Documentación del proyecto centrado en el rendimiento de equipos y jugadores durante la temporada 2025/26 del Big 5

Introducción al proyecto

El objetivo de esta aplicación web es visualizar el rendimiento de los equipos y de los jugadores durante la temporada 2025/26. Los datos de los jugadores proceden de Opta The Analyst, FBref y Transfermarkt. Por lo tanto, este análisis se centra en la temporada 2025/26 para las siguientes competiciones: Ligue 1, Bundesliga, Premier League, La Liga y Serie A.
De manera concreta, esta aplicación permite evaluar el estilo de juego de cada equipo e identificar al jugador que el usuario desee en función de su rendimiento estadístico a lo largo de la temporada en curso.

Extracción de datos y preprocesamiento de los datos

- Equipo

Dadas las necesidades de datos relacionadas con el rendimiento de los equipos durante la temporada actual, hemos recopilado información procedente de varios proveedores de datos.

En el caso de **Opta**, realizamos scraping de los datos de las cinco grandes ligas recuperando las siguientes tablas: **Attack, Defense, Pressing, Sequence y Other**. Todas las estadísticas de estas tablas se recopilan, salvo las redundantes, y se renombran utilizando el nombre de su tabla correspondiente como prefijo. Además, estos datos se actualizan semanalmente mediante GitHub Actions para cada campeonato.

Posteriormente, estructuramos nuestra base de datos a partir de las estadísticas disponibles en **FBref,** que ofrece una amplia variedad de indicadores avanzados que cubren las fases ofensivas y defensivas, la posesión, la presión y las acciones del portero. FBref organiza sus datos en múltiples categorías: **Standard Stats, Shooting, Passing, Pass Types, Goal and Shot Creation, Defensive Actions, Possession, Playing Time, Miscellaneous Stats**, entre otras.

Para integrar esta información en nuestro modelo, primero obtuvimos los datos de los equipos mediante scraping en Python. Históricamente, todo este proceso estaba automatizado, pero actualmente utilizamos una extracción semi manual asistida por un script en Python con el fin de garantizar un mayor control sobre la limpieza y la estructuración de los datos. Este proceso incluye una fase de limpieza en la que se corrigieron los nombres de las ligas, se armonizaron diversas variables textuales y se eliminaron columnas redundantes como Rk, Squad, Comp, etc.

A esto se añaden los datos salariales de los equipos, integrados para aportar una dimensión económica al análisis, así como la posición de cada equipo en la clasificación de su campeonato.

Posteriormente, unimos todas estas variables a nivel de equipo mediante un **mapeo basado en los nombres de los equipos**, con el objetivo de obtener una base de datos completa. También creamos varias variables derivadas de la información obtenida previamente, como la clasificación de los equipos en su liga, el porcentaje de ataques según el estilo de juego de cada equipo, o el ajuste de las métricas defensivas en función de la posesión del balón.

A partir de este conjunto de datos, establecimos un sistema de puntuación por categorías en función de las estadísticas disponibles. Estas categorías se dividen en cuatro grandes grupos:

- **Estadísticas con balón**: Ocasiones creadas, Finalización, Balón parado ofensivo, Creación de ocasiones, Progresión, Centros, Regates.
- **Estadísticas sin balón**: Ocasiones concedidas, Acciones defensivas, Balón parado defensivo, Eficacia del portero.
- **Estadísticas relacionadas con el estilo de juego**: Juego directo, Contraataques, Posesión, Presión.
- **Otras**: Clasificación en la liga, Duelos terrestres, Duelos aéreos, Faltas recibidas, Faltas cometidas, Pérdidas de balón, Sustituciones.

Cada una de estas subcategorías contiene varias estadísticas asociadas, cada una con un **coeficiente** que refleja su importancia para caracterizar el grupo. Cabe destacar que la elección de estas métricas y coeficientes se realizó de forma conjunta por los dos miembros del proyecto y, por lo tanto, es totalmente subjetiva.

