**M5. Tarea**

Para la resolución de esta tarea, se recomienda partir de los diferentes casos de usos que se encuentran descritos en el manual. Partiremos del caso de uso introducido en el bloque de Procesamiento de Spark con Python.

**Se pide:**

1. Selección de un conjunto de datos con el que realizar los siguientes apartados. No se tendrá presente la calidad de los resultado obtenidos, sino el procedimiento hasta llegar a ellos y las conclusiones que se extraigan de ellos.
2. Preparación y descripción del conjunto de datos. Se valorará la extracción de estadísticos descriptivos, la aplicación de transformaciones y la preparación del conjunto de cara a la resolución de los siguientes apartados.
3. Implementación y análisis de algún modelo de clasificación desarrollados en la API de Spark API de Spark:

<https://spark.apache.org/docs/latest/ml-classification-regression.html#classification>

1. Implementación y análisis de algún modelo de regresión desarrollados en la API de Spark API de Spark:

<https://spark.apache.org/docs/latest/ml-classification-regression.html#regression>

1. Implementación y análisis de algún modelo de agrupación desarrollados en la API de Spark API de Spark:

<https://spark.apache.org/docs/latest/ml-clustering.html>

1. Uso de  y análisis de algún algoritmo de optimización de parámetros de alguno de los tres apartados anteriores, este conjunto de algoritmos se deben encontrar desarrollados en la API de Spark API de Spark:

<https://spark.apache.org/docs/latest/ml-tuning.html>

Recordemos que un posible conjunto de datos con el trabajar es el empleado en la Tarea (2) del Módulo Introducción al Big Data:

* Trabajamos con datos de la Liga 2017/2018 del portal

<http://www.football-data.co.uk/spainm.php>

* Los resultados que nos interesan están en el fichero CSV:

<http://www.football-data.co.uk/mmz4281/1718/SP1.csv>

* El significado de las columnas del fichero de datos está recogido en:

<http://www.football-data.co.uk/notes.txt>