



Práctica 1

Tipología y Ciclo de Vida de los Datos

María Aránzazu Romero Moreno
Ronnel Ovalles Guerra

URL de los datos: <https://www.resultados-futbol.com/>

URL repositorio: <https://github.com/RonnelOvalles/Web-scraping-Resultados-Historicos-de-Partidos-de-Futbol>

URL dataset: <https://zenodo.org/records/10118973>

URL vídeo: https://drive.google.com/file/d/1le5LXBBbzIEoBDMdNi09yzmgx060MIYm/view?usp=drive_link

TABLA DE CONTENIDO

| | |
|---|----|
| 1.CONTEXTO..... | 2 |
| LA FUENTE DE DATOS..... | 2 |
| 2.TÍTULO DEL DATASET..... | 3 |
| 3.DESCRIPCIÓN DEL DATASET..... | 4 |
| 4.REPRESENTACIÓN GRÁFICA | 5 |
| 5.CONTENIDOS DEL DATASET | 5 |
| 6.PROPIETARIO DE LA WEB Y DEL CONJUNTO DE DATOS | 6 |
| 7.INSPIRACIÓN | 8 |
| 8.LICENCIA | 9 |
| 9.CÓDIGO..... | 10 |
| 10.DATASET..... | 11 |
| 11.VIDEO | 11 |

CONTEXTO

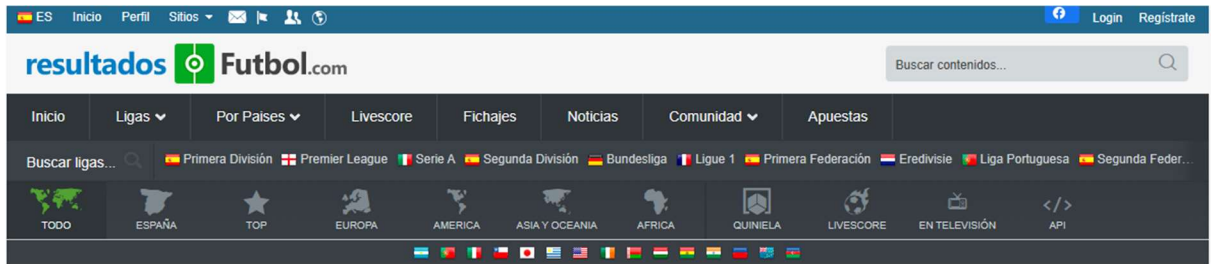
El mundo del fútbol profesional es apasionante para muchas personas y moviliza a millones de aficionados en todo el mundo, así como un gran volumen financiero. Los partidos, los equipos y los resultados son seguidos con gran interés, por otra parte, el análisis de los datos de las ligas y competiciones resulta fundamental tanto para los aficionados como para los profesionales del sector, además de periodistas, analistas y todo el entramado de las apuestas deportivas.

En este contexto, se ha decidido recolectar información sobre los marcadores de partidos de fútbol de diversas ligas alrededor del mundo. La recolección de estos datos permite tener una visión actualizada y detallada de los resultados de las competiciones, lo cual pensamos que es de utilidad para realizar, por ejemplo, análisis de rendimiento de los equipos, estadísticas para medios de comunicación o incluso para el desarrollo de estrategias en las apuestas deportivas.

LA FUENTE DE DATOS

El sitio web seleccionado para la recolección de estos datos es el portal *Resultados-Futbol.com*.

URL: <https://www.resultados-futbol.com/>



El portal se especializa en ofrecer información actualizada y muy diversa sobre las competiciones de fútbol más importantes del mundo. Es una fuente fiable y adecuada para recopilar información actualizada en tiempo real y obtener datos precisos sobre estas competiciones y las noticias relacionadas con el sector.

TÍTULO DEL DATASET

"Resultados Históricos de Partidos de Fútbol: Ligas Globales, Fechas y Marcadores"

El título ha sido seleccionado por su capacidad para encapsular de manera clara y concisa la esencia del dataset.

- Se establece el contexto general del dataset, indicando que se trata de un registro de resultados de partidos de fútbol.
- Sugiere que los datos abarcan múltiples fechas y no un periodo específico.
- Se destaca que no se restringe a una liga o país en particular, sino que puede incluir datos sobre diferentes competiciones del mundo. Con esto se sugiere la diversidad y amplitud del dataset.
- La inclusión de las diferentes variables en el título también sugiere que el dataset puede ser utilizado para realizar análisis temporales y comparativos de resultados.

DESCRIPCIÓN DEL DATASET

El dataset es una recopilación estructurada de los resultados de partidos de diferentes ligas y competiciones de fútbol.

Aunque el sitio web de origen ofrece gran variedad de datos, la selección realizada responde a la intención de capturar la esencia de cada partido, proporcionando información clave. Por lo tanto, se ha optado por capturar los resultados finales, los equipos participantes, la liga y la fecha, ya que son elementos fundamentales para entender el desempeño histórico de los equipos y las ligas y permiten realizar análisis comparativos y temporales de manera directa.