De la misma manera, cada categoría estadística dispone de un coeficiente que permite obtener una valoración global de cada equipo, comprendida entre 0 y 100. Esta valoración determina la clasificación del Power Ranking entre los 96 equipos del Big 5. Por último, tuvimos en cuenta la fortaleza de cada campeonato aplicando una **ligera penalización basada en el índice de potencia de las ligas** calculado por Opta en mayo de 2025.

- Jugador

Para el análisis individual de los jugadores, construimos nuestra base de datos a partir del conjunto de datos disponible en **Kaggle**: https://www.kaggle.com/datasets/hubertsidorowicz/football-players-stats-2025-2026

Anteriormente, esta información también se obtenía mediante scraping, pero este proceso ya no se utiliza. Actualmente nos apoyamos directamente en este archivo, lo que garantiza una mayor coherencia y una reducción del ruido en los datos iniciales.

Posteriormente, se llevó a cabo una fase de limpieza exhaustiva para armonizar y validar las variables. Entre las principales operaciones realizadas se incluyen:

- **Estandarización de las posiciones**: las abreviaturas fueron reemplazadas por su forma completa para obtener categorías más explícitas, por ejemplo:
  - DF -> Defender
  - MF -> Midfielder
  - FW -> Forward
  - GK -> Goalkeeper
- **Corrección y estandarización de los nombres de los países**, sustituyendo las formas abreviadas por su denominación completa.
- **Eliminación de duplicados**, conservando un único registro por jugador.

Además, integramos datos salariales procedentes de **Capology**, lo que enriquece la base de datos con una dimensión económica clave para el análisis del rendimiento y del valor de los jugadores.

Para completar nuestra base de información, también integramos datos de **Transfermarkt** utilizando la API no oficial disponible en: https://github.com/felipeall/transfermarkt-api

Siguiendo las instrucciones proporcionadas, instalamos **Docker** y configuramos el entorno necesario para realizar consultas a la API. Este proceso nos permite obtener, en una primera fase, la lista de clubes y posteriormente los jugadores pertenecientes a cada uno de ellos. A partir de esta información, realizamos consultas adicionales para obtener datos detallados de cada jugador.

Los datos proporcionados por Transfermarkt son especialmente ricos y aportan un nivel de precisión que otras fuentes no ofrecen. Entre las variables más importantes que integramos en nuestro modelo se encuentran:

- La **posición natural del jugador**, descrita de forma granular (por ejemplo Left Winger, Right Back, Attacking Midfielder), en lugar de una categoría general.
- Los **datos de representación** (agente o estructura que gestiona al jugador).
- El historial y la situación actual de las **lesiones**.
- Un conjunto de información contextual que incluye:
    - nombre del jugador (player_name)
    - posición (position)
    - fecha de nacimiento (dateOfBirth)
    - edad (age)
    - nacionalidad (nationality)
    - club actual (currentClub)
    - altura (height)
    - pierna dominante (foot)
    - año de llegada al club actual (joinedOn / joined)
    - club de procedencia (signedFrom)
    - fecha de finalización de contrato (contract)
    - valor de mercado (marketValue)
    - status, que indica si el jugador está lesionado, sancionado o es capitán

A partir de estos datos, **agregamos determinadas estadísticas por 90 minutos y ajustamos otras en función de la posesión del equipo**. El objetivo del ajuste por posesión es analizar de manera objetiva las acciones defensivas como si el equipo del jugador tuviera el balón el 50 por ciento del tiempo. Posteriormente, realizamos una unión entre los dos proveedores de datos utilizando varias estrategias para incluir al mayor número posible de jugadores minimizando los errores.

