Actividad 1 Módulo 6

A continuación, se detallan las respuestas a las preguntas planteadas en la actividad, utilizando los conceptos de aprendizaje supervisado y preprocesamiento de datos.

1. ¿La variable objetivo se presta para una tarea de clasificación o regresión?

La variable objetivo (target) se presta para una tarea de clasificación.

El problema consiste en asignar cada muestra de vino a una de tres categorías predefinidas.

El objetivo no es predecir un valor numérico continuo, sino una etiqueta de clase discreta, lo cual es la definición de un problema de clasificación.

2. Propuesta de una segunda tarea: predecir el nivel de alcohol.

Si se toma una columna numérica como el nivel de alcohol para que sea la nueva variable objetivo, la tarea se convertiría en un problema de regresión.

A diferencia de la tarea anterior, aquí el objetivo sería predecir un valor numérico continuo (el grado de alcohol), lo cual es característico de los problemas de regresión.

3. Aplicación de codificación a la variable *target*

Sí, es necesario aplicar codificación a la variable *target*.

Aunque la variable *target* ya viene codificada numéricamente en el dataset de vinos de sklearn (usualmente como 0, 1, 2), si estuviera en formato de texto (ej. "Clase A", "Clase B"), sería indispensable transformarla.

Si se transforma la variable *target* con LabelEncoder, se asignará un valor numérico único a cada una de las tres clases de vino. Por ejemplo, las etiquetas podrían convertirse en 0, 1 y 2. Esto es un paso fundamental para que los algoritmos de Machine Learning puedan procesar las etiquetas.

4. Aplicación de MinMaxScaler y StandardScaler

Para el preprocesamiento de los datos, se deben aplicar las técnicas de escalado a las variables numéricas del dataset, como el nivel de alcohol, ácido málico o magnesio.

- MinMaxScaler: Transforma cada característica escalando sus valores a un rango específico, comúnmente entre 0 y 1.
- StandardScaler: Estandariza las características al remover la media y escalar a la varianza unitaria.

5. Visualización y comparación de técnicas de escalado

Al visualizar los datos después de aplicar ambos escaladores (por ejemplo, mediante histogramas o diagramas de dispersión), se observarían los siguientes efectos:

- Con MinMaxScaler: Todas las variables compartirían el mismo rango de valores (ej. 0 a 1), lo que puede ser útil para algoritmos sensibles a la escala.
- Con StandardScaler: Las distribuciones de las variables se centrarían en torno a una media de 0 y una desviación estándar de 1. Esto es especialmente beneficioso para algoritmos que asumen que los datos están distribuidos normalmente.

6. Escalado más adecuado para un modelo KNN

Para un modelo como K-Nearest Neighbors (KNN), el escalado

StandardScaler sería generalmente más adecuado.

Justificación: KNN se basa en la medición de distancias (como la distancia euclidiana) entre los puntos de datos. Si las variables tienen escalas muy diferentes, aquellas con rangos mayores dominarán el cálculo de la distancia.

StandardScaler normaliza las variables, asegurando que todas contribuyan de manera equitativa al cálculo de la distancia, lo que suele mejorar el rendimiento y la precisión del modelo.

Reflexiones Adicionales

- Diferencia con el dataset Iris: El dataset de vinos probablemente contiene un mayor número de características y una mayor variabilidad en las escalas de esas características en comparación con el dataset de Iris. Esto hace que el paso de escalado sea aún más crítico para el rendimiento del modelo. El flujo de preprocesamiento, sin embargo, sería muy similar: inspección, codificación de la variable objetivo (si es necesario) y escalado de características.
- Aprendizaje sobre la formulación de problemas: Esta actividad enseña que el primer paso crucial en Machine Learning es entender la naturaleza de la variable objetivo (*target*). Si es categórica, se formula un problema de clasificación; si es numérica y continua, se trata de un problema de regresión. Esta decisión determina las técnicas de preprocesamiento y los modelos que se pueden aplicar posteriormente.