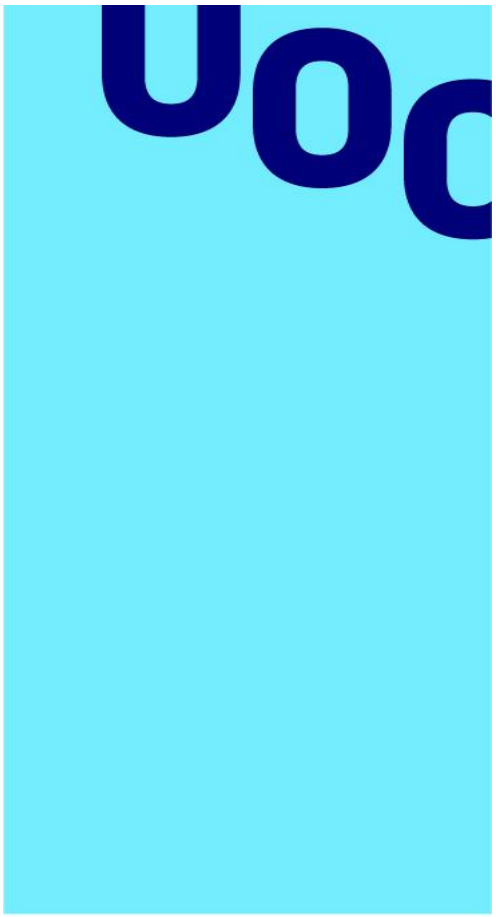
Tipología y Ciclo de Vida de los Datos

PRÁCTICA 1

Víctor Díaz Bustos: victordiazb@uoc.edu
Marc Moreno González: marcmg98@uoc.edu



Universitat Oberta de Catalunya

Índice

1. Contexto

Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

Siendo los deportes uno de nuestros principales hobbies, nos decantamos por realizar esta práctica orientada a algo que estuviese relacionado con ello.

Además, es sabido, que el futuro de los clubes y equipos profesionales, pasa por estudiar, analizar y sintetizar los datos que se generan de los partidos, empezando por lo básico que es el resultado y siguiendo con otra información como ocasiones de gol, porcentaje de posesión y en qué áreas del campo (para el fútbol) y porcentajes de tiro, asistencias (para el fútbol). Las conocidas como estadísticas avanzadas del deporte.

Pese a que el deporte es en gran parte impredecible, se pueden estudiar las tendencias y adaptar las tácticas de los equipos según el equipo, época del año e incluso racha del equipo rival al que te enfrentes.

Es por estos motivos, que hemos optado por recoger este tipo de datos procedentes de la web oficial del periódico diario As. Esta web se caracteriza por recopilar resultados, así como noticias relacionadas con los deportes y sus grandes ligas. A continuación facilitamos la dirección URL de la que hemos extraído los datos:

URL = "https://resultados.as.com/resultados/"

2. Título

Definir un título conciso y que sea descriptivo para el dataset.

Los títulos elegidos para los dos deportes trabajados son los siguientes:

- partidos_baloncesto.csv
- partidos_futbol.csv

3. Descripción del dataset

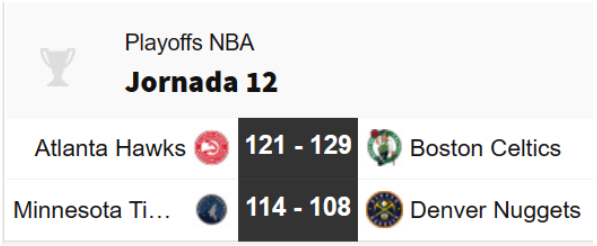
Desarrollar una breve descripción del conjunto de datos que se ha extraído. Es necesario que esta descripción sea coherente con el título elegido.

Estos conjuntos de datos recogen resultados de partidos de fútbol y baloncesto de varias ligas y competiciones. Contiene información sobre la competición, la jornada y la fecha del partido, así como los equipos locales y visitantes y los goles/puntos marcados por cada equipo.

Hay datos de varias ligas como Laliga Smartbank, Laliga Santander, Serie A en el caso del fútbol, o la Liga ACB y la NBA en el caso del baloncesto.

