Master en ciencia de datos
Universitat Oberta de Catalunya

Tipología y ciclo de vida de los datos
Práctica 1

Mateo Rodríguez Lavado

Eduard Conesa Guerrero

Contenido

1.- Contexto

El fútbol es un deporte que mueve a millones de personas, sobre todo en Europa y América. Entre las muchas ligas de fútbol existentes, la liga española de fútbol, no solo se sigue en España, sino que también es vista en otras partes ya que tiene a varios de los mejores equipos de Europa. Por ellos, los datos referentes a esta liga, son valiosos y, a su vez, consultados con frecuencia por la cantidad de dinero que mueve.

Si se quisiera realizar un análisis de los resultados de los partidos y las tablas de clasificación, que mejor sitio que la propia página de la liga española de fútbol donde sabemos que la información va a ser 100% fiable.

Por otra parte, cada vez más se está implantando la disciplina del Sports Analytics, que consiste en la aplicación de la ciencia de datos en el mundo del deporte. Juntamente con dicha disciplina y la implantación de una cultura de decisiones basadas en los datos, los clubes deportivos requieren de una mejor gestión de cualquier tipo de dato relacionado con los resultados de los encuentros, entrenamiento físico de los jugadores… Asimismo, un inicio para realizar un análisis muy simple del rendimiento o seguimiento del resto de los equipos puede ser el de obtener los datos de manera automatizada de la web de la liga.

2.- Título

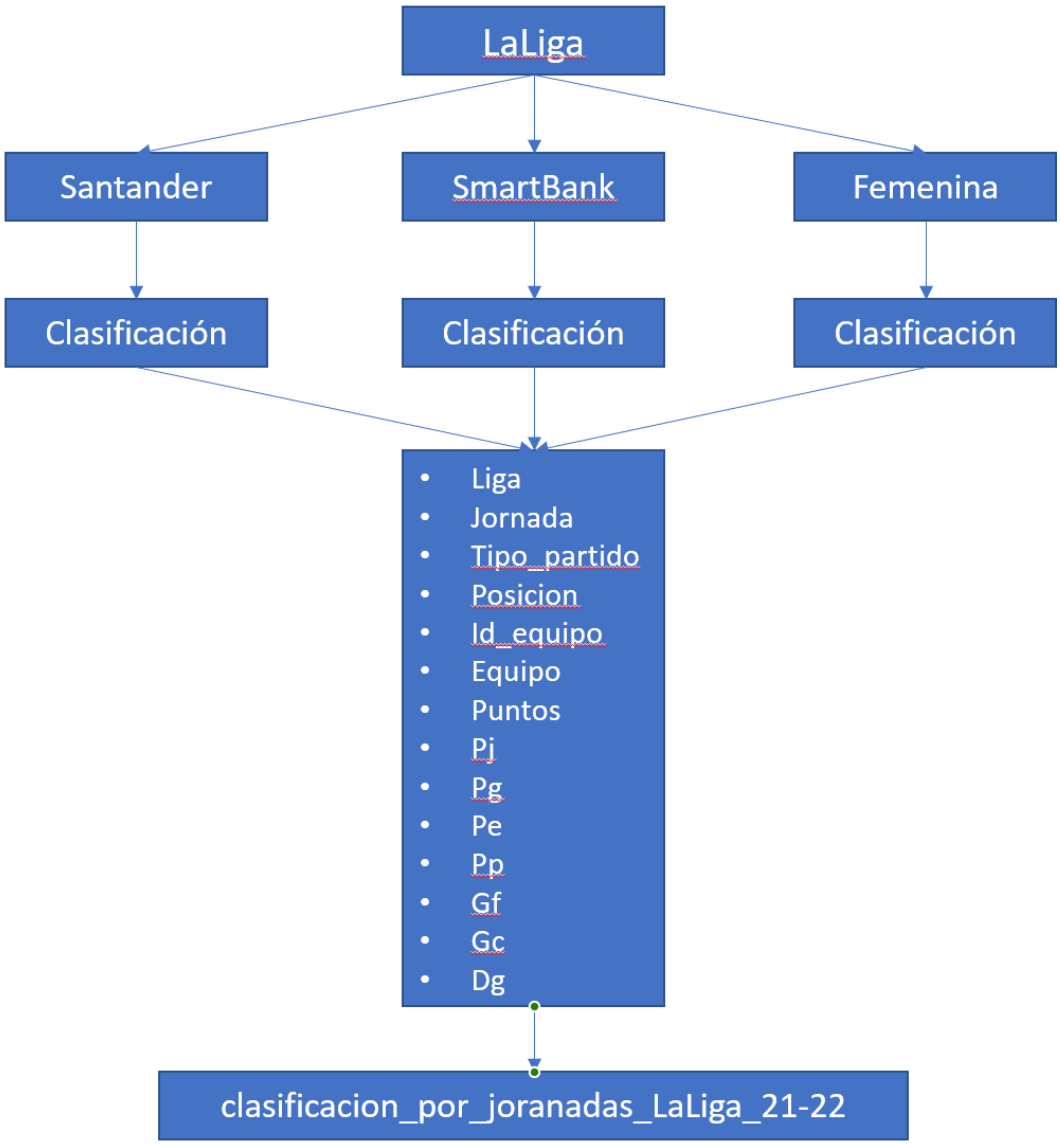
El título del dataset es clasificacion_por_joranadas_LaLiga_21-22 y está en formato csv.

3.- Descripción del dataset

Nuestro dataset contiene la información relativa a las tablas de clasificación de las diferentes ligas de fútbol español jornada a jornada. Es decir, por cada liga y jornada dentro de esa liga, obtenemos la posición, puntos, partidos ganados, perdidos y empatados; goles a favor y en contra; y la diferencia de goles. También se puede filtrar el dataset para mirar la clasificación según partidos en casa, partidos fuera de casa y, obviamente, partidos totales (fuera y en casa).

