitxer robots.txt del lloc web ens indica que no es no permet l'accés d'*scrapers* a certes direccions de la pàgina, però les URL que nosaltres visitem amb l'*scraper* no estan incloses dins d'aquestes.

Finalment, pel que respecta a anàlisis anteriors, podem trobar alguns llocs web que han utilitzat dades de Sports Reference per anàlisi de dades. Aquí[1], per exemple, podem veure un anàlisi on s'utilitza R per analitzar les estadístiques d'un jugador al llarg de la seva carrera. Es pot veure com evolucionen algunes de les seves mètriques durant aquests anys. En aquesta altra web[2], l'autor *scrapeja* dades de Basketball Reference també amb R per analitzar partits dels Lakers de Los Angeles. Per altra banda, aquí[3] se'ns parla d'algunes de les noves mètriques que han sorgit en els darrers anys (i que podem trobar a Basketball Reference) i de la seva importància a l'hora d'entendre el joc avui en dia. Finalment, també podem trobar dades sobre les tendències dels llançaments[4] en els últims darrers anys i la rellevància d'aquestes tendències sobre l'èxit de l'equip.

Exercici 7. Inspiració

Creiem que els anàlisis que citem a l'apartat 6 no entren en gaire profunditat en l'evolució del joc o es limiten a un període en concret. La nostra idea és anar una mica més enllà. Sabem que no es juga al bàsquet tal com es jugava als anys 50, però volem saber per què. Què ha canviat? I per què ha canviat? És evident que canvis com la introducció del llançament de tres punts a l'any 1979 van canviar el joc per sempre, però des dels anys 80 fins aquí hi ha hagut una evolució clarísima de l'esport. No cal anar als 80: un partit de principis dels 2000 té poc a veure amb el que podríem trobar-nos ara a la televisió.

Per això, no ens fixarem en un jugador ni en un equip en concret ni tampoc ens limitarem a una temporada determinada o a al rang d'anys de la carrera d'un jugador. Tampoc ens limitarem, a priori, a analitzar les tendències dels llançaments (hi ha moltes més estadístiques que poden influir en les victòries a més dels llançaments). Sabent que (com es diu al tercer link[3]) les noves mètriques disponibles poden donar-nos una imatge més real del que està passant a l'esport, volem fer una comparativa històrica a nivell estadístic que ens proporcioni una idea de com està evolucionant (i com ho ha fet al passat) el bàsquet nord-americà (amb el que comporta, ja que marca la tendència de les altres competicions mundials).

Amb aquest objectiu, agafarem les estadístiques de tots els equips al llarg de totes les temporades existents, tant de la NBA com de la WNBA. La idea final és ambiciosa: hi ha alguns patrons que distingeixin actualment (o hagin distingit al passat) als equips més exitosos? Quins són? I, en aquesta era on proliferen les mètriques, poden les dades estar darrere de l'evolució de l'esport?

Exercici 8. Llicència

El nostre treball serà publicat amb la llicència ***Creative Common Zero (CC0): Public Domain License***. Hem triat aquesta llicència ja que és la més permissiva en quant a copyright. La nostra intenció no és restringir l'accés ni la reproducció del nostre codi, sinó que aquest aconsegueixi la màxima distribució i que d'altres puguin fer-lo servir, ja sigui per aquesta mateixa temàtica o adaptar-lo amb altres finalitats. Creiem que és el més just, ja que nosaltres mateixos ens aprofitem a diari de la feina que altres comparteixen de forma altruista per inspirar-nos a l'hora de realitzar les nostres pròpies creacions.

Exercici 9. Codi

El codi que s'ha realitzat permet extreure les dades de la pàgina web https://www.basketball-reference.com/. Dins d'aquesta ens centrem en les dues lligues a analitzar, la WNBA i la NBA.

Per tant, el primer que fem és buscar a partir de quin enllaç hem d'iniciar el web scraping. En les dues opcions, trobem una pàgina web inicial molt similar, en la qual ens trobem una taula amb totes les temporades que s'han jugat. A partir d'aquesta, analitzem el contingut de la taula per extreure tots els enllaços de les temporades per obtenir les seves estadístiques.

Les pàgines web de les estadístiques també disposen d'una estructura molt similar. Presenten la informació separada en diferents taules. Les taules que més informació ens han aportat són les de classificacions de la conferència (*Conference Standings*), estadístiques per partit (*Per Game Stats*) i estadístiques avançades (*Advanced Stats)*. A partir d'aquestes generem el conjunt de dades final.