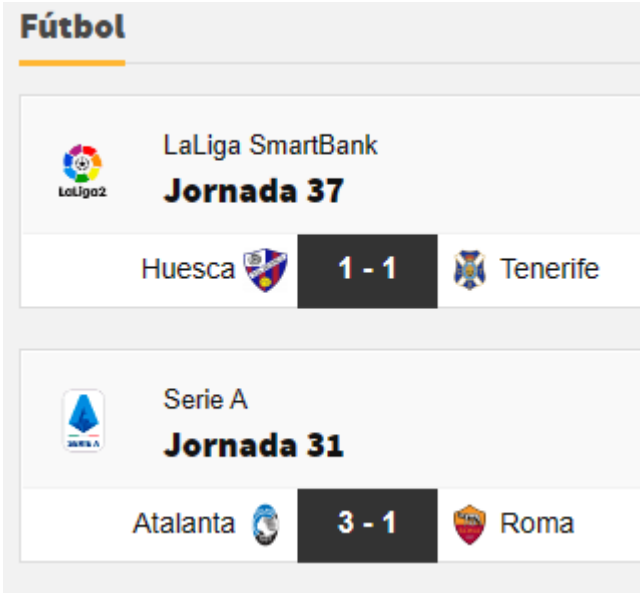
4. Representación gráfica

Dibujar un esquema o diagrama que refleje visualmente el dataset y el proyecto elegido.



DATASET

- Competición
- Jornada
- Fecha
- Equipo Local
- Equipo Visitante
- Goles / Puntos Local
- Goles / Puntos Visitante



5. Contenido

Explicar los campos que se incluyen en el dataset y el período de tiempo al que pertenecen los datos.

	Competicion	Jornada	Fecha	Equipo local	Equipo visitante	Puntos local	Puntos visitante
2	Playoffs Nba	Jornada 13	2023/04/24	Philadelphia 76Ers	Brooklyn Nets		
3	Playoffs Nba	Jornada 12	2023/04/24	Atlanta Hawks	Boston Celtics	121	129
4	Playoffs Nba	Jornada 12	2023/04/24	Minnesota Timberwolves	Denver Nuggets	114	108
5	Acb	Jornada 29	2023/04/23	Baxi Manresa	Ucam Murcia	93	83
6	Acb	Jornada 29	2023/04/23	Lenovo Tenerife	Rio Breogan	85	67
7	Acb	Jornada 29	2023/04/23	Casademont Zaragoza	Unicaja	70	74
8	Acb	Jornada 29	2023/04/23	Monbus Obradoiro	Real Betis	67	82
9	Acb	Jornada 29	2023/04/23	Joventut Badalona	Real Madrid	76	86
10	Acb	Jornada 29	2023/04/23	Carplus Fuenlabrada	Cazoo Baskonia	93	112
11	Playoffs Nba	Jornada 11	2023/04/23	Miami Heat	Milwaukee Bucks	121	99
12	Playoffs Nba	Jornada 11	2023/04/23	Los Angeles Lakers	Memphis Grizzlies	111	101
13	Playoffs Nba	Jornada 12	2023/04/23	New York Knicks	Cleveland Cavaliers	102	93
14	Playoffs Nba	Jornada 12	2023/04/23	Golden State Warriors	Sacramento Kings	126	125

El conjunto de datos contiene los siguientes campos:

- Competición: El nombre de la competición en la que se jugó el partido, como Laliga Smartbank, Laliga Santander, Serie A, Premier League y Superliga Turquia.

- Jornada: El número de la jornada en la que se jugó el partido.

- Fecha: La fecha en la que se jugó el partido, en formato AAAA/MM/DD.

- Equipo local: El nombre del equipo local que jugó el partido.

- Equipo visitante: El nombre del equipo visitante que jugó el partido.

- Goles local/Puntos local: La cantidad de goles que marcó el equipo local durante el partido. Si el campo está vacío, indica que el partido aún no se ha disputado y por lo tanto en el momento de recoger los datos, estos todavía no estaban disponibles.

- Goles visitante/Puntos visitante: La cantidad de goles o puntos que marcó el equipo visitante durante el partido.

Si el campo está vacío, indica que el partido aún no se ha disputado y por lo tanto en el momento de recoger los datos, estos todavía no estaban disponibles.

Los datos en el conjunto de datos pertenecen a un período de tiempo específico, que no se especifica claramente en el conjunto de datos, pero se puede deducir de la fecha de los partidos. La última fecha registrada en el conjunto de datos es el 24 de abril de 2023 y la web de AS proporciona estos datos al menos desde el año 2021.

6. Propietario

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en su defecto, justificar esta búsqueda con análisis similares. Indicar qué pasos se han seguido para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto elegido.

Los datos del conjunto de datos histórico de resultados de partidos de fútbol y baloncesto han sido recogidos de la página web del periódico As. Por lo tanto, se entiende que el propietario de los datos es el periódico As.

En el proceso de obtención de datos se trató de realizar web scraping a la web del diario Marca, sin éxito, ya que hemos tenido problemas con obstáculos que no hemos podido resolver como barreras de acceso denegado. Por ello, finalmente hemos optado por la web del diario As.

