

PRIMERA PRACTICA:

Autores: Francisco Lora Hernanz; M. David Miralles Nieto

Fecha: 2023-03-17

INFORMACIÓN GENERAL

A continuación, se muestran los pasos a seguir para realizar la primera práctica. Para ello empecemos dando **contexto a nuestro problema** y fijando un **objetivo** o *target*.

Problema:

Trabajamos en una empresa dedicada a las apuestas deportivas. Y se está barajando incorporar la F1 dentro de dichas de apuestas.

Se necesita poner unas cuotas iniciales y para ello se necesita hacer un estudio a grosso modo que comparé la puntuación del mundial de pilotos entre los años 2021 y 2022 para ver si ha mejorado en dichos años.

Nuestra empresa nos da 3 datasets para poder ejecutar dicho estudio.

Datasets:

Nos encontraremos dentro de la carpeta DataF1CSV tres CSV. Veámoslos:

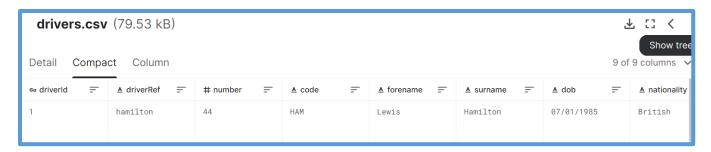
 races.csv: Es un dataset donde se encuentran todas carreras desde el 1950 hasta el 2023 donde el id(único) es el llamado raceid.

Visualmente: (En nuestro programa no se va a ver así de bonito)



drivers.csv: Es un dataset donde se encuentran todos los pilotos desde el 1950 hasta el 2023 donde el id(único) es el llamado driverid.

Visualmente: (En nuestro programa no se va a ver así de bonito)



• **results.csv**: Es un dataset donde se encuentran todos los resultados desde el 1950 hasta el 2023 donde el id(único) es el llamado **resultid**.

Visualmente: (En nuestro programa no se va a ver así de bonito)



Objetivo:

Necesitamos al final una tabla que refleje los datos de la siguiente manera:

driverid	driverRef	2021	2022	Mejora
1	Hamilton	236	221	bad
3	Verstappen	231	245	good
5	Alonso	145	180	good



¿Cómo se hace? (This is the way)

- 1. Importamos los 3 csv dentro de nuestra aplicación Spark. A los que llamaremos
 - CarrerasDF,
 - PilotosDF,
 - ResultadosDF,
- Dentro de CarrerasDF tenemos que crear dos nuevos dataframes (Carreras2021DF, Carreras2022DF) donde solo nos quedamos con las columnas (raceid, year) para cada uno de los dataframes, esto es, eliminar las columnas sobrantes. Para hallar dichos dataframes debemos filtrar por año, esto es, year==2021 (resp. year==2022).
- 3. Dentro de **PilotosDF** tenemos que quedarnos solo con los dos campos (**driverid**, **driverRef**), esto es, eliminar las columnas sobrantes.
- 4. Dentro de ResultadosDF vamos a:
 - Quedarnos con el las columnas (raceid, driverid, points).
 - Hacer un join (debéis ver de qué tipo es, inner, left, right) mediante el raceid entre el
 dataframe Carreras2021DF y ResultadosDF para conseguir un dataframe que nos de
 los resultados de la sesion 2021. A este dataframe lo llamaremos Resultados2021DF
 - Hacer un join (debéis ver de qué tipo es, inner, left, right) mediante el raceid entre el
 dataframe Carreras2022DF y ResultadosDF para conseguir un dataframe que nos de
 los resultados de la sesion 2022. A este dataframe lo llamaremos Resultados2022DF
 - Agrupar dichos dataframes por piloto (driverid) sumando los points para obtener la suma de la temporada.
 - Hacer un join (debéis ver de qué tipo es, inner, left, right) entre Resultados2021DF y
 Resultados2021DF mediante el driverid. Lo llamaremos Resultados2021vs2022DF.
 - En Resultados2021vs2022DF añadiremos una columna a la que llamaremos diferencia, que es a diferencia de 2022 2021. También crearemos otra columna que llamaremos Mejora y, que se calcurará según si el valor de diferencia es positvo o negativo podrá good o bad.
 - Por último haremos un join según el *driverid* entre Resultados2021vs2022DF y
 PilotosDF. A dicho dataframe lo llamaremos ResultadoFinalDF
- 5. Por último guardaremos el ResultadoFinalDF en formato CSV.

¡Esto es todo! Ya lo has conseguido 🚱

