**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*).

Este sería el formato con el que se describen los diferentes atributos de nuestros datasets, que no explican el sentido ni significado de los atributos en sí; esto se desarrollará en el análisis exploratorio de datos más adelante.

**Preprocesado de datos**

La preparación de los datos es una fase esencial en cualquier modelado predictivo que se realice. Es necesario que los datos vayan acordes con la tecnología y estrategias que se usarán en el futuro. El paradigma que utilizaremos en el modelado, en principio, únicamente permite datos en formato numérico, en este caso números de coma flotante.

Existen varias tareas genéricas para preparar los datos disponibles. En primer lugar, la *selección del conjunto de datos* que se nos proporcionan y que tomaremos como veraces y de calidad, antes de realizar ciertas modificaciones necesarias para el modelado. Son 8 datasets que describen el comportamiento de los diferentes atributos, pero, 4 de ellos, la clase será la cantidad de olivas picadas y en los restantes serán la cantidad de moscas de la oliva que se recogen en una trampa de las plantaciones. Son valores representativos, no exactos.

La siguiente tarea consiste en una *limpieza de datos*, las modificaciones necesarias mencionadas anteriormente. Se da la posibilidad de aplicar multitud de técnicas diferentes para transformar los datos de tal manera que faciliten y mejoren el proceso siguiente de modelado predictivo. Principalmente, este proceso se lleva a cabo para corregir o eliminar registros inexactos en un conjunto de datos. De forma general, se identifica y sustituye los datos o registros incompletos, inexactos, corruptos o irrelevantes. Tras este proceso, los datos deben ser coherentes y estar libres de errores, algo esencial para la explotación de los datos.

Las técnicas utilizadas en el conjunto de datasets disponibles relacionadas con la limpieza de datos se enumeran y desarrollan a continuación.

En multitud de situaciones, obtenemos un conjunto de datos en los que no todas las instancias tienen valores para sus atributos. En estos casos, es imprescindible actuar para poder completar nuestro dataset y que no conste falta de datos en él. La implementación para el modelado no permite valores ausentes así que esta acción es indispensable. Para poder mantener el mayor número de instancias, procederemos a asignar un valor representativo que sustituirá los valores ausentes. Este valor es la media aritmética del atributo en cuestión, así no genera un gran impacto a la hora de analizar los datos. Si existe alguna columna que no tenga ni un solo valor, se eliminará el atributo ya que no aporta nada al modelado ni predicción.

Al igual que realizamos esta imputación, también imputamos cualquier atributo que no sea de tipo numérico, ya que no nos aportará ningún tipo de información en este caso.

Para poder tener unos atributos con valores numéricos dentro de un rango continuo [0, 1] aplicamos una operación matemática a cada columna de manera independiente, llamada homotecia y/o normalización. Para cada valor xi, su normalización xiN se define como:

Al igual que la operación de estandarización, esta transformación es muy recomendable si queremos aplicar modelos de aprendizaje automático, pues las diferencias de escalas y rangos de valores entre los atributos puede perjudicar el entrenamiento de modelos. Sklearn nos facilita procedimientos y funciones para realizar este cálculo.

Otro aspecto interesante que aplicar es la detección de outliers. Esto son valores anómalos que se presentan fuera de la distribución normal de datos. Nos referimos a instancias que difieren mucho en sus atributos a las demás instancias de nuestro conjunto de datos, lo que quiere decir que proporcionan información ‘desproporcionada’ ya que no sigue la tendencia normal, y puede inferir negativamente en la predicción resultante.

Nos centraremos en la aplicación del algoritmo de envolvente elíptica, que presupone que los datos siguen una distribución normal (distribución gaussiana). En nuestro caso, supondremos que el conjunto de datos relativo a las plantaciones sigue dicha distribución normal, aplicando de forma fiable el algoritmo propuesto.

Haremos uso de funciones que ofrece Sklearn para proceder con la eliminación de outliers. En primer lugar, entrenaremos un modelo basado en el dataset que queremos analizar, después calcularemos una puntuación sobre la normalidad de cada instancia respecto a la distribución centrada y escalada en la mediana e IQR, respectivamente. Estableceremos un umbral según el percentil que deseemos para obtener el valor límite que diferencia entre valor aceptable y outliers. Y, por último, eliminaremos aquellas instancia que su puntuación supere ese umbral.