En total obtenemos la información de tres ligas españolas de fútbol: primera división masculina, segunda división masculina y primera división femenina. A medida que avanza el año, se irán ampliando el número de jornadas de las que podemos recolectar la información hasta llegar a la jornada 38, última jornada de la liga donde se podrá ver quién es el campeón de la misma.

4.- Representación gráfica



5.- Contenido

Los datos recogen la clasificación de las diferentes ligas de fútbol español en la temporada 2021/22, desde la jornada inicial hasta la más actual del momento de ejecución. Los campos que contiene nuestro dataset son:

- liga (String): indica la liga a la que pertenecen los datos (LaLiga Santander, LaLiga Smartbank o Primera División Femenina).
- jornada (String): señala la clasificación de una liga según la jornada en la que se encuentre.
- tipo_partido (String): Indica si la clasificación es contado solo los partidos en casa, contando solo los partidos fuera de casa o contando todos los partidos.
- posición (Int): indica la posición dentro de la tabla de clasificación.

- id_equipo (String): Identificador de cada equipo en 3 letras.
- equipo (String): nombre completo del equipo.
- puntos (Int): puntos que lleva un equipo en la jornada seleccionada.
- pj (Int): partidos que ha jugado un equipo.
- pg (Int): partidos que ha ganado un equipo.
- pe (Int): partidos que ha empatado un equipo.
- pp (Int): partidos que ha perdido un equipo.
- gf (Int): goles a favor, es decir, goles marcados que lleva un equipo hasta el momento.
- gc (Int): goles en contra, es decir, goles que le han marcado a un equipo hasta el momento.
- dg (String): diferencia de goles calculada como los goles a favor menos los goles en contra.

6.- Agradecimientos

Como antes hemos mencionado, este conjunto de datos se ha recogido de la página web de LaLiga. LaLiga es una asociación deportiva privada en la que se integran los clubes de las diferentes ligas de fútbol y es la responsable de la organización de las competiciones futbolísticas de carácter profesional y ámbito nacional en España.

Estos mismos datos pueden estar recogidos en otros lugares como periódicos deportivos, que lo utilizan para poder ofrecer información y análisis deportivos de la situación actual de la liga. Un ejemplo de ello podría ser el periódico Marca (https://www.marca.com/futbol/primera-division/clasificacion.html)

También pueden ser utilizados por casas de apuestas deportivas para poner precios a las apuestas que se hagan sobre posibles resultados.

Para garantizar la legalidad del proyecto y seguir los principios éticos de este proyecto, lo primero que se ha hecho ha sido investigar el robots.txt de la página de la liga y comprobando que, efectivamente, se puede realizar web scraping sobre las páginas que deseamos. A continuación, expongo una parte del robots.txt para que se vea:

########### Allows Robots all user agents ############

Allow: /

Allow: /laliga-santander/*

Allow: /laliga-smartbank/*

Allow: /futbol-femenino/*

Allow: /laliga-genuine-santander/*

Allow: /laliga/*

Tiempo en segundos entre dos solicitudes

Crawl-delay: 30

Sitemap: https://www.laliga.com/sitemap.xml

Podemos comprobar que las páginas de /laliga-santander/*, /laliga-smartbank/*, /futbol-femenino/* y /laliga/* son accesibles para cualquier bot y son precisamente estás las que hemos utilizado. Además, se puede observar un crawl delay entre peticiones de 30 segundos algo que se ha respetado añadiendo un espaciado de 30 segundos entre peticiones.

7.- Inspiración

La predicción y análisis de la evolución de los encuentros deportivos es interesante tanto para los fans, casas de apuestas así como para los propios clubes de fútbol.

Este conjunto de datos obtenido permite:

- Obtener un análisis de la evolución de los clubes deportivos.
- Realizar un análisis de los puntos fuertes y débiles de de los equipos, comparando los goles anotados y los encajados.
- Analizar qué encuentros han sido más críticos con los equipos adversarios.
- En caso de disponer otros datos de estrategias, analizar la efectividad de estos, integrando los datos con las fuentes propias del club y el entrenador.
- Predecir resultados de los encuentros deportivos.
- Analizar el efecto del apoyo del público en el equipo, analizando los partidos jugados en casa y como visitante.
- Obtención de ranking de club con más goles anotados.
- Obtención de ranking de club con menos goles encajados.

8.- Licencia

La licencia utilizada para el dataset generado será CC BY-NC-SA 4.0, ya que según el apartado de información legal de esta sobre los derechos de propiedad intelectual e industrial, se reconoce la titularidad de los datos a LaLiga y se detalla que “El Usuario, única y exclusivamente, puede utilizar el material que aparezca en este sitio Entorno LaLiga para su uso personal y privado, quedando prohibido su uso con fines comerciales o para incurrir en actividades ilícitas”.

Asimismo, LaLiga expone que cualquier efecto provocado por la modificación de los datos constituye una infracción, por lo que los datos deben reproducir el mismo contenido, ya que por ejemplo se podrían falsear los resultados, afectando a otros usuarios que utilizarán estos datos.

9.- Código

El código se encuentra en un repositorio de GitHub. El enlace es el siguiente:

https://github.com/mlavador/Practica1TD

10.- Dataset

10.5281/zenodo.5652052

Firmas

Contribuciones	Firma
Investigación previa	ECG, MRL
Redacción de las respuestas	ECG, MRL
Desarrollo del código	ECG, MRL

Referencias

LaLiga. Obtenido de https://www.laliga.com/