La elección de estos campos específicos también se ha realizado pensando en la accesibilidad y la comprensión del público general, pues el dataset se convierte en una herramienta de amplia utilidad para una gran variedad de usuarios, desde entusiastas del fútbol hasta analistas deportivos profesionales.

REPRESENTACIÓN GRÁFICA



CONTENIDOS DEL DATASET

El dataset generado a partir del proceso de web scraping está estructurado para proporcionar una visión detallada de las competiciones de fútbol. A continuación, se detallan los campos que constituyen este conjunto de datos y el periodo de tiempo que cubren.

1. **Fecha:** Indica la fecha en que se disputó el partido
2. **Liga:** Indica la liga o competición en la que se celebró el partido.
3. **Equipo 1:** Nombre del equipo local o primer equipo participante.
4. **Equipo 2:** Nombre del equipo visitante o segundo equipo participante.

5. **Resultado:** Refleja el marcador final del partido, indicando los goles de cada equipo.

| Fecha | Liga | Equipo 1 | Equipo 2 | Resultado |
|----------|----------------|-------------------|--------------|-----------|
| 04/11/23 | Premier League | Manchester United | FC Barcelona | 0-0 |

El código extrae cada uno de estos datos como texto (string) y se almacena de esta manera en el archivo CSV que se genera tras finalizar el web scraping. Sin embargo, para un análisis posterior, algunos de estos campos podrían ser convertidos a diferentes tipos de datos, como la fecha o el resultado del partido, en los que podría ser más útil dividirlos en campos separados que permitan facilitar los cálculos estadísticos.

El dataset está diseñado para recoger datos de partidos de fútbol de hasta 100 días anteriores a la fecha actual. No obstante, esta cantidad puede ser ajustada por el usuario según sus necesidades, lo cual permite mayor flexibilidad en la cantidad de datos históricos recopilados.

PROPIETARIO DE LA WEB Y DEL CONJUNTO DE DATOS

El conjunto de datos original es propiedad de la empresa **RESULTADOS DE FÚTBOL, S.L.**, cuya plataforma en línea, Resultados de Fútbol, ofrece información detallada sobre los marcadores de partidos de fútbol. La empresa está registrada en Málaga, España, y puede ser contactada a través de su dirección de correo electrónico: info@resultados-futbol.com.

Según el Aviso Legal del sitio web, la información, textos, fotografías, gráficos, imágenes, y demás contenidos disponibles en la plataforma están

protegidos por derechos de propiedad intelectual e industrial. Por lo tanto, no pueden ser reproducidos, distribuidos, comunicados públicamente, transformados o modificados sin autorización expresa de su legítimo titular. Sin embargo, no se encontró una declaración explícita sobre la permisibilidad del web scraping. Por lo tanto, es fundamental asegurarse de realizar el web scraping de forma ética, respetuosa, y en concordancia con las leyes y normativas vigentes.

Para garantizar el cumplimiento de estos principios éticos y legales, en este proyecto se ha propuesto minimizar la intrusión en el sitio web, extrayendo únicamente los datos relevantes para el estudio y asegurando no causar ningún perjuicio a su rendimiento.

La web de Resultados de Fútbol no proporciona detalles específicos sobre cómo la plataforma recopila sus datos. Solo menciona que la información se obtiene de "fuentes oficiales", pero no especifica cuáles son estas fuentes ni describe el proceso de recopilación de datos. Además, el sitio web señala que no es responsable de la exactitud y veracidad de la información publicada en la plataforma y esto es así porque los usuarios de la plataforma tienen la capacidad de introducir, editar, gestionar y actualizar la información, lo que sugiere que estos usuarios son una de las fuentes de datos para la plataforma.

Dado que no hay más detalles disponibles en el Aviso Legal del sitio web, no es posible hacer afirmaciones definitivas sobre cómo la plataforma recopila y verifica sus datos. Sin embargo, es común que se obtengan a través de fuentes como federaciones deportivas, organizaciones de Ligas u otros servicios de datos deportivos, además de las aportaciones de los usuarios.

Existen proyectos de web scraping sobre la web Resultados-Futbol.com que son similares a este que se presenta. Por ejemplo:

- https://github.com/dnunezs/Scraping_Resultados_Futbol

Sin embargo, se centra en resultados para la Liga española de Primera y Segunda división y recopilan los datos por temporadas completas en lugar de permitir seleccionar el número de fechas e incluir diferentes tipos de competición.

INSPIRACIÓN

El conjunto de datos de la plataforma Resultados-Futbol.com es interesante debido a la constante popularidad e influencia cultural del fútbol en todo el mundo. La información sobre partidos, ligas, equipos y jugadores es relevante para varios perfiles de stakeholders, por lo que analizarlo puede proporcionar descubrimientos sobre tendencias, patrones y estadísticas valiosas.

Las preguntas que se pueden responder con este conjunto de datos son diversas, entre las que pueden incluirse las siguientes:

1. ¿Cuáles son los equipos con mayores logros en diferentes ligas y temporadas?
2. ¿Cómo ha evolucionado el rendimiento de un equipo a lo largo del tiempo?
3. ¿Existen patrones o tendencias en los resultados de los partidos que puedan ser útiles para las predicciones?
4. ¿Cómo se distribuyen los goles durante a lo largo de las temporadas?