La primera etapa consistió en asociar a los jugadores con el mismo nombre, la misma liga, el mismo año de nacimiento y la misma posición cuando se trataba de porteros, ya que es el único puesto en el que no existen diferencias entre proveedores. Esta fase permitió asociar a la mayoría de los jugadores, pero algunas variaciones en los nombres nos llevaron a profundizar más. Para ello, utilizamos una función de similitud textual (fuzzy matching) siguiendo la siguiente lógica: si el porcentaje de similitud es suficientemente alto y coincide la liga, el año de nacimiento y la posición en el caso de los porteros, entonces consideramos que se trata del mismo jugador.

Por último, realizamos asociaciones manuales para los jugadores cuyos nombres diferían considerablemente entre proveedores, pero cuya información restante coincidía perfectamente. Cabe señalar que cuanto menor es el porcentaje de similitud del nombre, mayor es el riesgo de error. Por esta razón, decidimos no bajar de un umbral del 65 por ciento para garantizar una asociación fiable.

A continuación, establecimos un **sistema de valoración** de los jugadores entre 0 y 100, tanto por categoría como a nivel global, en función del rendimiento de los demás jugadores que ocupan el mismo puesto y pertenecen a la misma categoría de posición. Una puntuación alta refleja un elevado nivel de rendimiento durante la temporada 2025/26. Por coherencia, algunas variables fueron invertidas, como los errores o las pérdidas de balón. Además, cada posición dispone de coeficientes específicos adaptados a las exigencias de su rol. Por ejemplo, un delantero con buen rendimiento debe presentar buenas estadísticas de finalización y creación de ocasiones. Del mismo modo, cada categoría estadística tiene un peso destinado a valorar a los jugadores que destacan en las dimensiones clave asociadas. Los porteros cuentan con su propio sistema de valoración debido a la especificidad de su posición. Por último, se pueden aplicar **penalizaciones a la nota global en función de la fortaleza del campeonato en el que juega el jugador y del porcentaje de minutos disputados durante la temporada**.

Lista de páginas

La aplicación se compone de dos grandes partes: análisis de equipos y análisis de jugadores.

Páginas dedicadas a los equipos (6 páginas)

- **Inicio:** Presentación del proyecto y de las fuentes de datos.
- **Análisis de un equipo:** Exploración detallada del equipo seleccionado a través de diferentes estadísticas.
- **Comparación entre equipos:** Comparación directa de dos equipos para un análisis comparativo.
- **Clasificación - Estadísticas básicas:** Clasificación de los equipos según una estadística simple seleccionada.
- **Clasificación - Estadísticas avanzadas:** Clasificación de los equipos según una estadística avanzada seleccionada.
- **Power Ranking:** Clasificación global de los 96 equipos del Big 5, así como una clasificación por campeonato.

Páginas dedicadas a los jugadores (6 páginas)

- **Inicio**: Presentación del proyecto y de las fuentes de datos.
- **Análisis de un jugador**: Análisis detallado del jugador seleccionado a través de diferentes estadísticas.
- **Comparación entre jugadores**: Comparación de dos jugadores que ocupan el mismo puesto.
- **Clasificación - Estadísticas básicas**: Clasificación de los jugadores según una estadística simple seleccionada.
- **Clasificación - Estadísticas avanzadas**: Clasificación de los jugadores según una estadística avanzada seleccionada.

- **Scouting**: Búsqueda de jugadores que cumplan los criterios definidos por el usuario (información general, estadísticas básicas, estadísticas avanzadas).

Estructura del proyecto para el desarrollo de la aplicación

Application
├── scripts (team, player, big_5_performance.py): Todos los scripts del proyecto por tipo
├── data (team, player): Todos los datos del proyecto agrupados por tipo
├── image: Imágenes utilizadas en el proyecto
├── README.md
├── .github: Todos los workflows utilizados para la automatización de datos
└── requirements.txt: Archivo de dependencias de librerías

Puesta en marcha de la aplicación

Es importante destacar que la aplicación contará con tres versiones: francés, inglés y español.

Inicio

Tal como se explicó anteriormente, la página de inicio presentará brevemente los componentes del proyecto y dará acceso a diversos recursos, como la documentación y el código fuente del proyecto.

