1. **Contexto**. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

   En la presente práctica, se ha elegido como caso de estudio las estadísticas proporcionadas por LaLiga que gestiona el campeonato nacional de primera división de España. Información que podría ser muy útil para evaluar el rendimiento de cada jugador a nivel de equipo como su desempeño durante la presente temporada 2021/22, asimismo información general del equipo como apoyo para la toma de decisiones al momento de generar estrategias para los próximos partidos, evaluar desempeños y alineaciones para posibles traspasos, transferencias de jugadores o cambios de estrategias.

   En este ejercicio, se ha elegido la web site https://www.laliga.com/, el cual facilita información de estadísticas y análisis de los equipos de las ligas como; LaLiga Santander, LaLiga SmartBank y el futbol femenino. En este caso se ha elegido la tabla de clasificación de LaLiga Santander 2021/22 para obtener información a tiempo real de los equipos, su posición en la tabla, y los siguientes indicadores; puntos (PTS), partidos jugados (PJ), partidos ganados (PG), partidos empatados (PE), partidos perdidos (PP), goles a favor (GF), goles en contra (GC) y la diferencia de goles (DG +/-).

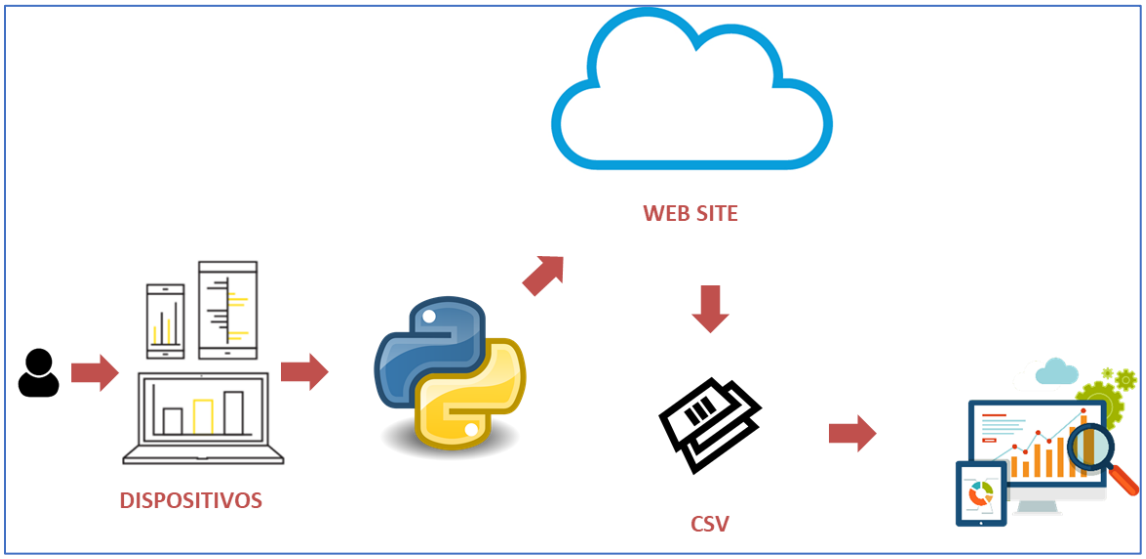
2. **Título**. Definir un título que sea descriptivo para el dataset.

   Tabla de clasificación de LaLiga Santander 2021/22.

3. **Descripción del dataset**. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

   El conjunto de datos (dataset) contiene la tabla de clasificación de LaLiga Santander 2021/22, qué se actualiza según la jordana y/o calendario programado, que lo normal es semanal, pero puede darse el caso de partidos postergados por programación de los equipos en otras competiciones como; copa del rey o champions. El fichero generado utiliza la extensión CSV (comma separated values) que es la más utilizada en la actualidad por su facilidad para interactuar con otros sistemas de visualización.

4. **Representación gráfica**. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



5. **Contenido**. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

   En este conjunto de datos (dataset) se presentan los siguientes campos e indicadores:

   - **Equipo**.- Nombre del equipo de futbol de primera división.
   - **PTS**.- Indicador de puntos del equipo.
   - **PJ**.- Indicador de partidos jugados.
   - **PG**.- Indicador de partidos ganados.
   - **PE**.- Indicador de partidos empatados.
   - **PP**. – Indicador de partidos perdidos.
   - **GF**.- Indicador de goles a favor.
   - **GC**. – Indicador de goles en contra.
   - **DG**.- Indicador de diferencia de goles.

La información fue extradía mediante la técnica de web scraping con el lenguaje de programación Python sobre la web site de LaLiga. Para que se guarden los datos extraídos en un fichero CSV.

