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.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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 x_i, su normalización x_i^N se define como:

$$x_i^N = \left\{ \frac{x_{i,j} - \min(x_i)}{\max(x_i) - \min(x_i)}, \forall j \in 1, \dots, n \right\}$$

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 instancias que su puntuación supere ese umbral.

Modelado supervisado

Tras comprender y tratar los datos a disposición, es el momento de utilizarlos para un fin determinado. Como se explicó al inicio del análisis, el objetivo principal es el uso del conocimiento disponible para la inferencia de una variable esencial en el tratamiento de la oliva, la cantidad de oliva picada en un momento determinado, así la toma de decisiones es más acertada según en qué fase o situación se encuentre nuestra plantación.

Para poder aprender de los datos, debemos conocer que tipo de aprendizaje automático nos conviene.

- Aprendizaje supervisado. Este consiste principalmente en el entrenamiento de un modelo para poder reconocer patrones y predecir una variable determinada, llamada clase. Dentro de este aprendizaje existen dos tipos de problemas de inferencia según como se representa la variable de salida. De regresión si esta es numérica o de clasificación si nuestra clase es categórica.
- Aprendizaje no supervisado. Es la parte del aprendizaje automático que se ocupa de los problemas donde no hay una variable de especial interés, sino que busca patrones genéricos del conjunto de datos e intenta identificarlos.

Podemos determinar que es de mayor interés la aplicación de un modelado supervisado para predecir nuestra variable útil. Podemos definir dicha variable como numérica, de manera que el aprendizaje será de regresión.

El modelado es la fase central del proyecto, en el que entrenaremos y pondremos en práctica un modelo de machine learning que nos proporcionará resultado con los que podremos decidir en base a datos obtenidos.

Los algoritmos de aprendizaje automático deben crear un modelo de inferencia a partir del conjunto de datos que relacione los atributos de entrada con el de salida. Por ello, todas las instancias deben tener un valor determinado para cada atributo, tarea de la que nos hemos encargado en el apartado de preprocesamiento de datos rellenando los valores nulos de instancias con la media aritmética de dicho atributo. Para poder realizar el modelado de predicción, deberemos separar dos conjuntos diferenciados de los datos disponibles, que tendrán funcionalidades esenciales. El primer conjunto de datos es el de entrenamiento con el que entrenaremos al modelo para la inferencia, y, el segundo conjunto, será el de test que servirá para probar la eficacia del modelo entrenado. Los dos conjuntos no pueden compartir ninguna instancia, este aspecto es fundamental para evaluar correctamente el aprendizaje.

Llamamos validación al proceso mediante el cual se divide el conjunto de datos en subconjuntos de entrenamiento y test con el objetivo de evaluar de forma adecuada la bondad de los algoritmos de aprendizaje supervisado. La estrategia de validación que seguiremos en el análisis, aunque no sea la más eficaz, es suficiente para este proyecto, será la validación cruzada. Esta no conlleva el inconveniente principal de la validación hold–out, la cual es escasa en representatividad de los resultados al evaluar únicamente un conjunto de test. Esta estrategia elegida posee las características de generalidad y representatividad. Principalmente, dividiremos el conjunto total de datos en K subconjuntos, e iremos rotando, asignando a uno de estos la función de conjunto de test y los demás serán el conjunto de entrenamiento. La idea es predecir cada subconjunto de datos usando como entrenamiento el resto.

Evaluaremos la eficacia de nuestro modelo mediante métricas de evaluación. Las métricas difieren según sea un problema de regresión o clasificación. El error medio absoluto es una medida de interpretación fácil y directa de los resultados. Nótese que la métrica utilizada es un cálculo de valores absolutos.

La métrica RMSE se basa en las diferencias de valor absoluto entre las predicciones y valores reales. A diferencia que el MAE, que haya una mayor diferencia entre ambos valores provoca un mayor impacto en la medida.

En ocasiones, conocer una métrica de evaluación relativa como el MAPE, que representa la magnitud del error cometido del valor real al que se predice, en forma de proporción.