Tal como se ha mencionado anteriormente, existen diversos análisis y proyectos que tratan sobre datos deportivos. Sin embargo, muchos de ellos no permiten acceder al proceso de recopilación de datos o no facilitan la data en un formato que otorgue la flexibilidad necesaria para hacer análisis personalizados.

Es por este motivo que el presente proyecto busca, no sólo recopilar los datos necesarios para responder a las preguntas descritas en este mismo apartado,

sino también permitir que otros puedan acceder al proceso de recopilación, que puedan personalizar el número de registros deseado y así puedan utilizarlos para realizar sus propios análisis.

LICENCIA

El sitio web Resultados-futbol.com establece en sus términos de uso una restricción comercial sobre los datos que recoge. Sin embargo, centrándonos en los datos de resultados deportivos y no en otros que pueda contener el sitio web, se entiende que estos datos provienen originariamente de una fuente pública como las páginas web de los clubes y ligas de fútbol o de la prensa deportiva, entre otros.

Considerando que el propósito del web scraping es facilitar el acceso y análisis de estos datos para la comunidad, una licencia que permita la distribución y uso, sin fines comerciales, y que, al mismo tiempo, reconozca el esfuerzo del creador del dataset, sería la mejor opción.

En este sentido, una opción adecuada podría ser: **CC BY-NC-SA 4.0 Licence**, que permite el uso de los datos para propósitos no comerciales, requiriendo atribución y compartiendo bajo términos similares.

- **Reconocimiento** (BY): Se debe proporcionar el crédito correspondiente, proporcionar un enlace a la licencia e indicar si se realizaron cambios.
- **No Comercial** (NC): El material no puede ser utilizado con fines comerciales, lo cual estaría en línea con las restricciones de la fuente original.
- **Compartir bajo mismos términos** (SA): Si se combinan, transforman o crean nuevos datos a partir del material, se deben distribuir estas contribuciones bajo la misma licencia que el original.

CÓDIGO

El código implementado está escrito en lenguaje Python y utiliza la biblioteca Selenium(**selenium==3.141.0**) para realizar el web scraping. El código fuente se halla en la carpeta `/source` del repositorio de Github y las dependencias y sus respectivas versiones están detalladas en el archivo *requeriments.txt*.

URL repositorio: <https://github.com/RonnelOvalles/Web-scraping-Resultados-Historicos-de-Partidos-de-Futbol>

El proceso de recolección de los datos comienza por acceder a la URL de Resultados-futbol.com usando Chrome como navegador web automatizado, sin interfaz gráfica, a través del driver proporcionado por Selenium. Una vez en la página web, se extrae la fecha de la jornada actual y se recopilan los datos de los partidos finalizados, configurado por defecto hasta un máximo de 100 jornadas, y que incluye datos como la liga, los equipos que participan y el resultado final del evento. La selección del intervalo de tiempo es flexible, permitiendo al usuario especificar la cantidad de días para los cuales extraer los datos.

El código utiliza intervalos de tiempo de 0,1 segundos para cargar los datos entre dos fechas. No obstante, está diseñado para manejar situaciones donde la web puede sufrir retrasos, intentando recuperarlos nuevamente en caso de fallos temporales en intervalos de 1 segundo. Esta estrategia de manejo de errores proporciona robustez, asegurando intentos adicionales para recuperar los datos que no han cargado correctamente.

Una vez que los datos se han extraído, se organizan en una lista de listas en Python, donde cada sublista representa un registro con la información de un partido, reservando la primera fila para los encabezados que describen cada campo. La estructura de cada registro es la siguiente:

Una vez que todos los datos se han recopilado y estructurado de esta manera, se escriben en un archivo CSV, y, para ello, se utiliza la librería csv de Python. El archivo CSV se crea en la carpeta /dataset del mismo directorio donde se encuentra el script, y se nombra como "resultado_de_partidos.csv".

Los datos en este formato son fácilmente accesibles y manipulables para las etapas posteriores del análisis, ya que el formato CSV es compatible con la mayoría de las herramientas y lenguajes de programación.

DATASET

Enlace Zenodo: <https://zenodo.org/records/10118973>






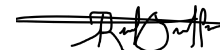


Enlace del DOI: **10.5281/zenodo.10118973**

Repositorio Github: <https://github.com/RonnelOvalles/Web-scraping-Resultados-Historicos-de-Partidos-de-Futbol>

VIDEO

Enlace Drive UOC:

[https://drive.google.com/drive/folders/1CuAUQSkZSu7Fz7g34\\$4arePH8Evelg_F](https://drive.google.com/drive/folders/1CuAUQSkZSu7Fz7g34$4arePH8Evelg_F)

| Contribuciones | Firma |
|-----------------------------|--|
| Investigación previa |   |
| Redacción de las respuestas |   |
| Desarrollo del código |   |
| Participación en el vídeo |   |