A3
Meta learning

Data

- Data: https://archive.ics.uci.edu/ml/machine-learning-databases/**heart-disease**/cleve.mod
  - —> Download cleve.mod data
- Binary classification between healthy (buff) or with heart-disease (sick).

Grade 4

- Choose 5 classification algorithms
- Preprocess data (standardisation, missing value handling, categorical feature handling, etc.)
- Find optimal set of hyper-parameters for each of the 5 classification algorithms. For each algorithm, do:
  - Employ Bayesian optimisation using Gaussian Processes to identify best hyper-parameters for given algorithm on data set (use `gp_minimize` from `scikit-optimize`).
  - Document search space boundaries and best fit values. What hyper-parameters did you find. Are some of them on the boundary?

Grade 3

- Compare optimal models across 5 classification algorithms (all pairwise comparisons):
  - Use McNemar's test for classifier comparisons and report “best” models (using Edward’s correction); using `mlxtend`.
  - Compare the results using cross validation (CV) on accuracy.
  - Plot accuracy distribution across 10-fold CV for 5 classification algorithms.
  - Compare results from CV and McNemar’s test.
  - Is there a single best algorithm? If not, what are the algorithms outperforming the rest?

Grade 2

- Using “the best” algorithm (if there are multiple ones, choose one):
  - Perform a greedy feature selection
    - From 1…N features do:
      - Always add feature that improves performance (using hyper-parameters determined earlier).
      - Stop when accuracy improvement is very small, or performance gets worse.
  - Compare performance to PCA with features that retain 95% of explained variance.

Grade 1

- Using “the best” algorithm with the greedy feature selection:
    - Use the `SHAP` library to explain the predictions
        - Create a `waterfall` and `beeswarm` plot and interpret them — what features are morst important
        - Does this correlate to the feature selection you have used earlier? I.e., are the features which have been identified first/earlier by the greedy procedure also the most important ones determined by `SHAP`?