Exploración de datos y clasificación

Fabián Villena

Enero 2024

Se le solicita ajustar (o entrenar) un modelo de aprendizaje automático para predecir etiquetas sobre ejemplos nuevos.

En particular, trabajaremos con el *Pima Indians Diabetes Dataset*. Este conjunto de datos tiene como objetivo predecir si un paciente tiene o no tiene diabetes en base a una serie de características. La muestra consiste en datos de mujeres de al menos 21 años de descendencia indígena del grupo *Pima*. Específicamente, existen varios atributos médicos y una variable objetivo, la columna *Outcome*. Esta columna alcanza un valor igual a 1 cuando la paciente posee diabetes, mientras que es 0 cuando no posee la enfermedad.

Deben realizar un flujo clásico al momento de crear modelos de aprendizaje automático.

El conjunto de datos se encuentra en la siguiente dirección:

`https://github.com/fvillena/biocompu/blob/2023/data/diabetes.csv` [1]

Preguntas

Responda las siguientes preguntas en un *Jupyter Notebook* con código desarrollado en el lenguaje de programación Python.

1. ¿Cuántas instancias y cuántos atributos contiene el conjunto de datos?
2. ¿Cuál es la distribución de clases en la variable objetivo *Outcome*?
3. ¿Cuántos pacientes mayores a 40 años padecen de diabetes?
4. Separe el conjunto de datos en un subconjunto de entrenamiento y uno de prueba
5. Ajuste un modelo de clasificación con el subconjunto de entrenamiento.
6. Prediga utilizando el modelo ajustado sobre el subconjunto de prueba.
7. Calcule la métrica de *accuracy* del modelo sobre el conjunto de prueba.

[1] Para importar el conjunto de datos, probablemente usted necesite el enlace directo al archivo, el cual está en `https://raw.githubusercontent.com/fvillena/biocompu/2023/data/diabetes.csv`