Sección de análisis

Encabezado del proyecto

Previamente, crearemos varias funciones (construcción de tablas de análisis, glosario, traducciones, etc.), así como diversas variables (glosario, listas de datos por categoría), con el fin de evitar una longitud excesiva del proyecto, teniendo en cuenta las tres lenguas disponibles y la similitud entre los distintos tipos de análisis.

Visualización de la aplicación

- **Equipo**

Para el **análisis de equipos**, el usuario deberá seleccionar primero el campeonato de su elección para poder elegir el equipo que desea analizar. El análisis se estructura en varias secciones: presentación de la información general del equipo (nombre, campeonato, clasificación, power ranking), visualización del top 5 de sus jugadores de la temporada según nuestros criterios de rendimiento, lista de los cinco jugadores con mayor valor de mercado según Transfermarkt, así como varios gráficos comparativos (ofensivos, defensivos, estilo de juego y otros) que permiten situar al equipo en relación con el Big 5 o con su propio campeonato. También se propone una lista de cinco equipos similares a partir de las estadísticas disponibles.

La sección **Duel** ofrece una comparación entre dos equipos seleccionados por el usuario, enfrentando sus estadísticas para facilitar el análisis comparativo.

La sección **Stats+** permite obtener una clasificación basada en una estadística agregada por categoría seleccionada (creación de ocasiones, finalización, regates, etc.), con la posibilidad de filtrar por campeonato.

La sección **Stat** sigue la misma lógica, pero aplicada al conjunto de estadísticas brutas disponibles (xG, PPDA, número de contraataques por 90 minutos, etc.). El usuario debe seleccionar previamente la categoría de la estadística deseada para facilitar la navegación. También se pone a disposición un glosario para explicar cada estadística.

Por último, la sección **Power Ranking** permite mostrar de forma clara la clasificación de los 96 equipos, con la posibilidad de filtrar por campeonato y comparar rápidamente sus estadísticas respectivas.

- **Jugador**

Para el **análisis de jugadores**, en primer lugar se pedirá al usuario que seleccione al jugador de su elección. Esta página mostrará el perfil del jugador con su información básica (nombre, foto, posición, club, etc.), su radar estadístico y una tabla con los cinco jugadores más similares desde el punto de vista estadístico. El radar puede adaptarse según el país, el campeonato o el grupo de edad. Las estadísticas mostradas en el radar son las más relevantes para la posición del jugador, representadas en azul, y se comparan con la mediana del grupo seleccionado, representada en rojo. Además, se dispone de un glosario de estadísticas desplegable si es necesario.

La página de **comparación de jugadores** sigue la misma lógica, permitiendo seleccionar dos jugadores del mismo puesto que el usuario desea analizar. Sus perfiles se muestran uno junto al otro, seguidos de sus respectivos radares en azul y rojo.

En la página de **clasificación de estadísticas básicas**, el usuario puede obtener un ranking según la categoría de su elección (finalización, regate, etc.), con la posibilidad de filtrar por posición, grupo de edad, país, entre otros. Se muestra un podio de los mejores jugadores según estos criterios, junto con información general de todos los jugadores que cumplen con la búsqueda del usuario.

De manera similar, en la página de **clasificación de estadísticas avanzadas**, se solicita al usuario la estadística que desea analizar. A partir de esta elección, se muestra una clasificación de los mejores jugadores según dicha métrica, con un podio para los tres primeros, seguido de una tabla con su información básica. La barra lateral incluye filtros opcionales (posición, club, campeonato, grupo de edad, valor de mercado), así como un glosario de estadísticas. Además, se muestra una imagen antes de la selección del usuario para evitar que la página quede vacía.

Por último, la página **Scout** permite obtener una lista de jugadores según los criterios definidos por el usuario (información básica, estadísticas básicas y avanzadas). Un resumen de las selecciones realizadas está disponible en la barra lateral, junto con un glosario de estadísticas.