Máster en Ciencias de Datos

Tipología y ciclo de vida de los datos

Web scraping de ligas de fútbol europeas

Gerard Casanovas Jiménez,
Sergio Merino Premió

PR 1
20 de abril de 2023

Índice

1. Contexto

El contexto de los datos son las temporadas actuales de fútbol de distintas ligas. La página presenta la clasificación y distintas estadísticas organizadas en tablas, todo de una manera que simplifica el poder recoger todas las tablas de la propia página.

2. Descripción del dataset

Para nuestro ejemplo hemos decidido seleccionar las dos tablas más importantes para nuestro objetivo.

- **team**: Equipo.
- **games**: Partidos Jugados.
- **points**: Puntos.
- **wins**: Partidos Ganados.
- **ties**: Partidos Empatados.
- **losses**: Partidos Perdidos.
- **goals_for**: Goles a Favor.
- **goals_against**: Goles en Contra.
- **goal_diff**: Diferencia de Goles (goals_for - goals_against).
- **xg_for**: Expected Goals (Goles Esperados).
- **xg_against**: Expected Goals Against (Goles Esperados en Contra).
- **xg_diff**: Expected Goal Difference (Diferencia de Goles Esperada).
- **players_used**: Jugadores utilizados.
- **avg_age**: Edad promedio.
- **possession**: Posesión.
- **games_starts**: Partidos iniciados como titular.
- **minutes**: Minutos jugados.
- **minutes_90s**: Minutos jugados divididos entre 90 (para saber cuántos partidos completos se jugaron).
- **assists**: Asistencias.
- **goals_assists**: Suma de goles y asistencias.
- **goals_pens**: Goles anotados de penal.
- **pens_made**: Penales convertidos.
- **pens_att**: Penales intentados.
- **cards_yellow**: Tarjetas amarillas recibidas.
- **cards_red**: Tarjetas rojas recibidas.

3. Representación

En cuanto a la representación gráfica, nuestro dataset gira alrededor del equipo. Todas las columnas hacen referencias a estadísticas del equipo, por lo que visualmente queda un poco saturado.

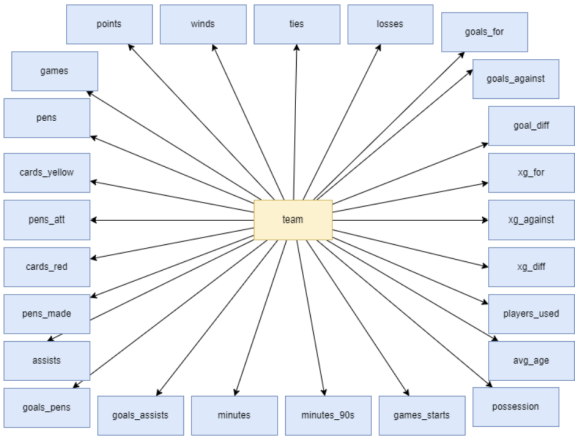

Figura 1

4. Contenido

Los datos pertenecen a la temporada actual, es decir, de agosto de 2022 hasta la fecha en la que se realiza la ejecución del código.

Los campos que se incluyen en el dataset buscan obtener información acerca de:

- **team**: Nombre del equipo de fútbol al que se refiere la tabla
- **games**: El número total de partidos que el equipo ha jugado en la competición.
- **points**: Los puntos acumulados por el equipo en la competición, que suelen otorgarse de acuerdo con las reglas de puntuación de la liga específica. Por lo general, se otorgan 3 puntos por una victoria, 1 punto por un empate y 0 puntos por una derrota.
- **wins**:El número de partidos que el equipo ha ganado en la competición.
- **ties**: El número de partidos que el equipo ha empatado en la competición.
- **losses**: El número de partidos que el equipo ha perdido en la competición.
- **goals_for**: El número total de goles marcados por el equipo en la competición.
- **goals_against**: El número total de goles recibidos por el equipo en la competición.
- **goal_diff**: La diferencia entre los goles a favor y los goles en contra del equipo en la competición (GF - GC).
- **xg_for**: La cantidad total de goles esperados que se estima que el equipo debería haber marcado en función de la calidad y cantidad de los disparos realizados.
- **xg_against**:La cantidad total de goles esperados que se estima que el equipo debería haber concedido en función de la calidad y cantidad de los disparos recibidos.

