**Tipología y ciclo de vida de los datos**

**Práctica 1: Webscrapping**

Icono

Descripción generada automáticamente con confianza baja

Por:

Adrián García Rodríguez

Tomás Jesús Luna López

Índice

[1. Contexto 2](#_Toc100166691)

[2. Título 3](#_Toc100166692)

[3. Descripción del dataset 4](#_Toc100166693)

[4. Representación gráfica 5](#_Toc100166694)

[5. Contenido 5](#_Toc100166695)

[6. Agradecimientos 5](#_Toc100166696)

[7. Inspiración 6](#_Toc100166697)

[8. Licencia 6](#_Toc100166698)

[9. Código 6](#_Toc100166699)

[10. Dataset 6](#_Toc100166700)

[11. Video 6](#_Toc100166701)

# Contexto

En los últimos años hemos visto gestas deportivas que podían parecer imprevisibles, cómo por ejemplo la consecución del título de la Premier Ligue por parte del Leicester C.F., o el título de la NBA de los Toronto Raptors, un equipo que no hace tanto era el peor de su división.

Además, en los últimos tiempos han proliferado las casas de apuestas, y aunque no nos posicionamos a favor de las apuestas y sabemos el gran problema que supone la ludopatía, ¿seríamos capaces de predecir los resultados mejor que las casas de apuestas y ganarles en su propio juego?

Supongamos que queremos ser capaces de ajustar un modelo para predecir resultados deportivos, y si el modelo es lo suficientemente preciso, realizar apuestas deportivas en función de las recomendaciones del sistema. Aunque para esto podríamos elegir muchas disciplinas deportivas como el baloncesto o el tenis, para nuestro caso hemos elegido el futbol, concretamente La Liga española.

Para ello, suponemos que los jugadores disponibles de cada equipo, así como su desempeño a lo largo de la temporada, pueden ayudarnos en esta tarea. Por ejemplo, para dos equipos de calidad similar, si un equipo tiene disponibles jugadores que han estado jugando muy bien y el otro equipo tiene jugadores con mala racha de partidos, el modelo debería predecir que ganará el primero.

El objetivo final (no se incluye en esta práctica) sería tener un resultado estimado para cada partido (victoria local, empate, o derrota local), con una probabilidad para cada resultado. Posteriormente compararíamos nuestras probabilidades con las cuotas ofertadas por las casas de apuestas, y en caso de que en los testeos el modelo hubiese resultado robusto, apostaríamos en las apuestas que tuvieran más esperanza matemática.

Para conseguir tal información comprobamos distintas páginas webs, entre ellas, de valoración de jugadores de ‘ligas fantásticas’, por ejemplo:

[Estadísticas LaLiga 2021/22 - Alineaciones probables - FútbolFantasy (futbolfantasy.com)](https://www.futbolfantasy.com/laliga/estadisticas)

[Estadísticas de Courtois en clave Fantasy LaLiga (analiticafantasy.com)](https://analiticafantasy.com/estadisticas/courtois)

Sin embargo, la página que nos aportaba la información más detallada y precisa para nuestro objetivo fue:

[Estadísticas e Historia del Fútbol | FBref.com](https://fbref.com/es/?lang=es)

Además, decidimos que sería enriquecedor complementarla con datos oficiales de LaLiga:

[Avanzadas | LaLiga](https://www.laliga.com/estadisticas-avanzadas)

# Título

El título de un proyecto puede parecer algo superfluo y trivial, pero va a ser su imagen de marca, y el hecho de tener un nombre con gancho puede hacer que un proyecto se aprecie más, por tanto, nos tomamos esta tarea en serio y pusimos sobre la mesa diversas opciones:

* Coach’s notebook
* Net scraper
* FootData
* Players summary
* DataKeeper
* Data equaliser
* Goal data keeper

Sin embargo, finalmente nos quedamos con el que mejor definía nuestro proyecto y más atractivo nos parecía:

* **LaLiga stats**

# Descripción del dataset

# Representación gráfica

# Contenido

# Agradecimientos

En una tarea tan ambiciosa como esta podemos suponer que no habremos sido los primeros en intentarlo, algunos ejemplos que nos sirvieron de inspiración son:

[A simple method to predict player performance using Fantasy Football data | by Davide Totaro | Towards Data Science](https://towardsdatascience.com/a-simple-method-to-predict-player-performance-using-fantasy-football-data-8b2d3adb3a1a)

[project-report (4).pdf (imperial.ac.uk)](https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf)

[Machine Learning Algorithms for Football Predictions | by Matheus Kempa | Towards Data Science](https://towardsdatascience.com/machine-learning-algorithms-for-football-prediction-using-statistics-from-brazilian-championship-51b7d4ea0bc8)

Podemos ver que incluso uno de ellos se basa en datos de FBREF para su análisis.

**Consideraciones previas**

* *FBREF:*

Una vez decididas las fuentes de los datos el primer paso realizado ha sido la comprobación del fichero robots.txt, en el cual no aparece expresamente las rutas concretas de la web que nosotros utilizamos, sin embargo, cabe recordar que este documento es tan sólo una recomendación, y los recolectores de datos no están obligados a respetarlos. A pesar de ello es recomendable para evitar posibles bloqueos de la IP o u otro tipo de problemas con la página web.

Enlace a fichero robots.txt: <https://fbref.com/robots.txt>

Leyendo los términos de uso de la página descubrimos que el hacer recolección de datos de su página podía ser un problema para ellos, por ello contactamos para la página para pedirles su permiso explícito.

[Sports Reference Terms of Use | Sports-Reference.com](https://www.sports-reference.com/termsofuse.html)

Adjuntamos el correo preguntando si podemos utilizar los datos y la respuesta de la página web.



Cómo podemos ver, la página autoriza la recolección de los datos mientras el flujo de llamadas no produzca problemas en la página web, el cual hemos comprobado que no es perjudicial para la misma. Por tanto continuamos con nuestro proyecto.

* LaLiga

La página de LaLiga no pone trabas al uso de robots, tan sólo una latencia entre peticiones de 30 segundos, la cual respetamos para evitar ser bloqueados.

<https://www.laliga.com/robots.txt>

Podemos observar cómo en los términos de uso de la página tampoco impiden la recolección de datos para nuestro fin.

<https://www.laliga.com/informacion-legal/legal-web>

# Inspiración

En la introducción del proyecto hemos presentado nuestros objetivos, a pesar de ello a continuación los explicaremos con más detalle.

Las preguntas que queremos responder son:

* ¿Cuál es el rendimiento de los distintos jugadores en función de variables cómo la hora del partido, racha, jugar en casa o cómo visitante?
* ¿Cuál es el rendimiento de los jugadores en función de los compañeros con los que juegan?
* ¿Cuál es el rendimiento esperado de los jugadores para el próximo partido?
* En función de lo anterior, ¿Qué resultado esperamos? ¿Bajo que probabilidad?
* ¿Merece la pena con los datos obtenidos apostar en futuros partidos?

Enlazar con el punto 6

# Licencia

# Código

Enlace al repositorio de GIT: <https://github.com/KoNGoL/Tipologia_prac1.git>

El código ha sido desarrollado en lenguaje Python, usando las librerías BeautifulSoup y Request

# Dataset

# Video

# Contribuciones

|  |  |
| --- | --- |
| *Contribuciones* | *Firma* |
| *Investigación previa* | Adrián García Rodríguez  Tomás Jesús Luna López |
| *Redacción de las respuestas* | Adrián García Rodríguez  Tomás Jesús Luna López |
| *Desarrollo del código* | Adrián García Rodríguez  Tomás Jesús Luna López |