Guía del Dataset Diabetes (UCI) – Dificultad: baja

Descripción del dataset

El conjunto de datos con el que trabajaremos proviene originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales de Estados Unidos. El objetivo del ejercicio es predecir si un paciente tiene diabetes o no, basándose en ciertas mediciones diagnósticas incluidas en el conjunto de datos. Se impusieron varias restricciones en la selección de estas instancias a partir de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de edad y de ascendencia indígena Pima.

Contenido

El conjunto de datos consta de varias variables predictoras médicas y una variable objetivo, denominada Outcome (resultado). Las variables predictoras incluyen el número de embarazos que ha tenido la paciente, su índice de masa corporal (IMC), nivel de insulina, edad, entre otros.

Referencia original (por si alguien tiene curiosidad)

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

Atributos del dataset

Ver https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Recomendaciones para resolver el problema (rúbrica)

1. En función de la descripción del problema, qué tipo de problema debemos resolver? es un problema de regresión o de clasificación?

2. Lee en un dataframe el archivo de datos que se encuentra en https://raw.githubusercontent.com/gustavovazquez/datasets/refs/heads/main/diabetes_copy.csv
3. Muestra el contenido inicial del dataframe (head)
4. Haz una descripción del contenido (describe)
5. Muestra los tipos de las variables del dataframe (dtypes)
6. Verás que hay una columna que tiene un ID para cada paciente. Como hemos mencionado, las columnas que hacen referencia a IDs no deben ser parte del modelo. Puedes borrarla -drop- o mejor, copia los valores de los IDs en los nombres de las filas (la segunda opción sería la indicada en caso de que los IDs hagan referencia específicamente al paciente, por ej. usando un nro de cédula).
7. Obtén la matriz de correlación y revisa si hay predictores correlacionados. Si los hay, elimina alguno con un criterio definido.
8. Convierte las columnas categóricas a su codificación one-hot-ecoding
9. Arma un pipeline con dos etapas: normalización (minmaxscaler) y el método de machine learning indicado para el tipo de problema
10. Calcula las métricas clásicas (según el tipo de problema supervisado que estamos considerando)
11. Discute para el tipo de problema que tenemos cómo se interpretan las métricas fundamentales de clasificación