

Aprendizaje Automático

Trabajo Práctico 4

Para la realización de los siguientes ejercicios, se deben descargar los datos mediante el enlace <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>.

Ejercicio 1. A partir del conjunto de datos, se pide lo siguiente.

1. Describir el tipo de los atributos involucrados y la cantidad de registros. ¿Existen registros con datos faltantes? ¿Qué se debería hacer en tal caso?
2. Eliminar los registros para los cuales la calidad del vino tiene un puntaje igual a 6. Luego, reemplazar este atributo por uno que represente la función objetivo para una instancia \mathbf{x} ,

$$f(\mathbf{x}) = \begin{cases} \text{REGULAR} & \text{si la calidad tiene un puntaje inferior a 5,} \\ \text{BUENO} & \text{si la calidad tiene un puntaje igual a 5,} \\ \text{EXCELENTE} & \text{si la calidad tiene un puntaje superior a 5.} \end{cases}$$

Ejercicio 2. Se desea conocer si un vino es regular, bueno o excelente.

1. Dividir aleatoriamente el conjunto de datos para utilizar una parte de los mismos como conjunto de entrenamiento (el 80 %) y otro como conjunto de prueba (el 20 %).
2. Aplicar el método de vecinos más cercanos con $k = 3, 5, 7$. ¿Cuál es el valor de k recomendado? Justificar la respuesta.
3. Para el valor de k recomendado en el ítem anterior, aplicar el método de vecinos más cercanos ponderado. ¿Se logra una mejora? Justificar.

Ejercicio 3. En este punto se aplica un aprendizaje no supervisado.

1. Aplicar el método de k -medias para el conjunto de entrenamiento obtenido en el [Ejercicio 2](#), utilizando $k = 2, 3, 4, 5$.
2. ¿Cuál es la cantidad de *clusters* recomendada? Justificar la respuesta.