UNIVERSIDADE DA CORUÑA

## Máster Universitario en Bioinformática para Ciencias de la Salud
## Inteligencia Computacional para datos de alta dimensionalidad

# Práctica Librerías

## 1. Introducción

MLlib es una librería básica de Spark que proporciona numerosas utilidades que resultan prácticas para las tareas de aprendizaje automático, como clasificación, regresión, clustering o extracción de características. Está formada por dos paquetes: la API original, construida sobre RDDs y una API de más alto nivel, construida sobre DataFrames para la construcción de pipelines. Básicamente, un DataFrame es un RDD que posee un esquema interno, es decir, los datos se almacenan como si se tratase de la tabla de una base de datos.

## 2. Conjunto de datos

Para realizar la experimentación utilizaremos el conjunto de datos 'Reef Life Survey (RLS): Global reef fish dataset', disponible en el repositorio AOC (Australian Ocean Data Network). Este conjunto de datos contiene registros de peces óseos y elasmobranquios recolectados por buceadores de Reef Life Survey (RLS) a lo largo de transectos de 50 metros en arrecifes de coral, localizados alrededor del mundo.

Podéis descargaros el conjunto de datos en formato .csv en el siguiente enlace:
https://portal.aodn.org.au/search?uuid=9c766140-9e72-4bfb-8f04-d51038355c59

## 3. Objetivos

La tarea consiste en comprobar si se puede utilizar el recuento por especies de peces de una zona para estimar su latitud y longitud. Para ello será necesario manipular los datos para ponerlos un formato adecuado para el algoritmo de aprendizaje que utilicemos, tras lo cual habrá que realizar el entrenamiento y validación de los modelos de regresión que, dado un recuento de peces, estimen el valor de latitud o longitud.

Deberán adquirirse o reforzarse estas destrezas:

- Manipulación de RDDs
- Transformación de RDD a DataFrame
- Trabajar con la librería MLlib de Spark
- Seguimiento de un esquema de trabajo de aprendizaje máquina adecuado.

### ¿Qué hay que hacer?

- Implementar el código que se indica en los TO-DO.
- Realizar un análisis preliminar de los datos, mostrando número de registros, surveys, localizaciones, especies y familias.
- Calcular la precisión del modelo de Regresión, en términos del error cuadrático medio, a la hora de predecir:
    - Longitud
    - Latitud

## 4. Indicaciones

Para calcular la predicción del modelo se utilizará el algoritmo de Regresión *Random forest regression,* disponible en la librería MLlib *(https://spark.apache.org/docs/latest/ml-classification-regression.html#random-forest-regression).*

(*) La clase *RegressionEvaluator* puede ser útil para evaluar el modelo de Regresión.

Se valorará positivamente el análisis con algún otro modelo de regresión.

## 5. Entrega

El ejercicio se realizará en grupos de dos personas (los establecidos en Moodle). En la entrega hay que incluir tanto el código de Spark como una pequeña memoria explicativa en la que se describa el código, una justificación del planteamiento de aprendizaje máquina y el estudio inicial de datos solicitado. Deberá incluir, además, una argumentación respecto a la validez de los modelos obtenidos para la tarea de predicción descrita.

La fecha límite de entrega de esta práctica es el **12 de Diciembre (23:55h)**