- **xg_diff**: La diferencia entre los xG a favor y los xG en contra del equipo (xG a favor - xG en contra)
- **players_used**:La cantidad total de jugadores diferentes que han participado en partidos oficiales para el equipo durante la temporada o competición.
- **avg_age**: La edad promedio de los jugadores que han participado en partidos oficiales para el equipo durante la temporada o competición.
- **possession**: Porcentaje de tiempo que el equipo ha tenido el balón en sus pies durante la competición
- **games_starts**: Partidos en los que el equipo ha empezado el partido
- **minutes**: Total de minutos en los que el equipo ha estado en el campo
- **minutes_90s**: Promedio que representa el número de minutos jugados por partido
- **assists**: Número de pases que han llevado a un gol anotado por otro jugador dentro del equipo
- **goals_assists**: Suma total de goles y asistencias del equipo
- **goals_pens**: Número de goles anotados por el equipo desde el punto de penalti
- **pens_made**: Número de penaltis cometidos por el equipo
- **pens_att**: Número total de penaltis que el jugador ha lanzado
- **cards_yellow**: Número de tarjetas amarillas recibidas por el equipo como sanción.
- **cards_red**: Número de tarjetas rojas recibidas por el equipo como sanción.

5. Propietario

Sports Reference LLC es una destacada empresa privada dedicada a la recopilación y análisis de datos deportivos en línea, reconocida por su fiabilidad como fuente de estadísticas deportivas en Internet. Ofrece una amplia gama de sitios web especializados en diversos deportes, como fútbol, béisbol, baloncesto, hockey sobre hielo, entre otros.

En el presente análisis, se describe el proceso de web scraping llevado a cabo en una página web perteneciente a Sports Reference LLC. Los detalles sobre dicho proceso de extracción de datos de fútbol utilizando técnicas de web scraping se encuentran disponibles en el repositorio de GitHub (`https://github.com/hoyishian/footballwebscraper`).

Se han seguido los siguientes pasos para actuar en concordancia con los principios éticos y legales:

1. Familiarización con las leyes y regulaciones: Se investigaron las leyes y regulaciones aplicables al web scraping en la región correspondiente. Además, se realizó una revisión previa de otros proyectos similares que llevaran a cabo web scraping en páginas similares.
2. Verificación del archivo robots.txt: Durante la ejecución de la práctica, se implementó una función llamada "_check_robots_file"que verifica el archivo robots.txt para asegurarse de que el web scraping esté permitido según las directrices del sitio web.
3. Limitación de datos recopilados: Se buscó no excederse en la recopilación de datos y se obtuvo únicamente la información necesaria para el proyecto, evitando la recopilación indiscriminada o excesiva de datos.

En resumen, se tomaron medidas para cumplir con los principios éticos y legales, asegurándose de que el web scraping se llevara a cabo de manera responsable y cumpliendo con las normas y regulaciones aplicables en la región correspondiente

6. Inspiración

La principal motivación detrás de este conjunto de datos radica en las oportunidades futuras que ofrece para llevar a cabo análisis de desempeño de equipos. Estos análisis son necesarios para que los equipos obtengan información valiosa sobre el rendimiento de sus jugadores y evalúen la efectividad de sus tácticas y estrategias, comparan su desempeño con el de otros equipos e identifiquen oportunidades de mejora.

Los conjuntos de datos generados pueden ser utilizados por departamentos de análisis de datos de equipos deportivos o por compañías especializadas en análisis de datos deportivos, como Statsbomb, que toman decisiones informadas basadas en datos en el ámbito deportivo.

7. Licencia

Este conjunto de datos se publica bajo la licencia CC BY-SA 4.0 License debido a que consideramos que es la más adecuada para el trabajo que he realizado. La licencia requiere que se proporcione el nombre del creador del conjunto de datos y se indiquen los cambios que se han realizado en relación con el trabajo original. De esta manera, se reconoce el trabajo ajeno y se indica en qué medida se han realizado aportaciones en relación con el trabajo original. La licencia permite también el uso comercial aumentando las posibilidades de obtener reconocimiento por el trabajo realizado. Asimismo, cualquier distribución posterior deberá distribuirse bajo esta misma licencia, asegurando la distribución del trabajo original bajo los mismos términos planteados inicialmente.

8. Dataset

`https://zenodo.org/record/7838607#.ZD2YbXbP1D8`