Por último, definir una métrica de evaluación para la regresión que se usa únicamente para evaluar modelos lineales. Este es el coeficiente de determinación, que recoge la cantidad de variabilidad de la clase que el modelo es capaz de predecir con respecto al total de variabilidad de la clase.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Explicada la validación y la evaluación de los algoritmos procederemos a explicar los dos que utilizaremos para nuestros modelados supervisados.

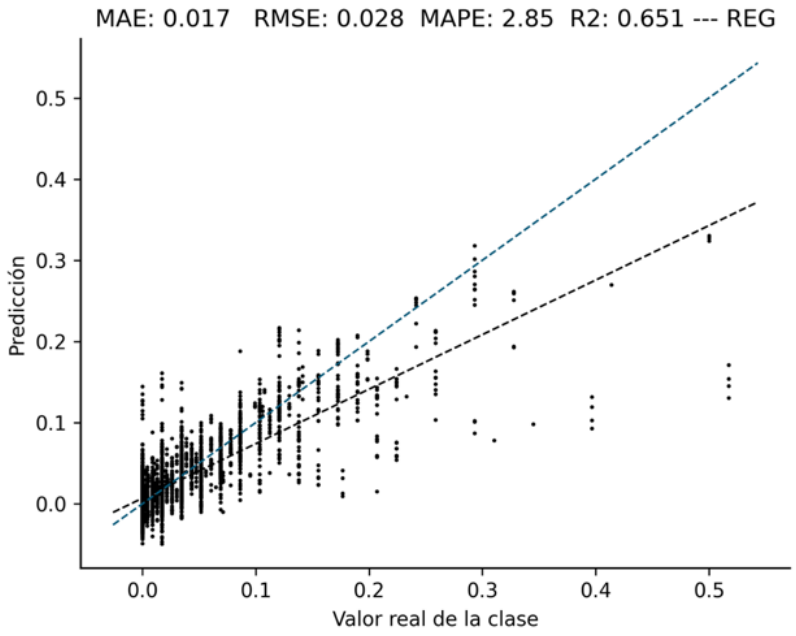
En primer lugar, tenemos la Regresión Lineal Múltiple, la cual pretende encontrar relaciones lineales entre los diferentes atributos y la clase. Definimos esta como múltiple porque se disponen de más de un atributo de entrada, si tuviésemos únicamente uno sería simple. Procederemos con OLS (*Ordinary Least Square*), que intenta calcular los coeficientes que multiplican los valores de los atributos que relacionan estos con el valor predicho de la clase de cada instancia.

$$\hat{y} = w_0 + w_1 \cdot x_1 + \cdots + w_p \cdot x_p$$

Esta función relaciona atributos, coeficientes y valor de la clase predicho, por lo que para el cálculo de los coeficientes debemos hacerlo minimizando la función de coste que relaciones coeficientes, valores de los atributos y valores de la clase real.

$$f(w_0, \dots, w_p) = \frac{1}{2n} \cdot \sum_{i=1}^{n}\left(\left(w_0 + \sum_{j=1}^{p} w_p \cdot x_{i,j}\right) - y_i\right)^2$$

Gracias a la librería de funciones que nos ofrece Sklearn podemos utilizar diferentes algoritmos ya programados para testear el modelo con la validación de nuestros datos sobre las plantaciones de olivar y evaluar su veracidad y eficacia en la inferencia. seguir el procedimiento de creación del objeto modelo de regresión lineal establecemos los parámetros necesarios para una correcta validación cruzada, los cuales son la división del conjunto, el algoritmo y los datos a tratar.



En la gráfica se muestran los valores predichos y los valores reales de la clase, acompañada de su propia recta de regresión y una recta de referencia que indica la idoneidad de inferencia, es decir, cuando los valores calculados son exactamente iguales que los valores reales (x = y).