6. **Agradecimientos**. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

   El propietario de la web site es LaLiga, al ser una información de acceso púbico no se requirió de un permiso específico, ni hubo la necesidad de programar un *bot* para su ejecución, al ser utilizado como un caso de estudio para el desarrollo de la práctica. A continuación, se presentan alguno casos similares y curso que apoyaron al desarrollo de la práctica:

   - https://www.youtube.com/watch?v=rhnMvvmfBFI
   - https://www.youtube.com/watch?v=XVv6mJpFOb0
   - https://www.youtube.com/watch?v=chPhlsHoEPo

   Para profundizar el análisis se podría extraer mayor información estadístico e integrarlo en una herramienta de visualización como Tableau o Power BI.

7. **Inspiración**. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

   Es cierto, que informaciones cómo estas se pueden visualizar en el buscador de Google y en diversas web sites, pero la motivación va enfocado en poder integrarlos con otros sistemas que apoyan a los tomares de decisiones tanto para el club (directivos), cuerpo médico, cuerpo técnico entre otros. Y por la pasión que despierta en mí el deporte rey, qué a pesar de ya no poder practicarlo más de 1 año (por estar a la espera de una cirugía

de ligamentos cruzados y un menisco), me motiva el poder apoyar con el tratamiento de información útil.

8. **Licencia**. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:
   - 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.
   - Other (specified above).
   - Unknown License.

   Una posible licencia para este conjunto de datos puede ser; Released Under CC BY-SA 4.0 License. La elección se basa en la idoneidad de las cláusulas que en ella se presentan en relación con el trabajo realizado en donde:

   - Se provee el nombre del creador del conjunto de datos generado y se indican los cambios realizados sobre este. De esta manera, se reconoce el trabajo de terceros y en qué medida se realizaron aportaciones con respecto al trabajo original.
   - Se permite su uso comercial, lo cual incrementa las posibilidades de que empresas puedan interesarse en los datos generados, permitiendo así, la realización de nuevos proyectos que reporten un reconocimiento al autor original.
   - Las nuevas contribuciones deben ser publicadas bajo la misma licencia, lo que permite que se le reconozca al autor original en todo momento y bajo los mismo términos que fueron planteados por él.

9. **Código**. Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

La práctica fue desarrollado con lenguaje de programación Python bajo la técnica de web scraping con la instalación de las siguientes librerías; BeautifulSoup, Requests y Pandas.

El código fuente se encuentra dentro de la carpeta Git: https://github.com/MMontalvoN/Practica1WebScraping

10. **Dataset**. Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

    El DOI al dataset en formato CSV en: https://github.com/MMontalvoN/Practica1WebScraping/WebScraping/CSV/

Contribuciones	Firma
Investigación previa	Montalvo Navidad, Miguel Ángel
Redacción de las respuestas	Montalvo Navidad, Miguel Ángel
Desarrollo del código	Montalvo Navidad, Miguel Ángel

(*) Si existe algún impedimento para publicar el dataset real, se deberá justificar esta situación y realizar y publicar en Zenodo un dataset simulado. En este caso, el dataset real se comunicará al profesor de forma privada (p.ej., enlace de Google Drive).

CÓDIGO:

```
# Paso 01: Instalación e importación de las librerías
from bs4 import BeautifulSoup
import requests
import pandas as pd

# Paso 02: Configuración de la web site para el scraping
url = 'https://www.laliga.com/laliga-santander/clasificacion'
page = requests.get(url)
soup = BeautifulSoup(page.content, 'html.parser')

# Paso 03: Extracción de los equipos y posición en la clasificación general
eq = soup.find_all('div', class_ = 'styled__ShieldContainer-lo8ov8-0 bkblFd shield-desktop')
equipos = list()
count = 0
equipos.append('EQUIPO')
for i in eq:
    if count < 20:
        equipos.append(i.text)
    else:
        break
    count += 1

# Paso 04: Extracción de los indicadores de clasificación
pts = soup.find_all('div', class_= 'styled__Td-e89col-10 gETuZs')
puntos = list()
count = 1

for i in pts:
    if count < 169:
        puntos.append(i.text)
    else:
         break
    count += 1

fila = 21
col = 8
M = [puntos[col*i: col*(i+1)] for i in range(fila)]

# Paso 05: Preparación y almacenamiento de los conjuntos de datos (dataset)
## Posición y equipos
de = pd.DataFrame({'Nombre':equipos}, index=list(range(1,22)))

## Indicadores de clasificación
dp = pd.DataFrame(data=M, index=pd.RangeIndex(range(1, 22)), columns=pd.RangeIndex(range(1, 9)))

## Unificación de los dataframes
frames = [de, dp]
result = pd.concat(frames, axis=1, join='inner')

# Paso 06: Exportación del data frames a un cojuntos de datos CSV
result.to_csv('WebScraping_Cladificacion.csv', index=False, header=None)
```