PRÁCTICA 1: TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

NOMBRE: DIEGO REFOYO MATELLÁN

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

En este proyecto, se ha decidido investigar sobre la clasificación de los mejores equipos de las grandes Ligas de fútbol europeo en sus respectivos países. En los últimos años, se ha abierto el debate de crear una "superliga" de fútbol con los mejores equipos europeos. En ella competirían los grandes clubes históricos como el Real Madrid, España, el Manchester United, Inglaterra, así como otros clubes importantes, aunque de más moderna afiliación como el PSG, Francia.

El propósito del proyecto es recoger información acerca de las estadísticas de estos equipos en sus ligas para poder crear una clasificación virtual que simule la superliga. Para ello, se ha obtenido la información del periódico deportivo 'AS' (https://as.com) y se ha guardado en un fichero con formato csv, en el que se recogen distintos puntos de interés como los puntos, los goles marcados o la posesión de cada equipo.

Los equipos seleccionados para esta clasificación virtual, son los cuatro primeros de las ligas de España, Alemania, Inglaterra, Italia y Francia. Si bien es cierto que no todos estos equipos serían parte de la llamada "Superliga", para este trabajo se ha decido escoger aquellos que en el momento de realizar la extracción de los datos ocupen el TOP 4 de cada una de las ligas mencionadas anteriormente.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

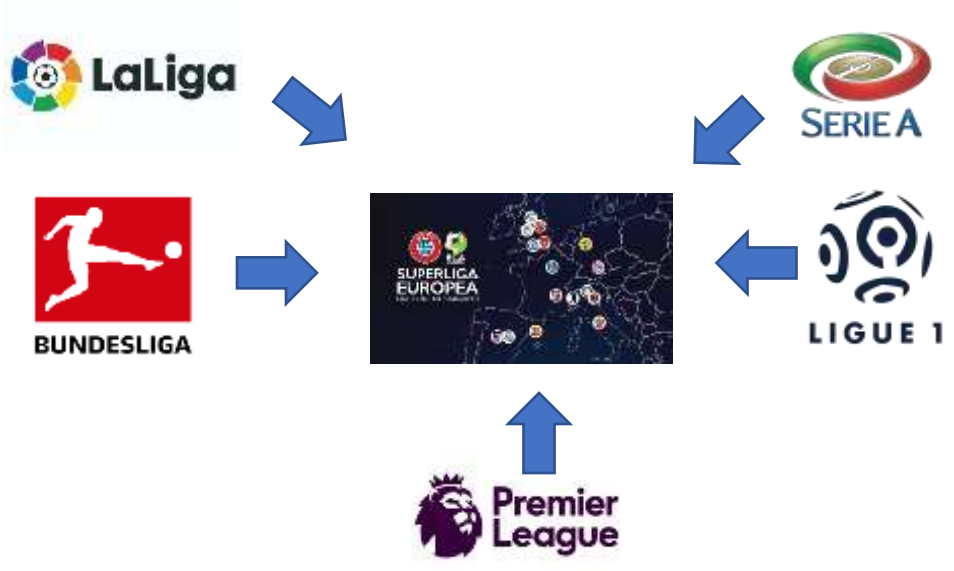
Clasificación en la Superliga de los equipos TOP4 de las grandes ligas de Europa.

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).

Tal como expresa el título dado en el apartado anterior, el dataset está basado en datos de los equipos situados en el TOP 4 de las cinco grandes Ligas Europeas. Este dataset recoge información real en el momento de la extracción y, por lo tanto, puede variar a lo largo del tiempo ya que las ligas de fútbol aún no han terminado. Sin embargo, por lo demás, el dataset se encuentra preparado para realizar un análisis directo, aunque siempre sería recomendable realizar un pequeño preprocesado para saber si es necesaria alguna tarea de limpieza previa. La descripción de las características extraídas es descrita en las siguientes

preguntas. El formato del dataset es un fichero CSV ya que facilita tanto su visualización como su tratamiento.

4. Representación gráfica. Presentar 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 dataset, se presentan dichos datos para los top4 equipos de cada una de las grandes cinco ligas de Europa. Las características extraídas son las siguientes:

A. Posicion: recoge el puesto de cada equipo en su respectiva liga.
B. Equipo: equipo al cual corresponde la información de la fila.
C. Puntos: puntos de cada equipo en su respectiva liga.
D. Pais: país al que pertenece cada equipo.
E. GolesMarcados: goles que ha marcado el equipo en su liga.
F. GolesRecibidos: goles que ha recibido el equipo en su liga.
G. Posesión %: posesión del equipo en su liga.

Los datos fueron recogidos mediante Web Scrapping en lenguaje de Python sobre diferentes páginas del periódico 'As' que contiene la información de la clasificación y de las estadísticas de cada equipo. Primero se recoge la información de los equipos que pertenecen al TOP 4 de cada liga y después las estadísticas de goles y posesión. Sobre cada página, se aplica el scraping para recolectar información. Finalmente se guardan 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.

El propietario del sitio web es el periódico deportivo 'AS' (https://as.com/?omnaut=1), del cual se han obtenido las estadísticas y análisis de cada uno de los equipos incluidos en la investigación.

Se ha buscado información acerca de la política de la página web sobre el web scrapping, pero no se ha encontrado ninguna referencia que no permita o en la que pidan que no se haga. Además, a diferencia de otros periódicos deportivos como 'Marca', no ocultan ninguna parte de su código web. Por lo tanto, se ha considerado adecuado obtener la información para el proyecto de esta página web, sobre todo teniendo en cuenta que la finalidad de este trabajo es meramente académico y no busca lucrarse en ningún momento.

7. Inspiración. Explique 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.

El interés de analizar este conjunto de datos nace de los gustos personales del alumno. Debido a que mi vida está muy ligada al mundo del fútbol al ser árbitro de este deporte, consideré un tema muy interesante simular una clasificación virtual de lo que podría ser la "Superliga". Salvando las diferencias acerca de los equipos integrantes de la misma, sirve para abrir el debate de si, finalmente se creará esta competición, la clasificación guardaría similitud con la presentada en este proyecto.

Por lo tanto, las preguntas que se desean responder son las siguientes:

- ¿Qué equipo sería campeón de la "Superliga"?
- ¿Habría algún país del cual destacaran sus equipos por encima del resto?
- ¿Influyen estadísticas como la posesión o los goles marcados en la posición de los equipos?

8. Licencia. Seleccione una de estas licencias para su dataset y explique 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 CC BY-SA 4.0 License. La elección se basa en la idoneidad de las cláusulas que tiene en relación con el proyecto realizado, en el cual se tiene en cuenta que:

- Se provee el nombre del periódico del cual se obtiene la información y se comentan los cambios realizados y cómo se ha adaptado al dataset final. De esta forma, se da reconocimientos a terceros, señalando cuál fue su aportación con respecto al trabajo original.
- El autor original se puede reconocer en todo momento y bajo los mismos términos que fueron planteados por él ya que las nuevas contribuciones han de ser publicados bajo la misma licencia.

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

El código fue hecho en lenguaje Python con la implementación de la librería BeautifulSoup. El código fuente se encuentra dentro de la carpeta completa de Git.

(https://github.com/DiegoRM8/SuperligaScraper)

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

El dataset se encuentra publicado en Zenodo en el siguiente link https://zenodo.org/record/4679318#.YHLJPugzbIU). El DOI obtenido es: 10.5281/zenodo.4679318.