# A1 - Preproceso de datos

Enunciado

Semestre 2020.2

## Índice

# Introducción

En esta actividad realizaremos el preprocesado de un fichero de datos que contiene el estilo de juego del videojuego de consola Fifa 2017, así como estadísticas reales de los jugadores de futbol. El conjunto de datos contiene más de 17500 registros y 53 variables.

Las principales variables que se usarán en esta actividad son:

- Name (Nombre del jugador)
- Nationality (Nacionalidad del jugador)
- Club_Joining (Fecha en la que empezó en el club)
- Contract_Expire (Año finalización del contrato)
- Rating (Valoración global del jugador, entre 0 y 100)
- Height (Altura)
- Weight (Peso)
- Preffered_Foot (Pie preferido)
- Birth_Date (Fecha de nacimiento)
- Age (Edad)
- Work_Rate (valoración cualitativa en términos de ataque-defensa)

La descripción de los atributos se puede consultar en https://www.fifplay.com/encyclopedia. La descripción de las abreviaturas de la posición del jugador en el campo se puede consultar en https://www.dtgre.com/2016/10/fifa-17-position-abbreviations-acronyms.html.

El objetivo de esta actividad es preparar el fichero para su posterior análisis. Para ello, se examinará el fichero para detectar y corregir posibles errores, inconsistencias y valores perdidos. Además, se presentará una breve estadística descriptiva y se hará un análisis de componentes principales (ACP) con algunas variables cuantitativas.

La actividad consta de tres partes diferenciades:

- En la primera parte (secciones 2, 3, 4 y 5), se realiza verificaciones y normalización de algunes variables, siguiendo los criterios que se especifican más adelante.
- En la segunda parte (secciones 6 y 7), se tratan los valores atípicos y los valores perdidos.
- En la tercera parte (secciones 8 y 9), se calculan algunas métricas de tendencia central y dispersión, que seria el primer paso del análisis descriptivo y, por último, se realiza un análisis de componentes principales (ACP) con algunas variables cuantitativas.

**Criterios de verificación y de normalización de las variables**:

A continuación se muestran los criterios con los que deben limpiarse los datos del conjunto:

1. En los datos numéricos, el símbolo de separador decimal es el punto y no la coma.
2. Verificar si hay registros duplicados con el valor ID. En caso de duplicación, seleccionar el registro con menor numero de NAs en las variables.
3. Las variables Name y Nationality no han de tener espacios en blanco antes o después de su valor. El valor para estas variables han de ser mayúsculas en la primera letra de cada palabra, tal como "Lionel Messi".
4. La variable Height se ha de expresar en cm con 3 dígitos sin decimales. Para facilitar la lectura y tratamiento del fichero deben ser numéricas, por lo tanto se debe quitar el símbolo de cm.

5. La variable Weight se ha de expresar en kg con 2 dígitos sin decimales. Si hay decimales, se ha de *truncar* el valor. Para facilitar la lectura y tratamiento del fichero deben ser numéricas, por lo tanto se debe quitar el símbolo de kg.
6. Verificar que la variable Club_Joining está en el rango de los años 1990 a 2017. En caso de haber algún registro que no compla la condición, indicar el número de registro, Name y Club_joining.
7. Verificar que el año de expiración del contracto (Contract_Expiry) no es inferior al año de inicio del contracto(Club_Joining). En caso de haber algún registro que no cumple la condición, indicar el número de registro, Name, Club_Joining y Contract_Expiry.
8. Verificar que la edad (Age) en la fecha 1/1/2017 corresponde a la calculada con la fecha de nacimiento (Birth_Date). En caso de haber algún registro que no cumpla la condición, modificar la edad en función del valor obtenido con la fecha de nacimiento.
9. Verificar que la variable Rating esté entre 0 y 100.
10. La variable Preffered_Foot ha de tener los valores Left y Right que corresponde a los valores actuales de 1 y 2, respectivamente.
11. La variable Work_Rate se basa en la combinación de dos de estas tres categorías: Low, Medium y High. Verificar que se cumple y en caso contrario, hay que corregirlo. Puedes encontrar los nombres de las categorías cortado con tres letras.

Para realizar el preproceso del fichero, seguir los pasos que se indican a continuación.

**Aspectos importantes a tener en cuenta para entregar la actividad**:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos. Es decir, se debe usar la terminología propia de la estadística.
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. El documento debe estar bien estructurado y ser conciso.

