Tipología y Ciclo de Vida de Datos

Web Scraping La Liga 19/20

Práctica 1

Alfonso Jiménez Hernández y Jorge Martin de la Calle
1-4-2020

Contexto

En este proyecto se ha decidido investigar sobre los datos que se generan a través del fútbol concretamente de la Liga Española de Fútbol profesional. Actualmente, con la aparición del Big Data se ha empezado a recopilar información para ser más competitivo, comprar jugadores más barato y explotar su rendimiento a un coste menor. Unido a esto encontramos un ejemplo visual que se observa en la película Moneyball. Esta película trata de un gerente general de un equipo de beisbol que con ayuda de un economista usa las estadísticas para fichar a jugadores para lograr un equipo competitivo.

Viendo cómo se encuentra el panorama actual y hacia donde van los equipos de futbol en el ámbito de los datos, se quiere observar basándonos en estadísticas recogidas durante este año si se pueden sacar alguna conclusión interesante.

Para este trabajo se ha elegido la página https://www.sofascore.com/es/ que recoge resultados de futbol de todo el mundo, además de clasificaciones y estadísticas. Recoge la información de una manera clara y precisa.

Título para el dataset

El título para el dataset es **Stats_Players_LaLiga_19_20** ya que con estas tres palabras describimos lo suficiente para saber los registros que contiene y además de la brevedad del título.

Descripción del dataset

Tal y como se expresa en el nombre del dataset, esta basado en la recopilación de las estadísticas más importantes que se tienen acerca de todos los futbolistas de la liga española de futbol del año actual 19/20. En el dataset se representa cada futbolista como registro único. Las unidades de las diferentes características se comentan en el apartado inferior según el caso.

Los datos, al ser extraídos de un proceso de web Scraping no han pasado un proceso de preprocesado o limpieza y puede existir valores que no se correspondan a sus atributos. También pueden venir valores no informados con el valor None , ‘?’ o vacío.

Descripción gráfica

[En construcción]

[Partido->Estadísticas->Análisis-> Beneficio]

Contenido

En el dataset se presentan las estadísticas más importantes de los futbolistas de primera divisón de la liga de futbol.

Name,Team,Nac.,Age,Height,Preffoot,Position,Number,ValuePlayer,matchesTotal

1. **Name:** Nombre del futbolista analizado
2. **Team:** Equipo del futbolista analizado
3. **Nac.:** Nacionalidad del futbolista
4. **Age:** Edad del futbolista cuando se realiza el Web Scraping
5. **Height:** Estatura del futbolista
6. **Preffoot:** Pie dominante del futbolista, este valor puede ser Left, Right o Both
7. **Position:** Posición que ocupa en el terreno de juego, se define con una letra, G,D,M o F.
8. **Number:** Número que lleva el futbolista en su camiseta
9. **ValuePlayer:** Valor del futbolista actualmente, puede estar expresado en Millones o Miles de euros,
10. **matchesTotal:** Numero de partidos disputados por el futbolista hasta la fecha

[….]

Los datos fueron extraídos utilizando Python sobre diferentes páginas que se encuentran en Sofascore, primero se extrajo de la página de cada equipo las diferentes páginas individuales de los futbolistas de manera automática por equipo. De la página individual del futbolista obtenemos la información de los atributos dos al nueve. Posteriormente, como hemos extraído el id del futbolista de la página del equipo al que pertenece, recorremos el JSON individual que la página tiene por futbolista para recoger las estadísticas de manera automática.

[En construcción]