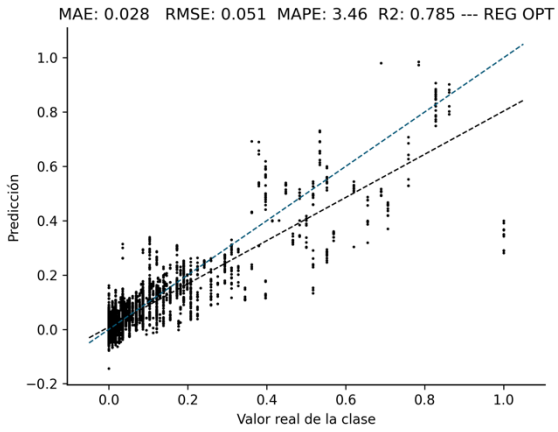
Dado que las métricas MAE y RMSE son absolutas, están en la misma unidad de magnitud que la clase, y representan de alguna manera la diferencia entre la inferencia y la realidad, por lo que cuando más cercanas a 0 estén, más acertada será la predicción. En este caso, los valores rondan entre 0.0 – 0.7 y las métricas muestran que la variabilidad no es alta, por lo que podemos fiarnos de los cálculos realizados. Gracias a la métrica porcentual, MAPE, podemos medir el error absoluto porcentual y tener una idea más clara de cuanto se alejan nuestras predicciones de la realidad en unidades relativas, es decir, en un tanto por ciento. Que este valor esté en torno a un 3% aclara, al igual que el MAPE, que nuestros errores son pequeños y no disparatados. Sin embargo, existen dos puntos a destacar aparte de las métricas:

- Nuestras predicciones son subestimaciones, puesto que de forma general toman valores menores a los reales.
- A valores reales más altos, la predicción disminuye en calidad y acierto.

La recta de regresión no acompaña en pendiente a la recta que determina una bondad perfecta. De aquí podemos deducir que la tendencia global de las predicciones no sigue a la tendencia de los valores reales y que como se ha comentado antes, a mayores valores reales, peor es la predicción. Esto viene dado por la cantidad de instancia que toman un valor relativamente bajo en vez de alto dentro del contexto del problema.

Por último, comentar el coeficiente de determinación que se ha obtenido con el algoritmo es aceptable. Podemos decir que un valor entre 0.6 – 0.75 define un modelo con una variabilidad decente respecto a la clase. Lo ideal sería tener un $R2 > 0.75$.

En un modelado que trata un gran conjunto de datos multidimensional, es esencial probar diferentes técnicas al igual que suprimir algunas que se consideraban apropiadas para cualquier tipo de problema. Si, en este caso, suprimimos el tratamiento de outliers y añadimos la selección de atributos mediante eliminación recursiva, quedándonos con el 75% de ellos, mejoramos todas las métricas estudiadas y gráficamente vemos como las predicciones son más acertadas. A continuación, se muestra el resultado de las modificaciones en el tratamiento de datos descrita.



Cabe destacar que las métricas absolutas como el MAE y el RMSE, al igual que el MAPE aumentan ya que se tienen en cuenta valores extremos que antes no contabilizaban, al haberse considerado outliers. Es importante tener en cuenta que los datos están normalizados, y que en la anterior gráfica no se consideraba el valor 1 de la clase ya que había sido considerado una instancia anómala. Es por eso por lo que la mayoría de las predicciones eran mejores en valores bajos. En este caso pasa lo mismo, pero ya nos aseguramos de que no es una causa directa de no haber tenido en cuenta todo el conjunto de datos. La principal razón parece ser la aglomeración de datos en los valores bajos de la clase, siendo escasas las instancia cuando la variable objetivo toma valores más altos.

El coeficiente de determinación es notablemente mejor, superando el umbral de 0.75 que determina una consistencia de los datos buena. Las predicciones son más precisas en términos de variabilidad, pero menos precisas en términos absolutos. Conseguimos una mayor consistencia a costa de alejarnos más de los valores reales. Estas métricas no se correlacionan, es decir, que una aumente o disminuye no significa que la otra métrica deba hacerlo ni inversamente.

Principalmente, el cambio de las métricas viene dado por un aumento en el número de datos tratados. Sigue habiendo una tendencia a subestimar las predicciones, pero menor al anterior modelado. Es posible que los datos previamente detectados como outliers hayan sido considerado errores, pero eran datos importantes y verdaderos para tener en cuenta, no anómalos. Aunque realicemos su detección con un percentil menor, sigue disminuyendo el coeficiente de determinación.

En definitiva, si hay que tomar una decisión sobre si incrementar el R2 o disminuir las demás métricas absolutas, elegiremos el primero ya que es una señal más real de que el modelo predictivo es fiable.