# 1. Carga del archivo

Se debe abrir el archivo de datos y examinar el tipo de datos con los que R ha interpretado cada variable. Examinar también los valores resumen de cada tipo de variable.

# 2. Verificar duplicación de registros

# 3. Normalización de los datos cuantitativos

Inspeccionar los valores de los datos cuantitativos y realizar las normalizaciones oportunas siguiendo los criterios especificados anteriormente. Estas normalizaciones tienen como objetivo uniformizar los formatos. Si hay valores perdidos o valores extremos, se tratarán más adelante.

Al realizar estas normalizaciones, se debe demostrar que la normalización sobre cada variable ha dado el resultado esperado. Por lo tanto, se recomienda mostrar un fragmento del archivo de datos resultante. Para

evitar mostrar todo el conjunto de datos, se puede mostrar una parte del mismo, con las funciones **head** y/o **tail**.

Seguid el orden de los apartados.

### 3.1. Rating

### 3.2. Height

### 3.3. Weight

## 4. Normalización de los datos cualitativos

Inspeccionar los valores de los datos cualitativos y realizar las transformaciones oportunas, siguiendo los criterios especificados. Al igual que en el apartado anterior, mostrar el resultado sobre un fragmento del conjunto de datos.

Se debe seguir el orden especificado de los apartados.

### 4.1. Name y Nationality

### 4.2. Preffered_Foot

### 4.3. Work_Rate

## 5. Posibles inconsistencias y variables tipo fecha

Verificar si existen inconsistencias entre algunas variables. Por otra parte, algunas de las variables es necesario indicar que son de tipo fecha (en r, tipo `Date`) para luego hacer las transformaciones adecuadas. Observar que la configuración del tipo fecha es mes/día/año. Al igual que en el apartado anterior, muestre el resultado sobre un fragmento del conjunto de datos.

Se debe seguir el orden especificado de los apartados.

### 5.1. Club_Joining

### 5.2. Contract_Expiry >= Club_Joining?

### 5.3. Revisar si la edad corresponde a la fecha de nacimiento

## 6. Valores atípicos

Revisar si hay valores atípicos en la variable Height y Weight. Si es así, y se trata de un valor *anormalmente* alto o bajo, se recomienda sustituir el valor por "NA".

## 7. Imputación de valores

Buscar en las variables Weight y Height donde haya valores perdidos (NA) y realice una imputación de valores.

Para realizar una imputación de valores necesita hacer una regresión lineal que tenga como variable a predecir la variable con la NAs. Esto significa que hay que realizar dos modelos de regresión lineal, uno para predecir los valores NAs en Weight a partir de la variable Height. El otro es precisamente al revés, predecir los valores NAs en Height a partir de Weight.

La función para hacer regresión lineal es `lm()`.

Muestre el resultado de la imputación y de la variable explicativa en aquellos casos donde había un NA.

## 8. Estudio descriptivo de las variables cuantitativas.

Realice un breve estudio descriptivo de las variables cuantitativas una vez depuradas. Hay que crear una tabla con medidas de tendencia central y de dispersión, tanto robustas como no robustas. Haga un breve comentario sobre los resultados obtenidos entre estos tipos de medidas en todas las variables.

## 9. Análisis de Componentes Principales (ACP)

Realizar el Análisis de Componentes Principales sobre las variables "Rating", "Height", "Weight" y "Age". Representar el gráfico biplot de dos dimensiones. Hacer un breve comentario indicando el porcentaje de variabilidad explicada en cada componente principal, la variabilidad explicada en las dos primeras dimensiones y qué variable original está más asociada a cada una de las dos primeras componentes principales. Interpretar.

Por ultimo, ¿hay algún punto más fuera de la nube de puntos?, si es así puedes mostrar los valores de las variables originales e indicar que tiene de especial este punto.

Para responder a este apartado, puede usar la función `prcomp` y `ggbiplot`.

## 10. Archivo final

Una vez realizado el preprocesamiento sobre el archivo, copie el resultado de los datos en un archivo llamado "fifa_clean.csv".

## 11. Avaluación de la actividad

- Apartados 1, 2, 3 y 4 (30 %)
- Apartado 5 (10 %)
- Apartado 6 (10 %)
- Apartado 7 (20 %)
- Apartados 8 y 9 (20 %)
- Calidad del informe dinámico (calidad del código, formato y estructura del documento, concisión y precisión en las respuestas) (10 %)