No obstante, al ver que la web de Marca nos ofrecía impedimentos para extraer información, nos planteamos extraer otro tipo de información como productos de comida y sus precios comparados para diferentes supermercados o de un modo similar, tipos de prendas y su precio para diferentes marcas de ropa como Zara o Mango.

7. Inspiración

Explicar por qué puede ser interesante este conjunto de datos y qué preguntas se pretenden responder con ellos. Es necesario comparar con los análisis anteriores o análisis similares presentados en el apartado 6.

La base de datos con el histórico de resultados de partidos de fútbol y baloncesto puede ser interesante por varias razones:

- **Análisis estadísticos**: Los datos pueden ser utilizados para realizar análisis estadísticos y generar estadísticas sobre los equipos y competiciones. Por ejemplo, se puede analizar la distribución de victorias y derrotas de un equipo en una competición específica y el promedio de goles por partido, entre otros.
- **Identificación de tendencias**: Los datos históricos también pueden ser utilizados para identificar tendencias a lo largo del tiempo. Por ejemplo, se pueden analizar los cambios en el número de goles marcados por partido en una determinada competición a lo largo de los años.
- **Predicciones**: Los datos históricos también pueden ser utilizados para predecir los resultados de partidos futuros. Al analizar los datos de equipos, se pueden utilizar modelos de aprendizaje automático y técnicas de minería de datos para predecir la probabilidad de que un equipo gane un partido contra otro equipo determinado.

8. Licencia

Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Otra (especificar cuál).

La licencia seleccionada es *Open Database License (ODbL)* con la cláusula de *"Solo Compartir-Añadir" (ODbL-SA)*, que establece que los usuarios pueden compartir, copiar y redistribuir la base de datos y agregar nuevos datos, pero no pueden modificar los datos ya existentes en la base de datos.

Se ha seleccionado esta licencia ya que el resultado de estos partidos es público, pues no nos interesa mantener esta información privada. Además, no tiene ningún efecto eliminar o modificar el resultado de partidos que ya se han jugado y cuyo resultado no se va a modificar, pero sí puede ser interesante que los usuarios puedan ir registrando los nuevos partidos que se disputen en el futuro.

9. Código

Código implementado para la obtención del dataset, preferiblemente en Python o, alternativamente, en R.

- El código deberá ubicarse en la carpeta /source del repositorio.
- Se deben indicar las librerías y versiones utilizadas. P. ej., en Python pueden obtenerse mediante el comando pip3 freeze > requirements.txt
- En la memoria en PDF, se deben comentar los aspectos más relevantes sobre cómo el código realiza el proceso de recolección de datos, qué dificultades presenta el sitio web elegido, y cómo se han resuelto.

El código está hecho en python. En primer lugar, se establece el *User-Agent* como *"Mozilla/5.0"* y se utiliza la librería *requests* para obtener el código HTML y la librería *BeautifulSoup* para convertir el HTML en un objeto de árbol de elementos, que se puede analizar y manipular para extraer información específica de la página web.

Posteriormente, se va navegando por la página web de las diferentes fechas. Para ello, se obtiene la URL ubicada en la etiqueta *<a class="slick-prev slick-arrow">*.

Para la web de cada fecha, se obtienen las distintas competiciones agrupadas por deportes, en este caso, fútbol y baloncesto, ubicadas en la etiqueta * * y *<h2 class="tit-decoration2">*, respectivamente. Se obtienen los partidos pertenecientes a cada competición mediante la etiqueta *<li class="list-resultado">* y se extraen los siguientes campos ubicados en sus respectivas etiquetas:

- nombre_competicion: * *
- jornada: * *
- equipo_local: * *
- equipo_visitante: * *
- puntos_local: *<a class="resultado">*
- puntos_visitante: *<a class="resultado">*

Finalmente se almacenan los datos obtenidos en un fichero mediante la librería *csv*.

Todo el código se puede encontrar en nuestro repositorio de github https://github.com/VictorDiazBustos/practica1_TCVD.

10. Dataset

Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción de la misma. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta /dataset del repositorio.

Si existe alguna circunstancia que impida publicar abiertamente el dataset real en Zenodo, se deberá: (1) comentar esta circunstancia y justificar el motivo en este apartado; (2) generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI; y (3) comunicar al profesor el dataset real de forma privada (p. ej., utilizando un repositorio privado).

https://doi.org/10.5281/zenodo.7860755

11. Video

Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

https://drive.google.com/file/d/19e44v5eunBizSeMUs_TEqOaZhTWX31KX/view?usp=sharing

CONTRIBUCIONES

Contribuciones	Firma
Investigación previa	MM, VD
Redacción de las respuestas	MM, VD
Desarrollo del código	MM, VD
Participación en el vídeo	MM, VD