**Análisis de datos de plantaciones de olivar y modelado predictivo usando técnicas de machine learning**

*Miguel Ángel Travado Muñoyerro*

*Gualberto Asencio-Cortés*

Data Science and Big Data Lab, Pablo de Olavide University, 41013 Seville, Spain

*Proyecto realizado por alumno interno de 3er año de Ingeniería Informática de los Sistemas de Información*

**Resumen**

La posibilidad de recopilar datos de múltiples variables relacionadas con las plantaciones de olivar permite el estudio y el tratamiento de estos con motivo de mejorar el proceso y la toma de decisiones relacionadas con el tratamiento de la oliva. Mediante una serie de fases se realizará un entendimiento y modelado de los datos disponibles, utilizando múltiples estrategias de machine learning. La metodología seguida para las diferentes fases será la CRISP-DM.

**Introducción**

La agricultura y la informática son dos disciplinas diferentes, pero existe la posibilidad de que una complemente a la otra. La primera, necesita de eficiencia y previsión para que los resultados sean mejores. En cambio, la informática requiere datos reales para poner a prueba sus capacidades. Gracias a la recogida de múltiples variables independientes o relacionadas del sector agrícola, tenemos la oportunidad de crear un modelo para la predicción y toma de decisiones que ayude a conseguir la mejora de esos aspectos determinantes a la hora de organizar y gestionar plantaciones de olivos. Nuestra tarea como informáticos trata de predecir la cantidad de olivas picadas y moscas de la oliva aparecerán en el plazo de una semana.

**Comprensión de los datos**

Tenemos a nuestra disposición un conjunto de datasets que recogen información relevante de 16 plantaciones de olivas diferentes situadas en Andalucía. Cada archivo consta de multitud de atributos recogidos de las plantaciones, que se recolectan de manera semanal, es decir, cada fila hace referencia a una semana. Es importante destacar que los valores de estos atributos no se recogen en igualdad de condiciones, refiriéndose así a que es posible que unos se calculen un día de la semana a una hora en concreto y otros a otra distinta.

Para ser precisos, estos atributos tienen un formato específico que determina la condición y el tipo de dato recogido. Dicho formato se indica de tal manera: *XXXX\_LAST/AVG\_bk\_w-X-sXX*. A continuación, se aclara el significado de cada etiqueta.

* *XXXX* es la etiqueta del atributo recogido. Puede ser cualquier conjunto breve de letras o acrónimo. Más adelante se nombrarán los más importantes.
* *LAST/AVG* son significan último y media señalando si el dato es el último tomado o la media de los datos según el tamaño temporal del atributo.
* *w-X* lo entendemos como el intervalo de tiempo en el que se recoge un dato y el siguiente. Tal y como ya hemos mencionado, cada fila del dataset es una semana, entonces, en nuestro caso el formato para todos los atributos obedecerá el *w-1* ya que cada semana tenemos una instancia nueva.
* *s-XX* determina el plazo temporal en días del atributo. Este es diferente a *w-X* ya que no expresa el intervalo de tiempo entre una instancia y otra del dataset, sino que determina de hace cuanto tiempo es el dato recogido. Por ejemplo, si tenemos un *XXX\_AVG\_bk\_w-1\_s-90* quiere decir que el valor del atributo *XXXX* definirá la media tomada (*AVG*) de hace 90 días (*s-90*).