Assignment 1

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

Tasks:

- Create a column called “BMI_category” using the “BMI” column and the general medical guidelines for BMI types.
- Split the data into 2 parts: train and val. Use 20% of the data for val.
- Apply Standard Scaler on the numeric features. Fit and transform on train and only transform on val.
- Apple One-hot Encoding to the categorical features. Fit and transform on train and only transform on val.
- Build a KNN classifier. Experiment with different values of k (3, 5, 7) and select the value with the highest f1 score.
- Build a Decision Tree classifier. Experiment with different values of max_depth (3, 5, 7) and choose the value with the highest f1 score.
- Build an inference pipeline:
  - Take the features as input for a test sample
  - Save the standard scaler and one-hot encoder
  - Save the best model
  - Load the above in the inference script and apply them on the test sample
  - Print the predicted class
  - Use 5 samples from the val set to demonstrate this script