Un dels problemes que ens hem trobat, és que als inicis de la NBA, aquesta no es dividia en conferències. Per tant, vam haver d'afegir una nova taula per obtenir les dades que trobàvem a la taula de conferència. Aquesta s'anomena classificacions de les divisions (*Division Standings).* Finalment, es va decidir deixar d'utilitzar la taula de classificacions per conferència, ja que en les pàgines de temporades actuals també es disposava de la taula de classificacions de divisions. D'aquesta forma, no calia controlar quines taules escollir en cada cas.

Seguidament, realitzarem una breu descripció del codi. Aquest s'ha col·locat dins de la carpeta **/source**. En aquesta, podrem trobar dos fitxers: **main.py** i **NBAWNBAStatsScraper.py**. Seguidament, comentarem cada un d'ells.

- **NBAWNBAStatsScraper.py:** Es tracta d'una classe en *Python*. Aquesta conté tots els mètodes encarregats de realitzar el *web scraping* segons l'enllaç base que es passa per paràmetre. En aquesta classe trobem els següents mètodes:
    - **Get_all_links**: És l'encarregat d'extreure tots els enllaços de les temporades regulars a partir de l'enllaç inicial de cada lliga. Finalment, retorna una llista amb llistes formades per cada url de cada temporada i l'any d'aquesta.

  - **Get_standings:** És l'encarregat d'extreure les estadístiques passades per paràmetre de les taules referents a les classificacions per divisió. Aquest retorna un *dataframe* amb la informació obtinguda.
  - **Get_per_game_stats:** És l'encarregat d'extreure les estadístiques passades per paràmetre de la taula referent a les estadístiques per partit. Aquest retorna un *dataframe* amb la informació obtinguda.
  - **Get_advanced_stats:** És l'encarregat d'extreure les estadístiques passades per paràmetre de la taula referent a les estadístiques avançades. Aquest retorna un *dataframe* amb la informació obtinguda.
  - **Get_all_stats:** És l'encarregat de cridar als anteriors mètodes per tal de recol·lectar totes les estadístiques i agrupar-les en un únic *dataframe*. Aquest retorna el *dataframe* final creat.
  - **NBA_WNBA_scraper:** És el mètode principal. És l'encarregat de cridar al primer mètode per recollir tots els enllaços de les temporades. A partir d'aquests crida N (longitud de la llista de les temporades tant per NBA i WNBA) cops l'anterior mètode per extreure totes les estadístiques. Finalment, agrupa tots els **dataframes** creats durant l'execució per tal de crear el conjunt de dades final.

- **Main.py**: Es tracta del fitxer *Python* a executar. Crida a la classe NBAWNBAStatsScraper amb l'enllaç base de la pàgina web com a paràmetre. Un cop inicialitzada, crida al mètode NBA_WNBA_scraper() per tal de començar amb el procés de *web scraping*.

Cal destacar que la llibreria de *web scraping* utilitzada ha estat *Beautiful Soup*.

Finalment, s'ha inclòs l'arxiu requeriments.txt amb les versions de les llibreries *beautifulSoup, bs4, pandas* i *requests* utilitzades per executar el nostre codi.

Exercici 10. Dataset

Podem trobar el dataset en el següent enllaç:

Statistical Evolution of NBA (1950-2022) and WNBA (1997-2022) | Zenodo

El DOI assignat al nostre dataset és: 10.5281/zenodo.7343400.

Exercici 11. Vídeo

El vídeo es troba a la següent direcció de Google Drive:

https://drive.google.com/file/d/1GBhzDMfkmYDNbXnnSuQCkf6RC0VcQM-n/view?usp=sharing

Taula de contribucions

Contribucions	Signatura
Investigació prèvia	CMT, XRC
Redacció de les respostes	CMT, XRC
Desenvolupament del codi	CMT, XRC
Participació al vídeo	CMT, XRC

Bibliografia

1. Bill Kapatsoulias (19 de desembre del 2021). *NBA Analytics Tutorial: Using R to Display Player Career Stats*. R-bloggers. https://www.r-bloggers.com/2021/12/nba-analytics-tutorial-using-r-to-display-player-career-stats/
2. Marios Kokkodis (6 de desembre del 2020). *Data Analytics in Practice*. Kokkodis. http://kokkodis.com/workbook/_book/index.html.
3. Ehran Khan (18 d'octubre del 2013). *Advanced NBA Stats for Dummies: How to Understand the New Hoops Math*. Bleacher Report. https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math
4. Robert Sandberg (24 d'octubre del 2021). *Data on 3-Point Shot in the NBA's Trends and Team Success*. NYC Data Science Academy. https://nycdatascience.com/blog/python/the-3-point-shot-in-the-nba-analysis-of-trends-and-correlations-with-team-success/