



Módulo 6

Sesión N° 1



ACTIVIDAD:



¿Clasificamos o predecimos? Explorando datos de vinos para
entrenamiento de modelo

Objetivo: Aplicar los conocimientos adquiridos sobre aprendizaje supervisado, codificación, escalado y análisis de tareas de Machine Learning para formular correctamente un problema de regresión o clasificación, utilizando un dataset real con múltiples variables numéricas.



Contexto

1. Importar e inspeccionar el dataset de vinos desde la librería `sklearn.datasets` (ver explicación abajo).
2. Identificar si la tarea es de clasificación o regresión según el objetivo planteado.
3. Aplicar codificación a la variable target si es necesario.
4. Aplicar escalado de variables numéricas con `MinMaxScaler` y `StandardScaler`.
5. Comparar visualmente los resultados del escalado.
6. Reflexionar sobre cómo estos pasos afectan la calidad del aprendizaje del modelo.



Explicación técnica:

A diferencia de otros ejercicios donde cargamos un archivo CSV desde una URL externa usando `pandas.read_csv()`, en esta actividad utilizamos un dataset embebido en la librería `scikit-learn`. Esto nos permite trabajar con datos estructurados sin necesidad de descargar archivos.

Tiempo estimado: 45 minutos

Formato de ejecución: grupal





Requerimientos:

Estás trabajando en un modelo para ayudar a una bodega a analizar distintas características químicas de sus vinos. El dataset contiene variables como el nivel de alcohol, ácido málico, magnesio, etc., y una columna target que indica a qué categoría de vino pertenece la muestra (tres clases en total).

Tu objetivo es proponer una tarea de aprendizaje automática con este dataset y aplicar los pasos previos al entrenamiento del modelo.

Tu misión:

1. ¿La variable target se presta para una tarea de clasificación o regresión? Justifica.
2. Propón una segunda tarea con una de las columnas numéricas como variable objetivo (por ejemplo, predecir el nivel de alcohol). ¿Sería clasificación o regresión?
3. Aplica codificación si consideras necesario. ¿Qué sucede si transformas target con LabelEncoder?
4. Aplica MinMaxScaler y StandardScaler a las variables numéricas.
5. Visualiza los resultados y compara ambas técnicas. (Opcional)
6. ¿Qué escalado considerarías más adecuado si aplicarás KNN con estos datos? ¿Por qué?

¿En qué se diferencia este dataset del usado en la actividad con Iris? ¿Cambiarías algo del flujo de preprocesamiento? ¿Qué aprendiste sobre cómo formular un problema de aprendizaje automático correctamente?

